Int J Artif Intell Educ (2017) 27:694–728
DOI 10.1007/s40593-017-0143-2

ARTICLE

Assessing Students' Use of Evidence and Organization in Response-to-Text Writing: Using Natural Language Processing for Rubric-Based Automated Scoring

Zahra Rahimi[1] · Diane Litman[1] · Richard Correnti[1] · Elaine Wang[1] · Lindsay Clare Matsumura[1]

Published online: 13 March 2017

Abstract This paper presents an investigation of score prediction based on natural language processing for two targeted constructs within analytic text-based writing: 1) students' effective use of evidence and, 2) their organization of ideas and evidence in support of their claim. With the long-term goal of producing feedback for students and teachers, we designed a task-dependent model, for each dimension, that aligns with the scoring rubric and makes use of the source material. We believe the model will be meaningful and easy to interpret given the writing task. We used two datasets of essays written by students in grades 5–6 and 6–8. Our experimental results show that our task-dependent model (consistent with the rubric) performs as well as if not outperforms competitive baselines. We also show the potential generalizability of the rubric-based model by performing cross-corpus experiments. Finally, we show that the predictive utility of different feature groups in our rubric-based modeling approach is related to how much each feature group covers a rubric's criteria.

✉ Zahra Rahimi
zar10@pitt.edu

Diane Litman
dlitman@pitt.edu

Richard Correnti
rcorrent@pitt.edu

Elaine Wang
elw51@pitt.edu

Lindsay Clare Matsumura
lclare@pitt.edu

[1] Learning Research and Development Center, 3939 O'Hara Street, Pittsburgh, PA 15260, USA

Keywords Automatic essay assessment · Analytical writing in response to text · Evidence · Organization · Task-dependent · Feedback · Natural language processing

Introduction

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 (Correnti et al. 2013). 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 (Correnti et al. 2013). 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.

Because scoring text-based writing assessments is typically labor intensive and requires extensive training and expertise on the part of raters to obtain reliable scores, automated essay scoring has been proposed as a fast, effective, and affordable solution to the problem of assessing student writing at scale. For example, a recent contrastive analysis of 9 state-of-the-art systems on 8 essay scoring prompts drawn from high-stakes assessments claimed that Automated Essay Scoring (AES) systems had as high a level of agreement with human graders as human graders had with each other (Shermis and Hamner 2012). However, critics of AES argue that AES scores typically under-represent the construct of writing (Condon 2013; Perelman 2013) and even ardent supporters of AES acknowledge its limitations (Shermis and Hamner 2012; Deane 2013).

First, many essay assessment systems rely on holistic rather than trait-based rubrics (Attali and Burstein 2006; Elliot 2003; Page 2003; Attali et al. 2013), and thus tend to focus on summative rather than formative assessment. While holistic methods are typically more efficient and provide more reliable scores, trait-based methods are better at providing diagnostic insight on student performance (Bacha 2001; Weigle 2002). Such insight is particularly useful for systems that not only score but also provide formative feedback. Even when systems do trait-based scoring, critics maintain that trait-based AES has focused on surface dimensions of writing such as grammar rather than more substantive dimensions (Attali and Powers 2008; Perelman 2012). Our system for automatically scoring the RTA is trait-based rather than holistic, scores two of the RTA's substantive writing traits (namely, Evidence and Organization), and is motivated by formative rather than summative assessment.

Second, in terms of writing tasks, most systems (whether holistic or trait-based) focus on assessing writing in response to open-ended prompts (Attali and Burstein 2006; Crossley et al. 2013; Elliot 2003; Lee et al. 2008; Page 2003; Klebanov and Higgins 2012) rather than in response to text. They usually use more generic rubrics instead of task-specific ones. One advantage of task-dependent rubrics is the ability

to provide feedback that is better aligned with the task. Existing systems 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 (Shermis and Burstein 2003). Our system for automatically scoring the RTA focus on assessing writing in response to text using task-dependent rubrics.

Third, in terms of scoring method, many AES systems do not consider construct validity (Condon 2013; Perelman 2012). Existing AES systems are limited in evaluation of higher-order aspects of writing, such as the quality of content and its organization. For example, AES achieves high reliability in evaluation of content and ideas mostly by using "bag-of-words" approaches that bear little relationship to the scoring rubric for the construct (Landauer et al. 1998; Attali and Burstein 2006; Attali 2011). In contrast, our model for automatically scoring the RTA is consistent with the rubric criteria and easily explainable. Others in the AES community are similarly arguing that automated scoring models should reflect important aspects of the construct being measured, following common practice in the measurement community. That is, dimensions of the construct should be well represented by the features used in the scoring model, and the features contained in the model should not be irrelevant to the rubric for the construct (Loukina et al. 2015). A model with construct validity has greater potential to generate useful formative feedback to students and teachers.

Finally, current AES systems typically score writing that is generated by upper middle-school, secondary, post-secondary students, or by adults for a high-stakes exam (Burstein et al. 1999; Deane et al. 2013; Klebanov and Higgins 2012). For example, the sample of essays in the contrastive analysis of Shermis and Hammer (Shermis and Hamner 2012) described above were from Grades 7, 8, and 10. Our work, in contrast, focuses on writing in Grades 5 through 8, which poses challenges for existing AES methods as RTA essays are typically shorter, contain more grammatical and spelling errors, and are less sophisticated in terms of use and organization of evidence. Our work thus tackles the challenge of using computational techniques on data that is particularly noisy given the stage of writing development of the students.

In the following sections, we first introduce the previous research on this topic. Next we talk about the data, the rubric dimensions and the prompt that we use in our study. Then, we explain the two models we designed to extract features for the Evidence and Organization dimensions of our rubric. Next, we discuss the experiments and results. Finally, we recap our conclusions and discuss future work. Our results show that in general, our rubric-based task-dependent model performs as well as (if not better than) the rigorous baselines we used. Moreover, the combination of our new features with the baseline features often yields better results than either the proposed or baseline features in isolation. Both within-corpus and cross-corpus experiments yield similar conclusions, supporting the robustness of our approach. Finally, feature ablation studies suggest that feature utility is related to rubric coverage.

Related Work

Natural Language Processing techniques have been used to evaluate both the content and organization of writing. One approach of evaluating the content of student

essays is to detect whether they are off-topic or on-topic (Louis and Higgins 2010; Higgins et al. 2006). Adherence to the prompt (Persing and Ng 2014) is another way to measure text topicality. Yet another approach to estimating the quality of content is to compare the essay to sets of training essays with different scores (Attali and Burstein 2006; Kakkonen et al. 2005; Xie et al. 2012). These prior studies differ from our response-to-text task in that they do not target source-based writing in which the quality of content should be measured with regard to how the essays use the source material.

Source-based writing refers to types of writing that require students to generate responses that are based on and that reference one or multiple source text(s). Generally, responses are expected to demonstrate close reading and deep comprehension of texts through effective use of evidence from the source text(s). For example, having read a novel, students might be asked to analyze the main theme, providing evidence from the novel. Or, having read two articles representing opposing viewpoints on a topic, students might be asked to write an opinion or argumentative essay in which they use points from the text to support their claim or rebut the opposing perspective. Professional standards for literacy in K-12 education are increasingly emphasizing such source-based writing (e.g., NCTE/IRA, 2012; NGAC/CCSSO, 2011).[1]

In contrast, quality of content is evaluated with regard to integrating information from the source materials in Kakkonen et al. (2005) and Lemaire and Dessus (2001). These studies also differ from our task. In our work, we care about pieces of evidence that students provide from the source material. So, the task is beyond simply deciding if the essay is semantically similar to the source material or not. To be able to score based on the criteria in the rubric and to have the ability of providing feedback based on the detailed information in the essays, we need to localize pieces of evidence. With an ultimate goal of scoring essays, Klebanov et al. (2014) evaluated different content importance models that help predict which parts of the source material should be selected by the students. This study is in a similar direction with our preliminary study (Rahimi and Litman 2016) for automatically extracting important pieces of evidence from the source material.

Another related area of research is to first find argumentation components using argumentation mining techniques, and then use the results of argumentation mining for scoring the essays (Ong et al. 2014; Burstein et al. 2003a; Song et al. 2014; Persing and Ng 2015). Mostly, argumentation mining in the domain of essay evaluation is applied to persuasive essay corpora written in response to a prompt (Stab and Gurevych 2014a,b) rather than to source-based writing. Similarly, the definition of Evidence in our task is related to source material and is different from more general definitions of Evidence, Premise, etc. in persuasive essays. Another difference from prior work is that in our study, the essays are written by young kids and we do not expect them to follow a sophisticated argumentation structure.

[1]International Reading Association/National Council of Teachers of English (IRA/NCTE; 2012). Standards for the English Language Arts. IRA/NCTE. National Governors Association Center for Best Practices, Council of Chief State School Officers (NGAC/CCSSO, 2011). Common Core State Standards English Language Arts standards. Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

As a construct, 'Organization' has figured in systems for scoring student writing for decades. When organization is considered as a separate dimension, some surface features of organization are considered. Such surface features include: effective sequencing; strong inviting beginning; strong satisfying conclusion; and smooth transitions.[2] Assessments aligned to the Common Core State Standards (CCSS), the academic standards adopted[3] widely in 2011 that guide K-12 education, reflect a shift in thinking about the scoring of organization in writing to consider the coherence of ideas in the text.[4] The consideration of idea coherence as a critical aspect of organization of writing is relatively new.

Notably, prior studies in natural language processing have examined the concept of discourse coherence, which is highly related to the coherence of topics in an essay, as a measure of the organization of analytic writing. For example, in Somasundaran et al. (2014) the coherence elements are adherence to the essay topic, elaboration, usage of varied vocabulary, and sound organization of thoughts and ideas. In Scott and McNamara (2011) the elements are effective lead, clear purpose, clear plan, topic sentences, paragraph transitions, organization, unity, perspective, conviction, grammar, syntax, and mechanics.

Many computational methods are used to measure such elements of discourse coherence. Vector-based similarity methods measure lexical relatedness between text segments (Foltz et al. 1998) or between discourse segments (Higgins et al. 2004). Centering theory (Grosz et al. 1995) addresses local coherence (Miltsakaki and Kukich 2000). Entity-based essay representation along with type/token ratios for each syntactic role is another method to evaluate coherence (Burstein et al. 2010) that is shown in Burstein et al. (2013) to be a predictive model on a corpus of essays from grades 6-12. Lexical chaining addresses multiple aspects of coherence such as elaboration, usage of varied vocabulary, and sound organization of thoughts and ideas (Somasundaran et al. 2014). Discourse structure is used to measure the organization of argumentative writing (Cohen 1987, Burstein et al. 1998, 2003b). All these works rely on lexical information to measure coherence. In contrast, our proposed model uses more coarse-grained topic information. Based on the rubric, we are interested in localizing the pieces of evidence for different topics in essays and evaluate the transition between these topics. For this purpose, we proposed the concept of topic-grid and topic-chain.

In previous studies, assessments of text coherence have been task-independent, which means that these models are designed to be able to evaluate the coherence of the response to any writing task. Task-independence is often the goal for automated scoring systems, but it is also important to measure the quality of students' organization skills when they are responding to a task-dependent prompt. One advantage

[2]Retrieved from http://www.rubrics4teachers.com/pdf/6TRAITSWRITING.pdf, February 25, 2015

[3]The CCSS were adopted by 46 states and the District of Columbia in 2010-2011. Since then, some states have withdrawn their adoption of the standards. Currently, 42 states + DC are still using the standards. These include the states that our writing samples are from.

[4]See, e.g., Grades 4 and 5 Expanded rubric for analytic and narrative writing retrieved from http://www.parcconline.org/sites/parcc/files/Grade 4-5 ELA Expanded Rubric FOR ANALYTIC AND NARRATIVE WRITING_0.pdf

of task-dependent scores is the ability to provide feedback that is better aligned with the task. Our model to evaluate the Organization dimension is task-dependent which means it is designed based on the detailed criteria in the rubric and makes use of the source material by evaluating the transition of important topics and pieces of evidence adopted from the source in essays.

Our preliminary studies addressing the task-dependent automatic scoring of both the Evidence (Rahimi et al. 2014) and Organization (Rahimi et al. 2015) dimensions of the RTA were motivated by the differences with prior work discussed above. Our initial method for localizing and analyzing the quality of Evidence in source-based writing was presented in Rahimi et al. (2014) and evaluated on a corpus of essays from grades 5–6. Here we extend this earlier work by taking advantage of a second corpus of essays from grades 6–8 (obtained from a different school district) to conduct new types of evaluations such as using cross-validation within each corpus separately and combined, and performing cross-corpus training versus testing. We also address an unbalanced score distribution issue that occurs in both corpora using an oversampling method, and conduct new feature ablation studies. Our initial method for analyzing the organization of ideas and evidence in source-based writing was presented in Rahimi et al. (2015). With the motivation of experimenting on a bigger corpus, in the current paper we conduct several new evaluations that combine our two available datasets from different grades and schools to create a third larger corpus.

Data

Our data consists of students' writings from the RTA introduced in Correnti et al. (2013). Specifically, we have datasets from two different age groups (grades 5–6 and grades 6–8) which represent different levels of writing proficiency. The two datasets are also from two different school districts.

The administration of the RTA involved having the classroom teacher read aloud a text while students followed along with their own copy. The text is an article from *Time for Kids* about a United Nations effort (the Millennium Villages Project) to eradicate poverty in a rural village in Kenya. After a guided discussion of the article as part of the read-aloud, students wrote an essay in response to a prompt that requires them to make a claim and support it using details from the text. A small excerpt from the article, the prompt, and three student essays from grades 5–6 are shown in Table 1.

Our datasets (particularly responses by students in grades 5-6) have a number of properties that may increase the difficulty of the automatic essay assessment task. The essays in our datasets are short[5] and have many spelling[6] and grammatical errors.

[5]This may be due to the fact that essays are written by students in grades 5–8 and also because the writing task is source-based. In the contrastive analysis discussed earlier (Shermis and Hamner 2012), source-based essays (Mean = 119.97, SD = 58.88) were reported to be much shorter than traditional essays (Mean = 354.18, SD 197.63).

[6]The effect of spelling correction on improving the quality of AES was investigated in our very first study on a subset of the grades 5–6 corpus (Rahimi et al. 2014). Perfect spelling correction was found to yield a small positive increase in system performance.

Table 1 A small excerpt from the *Time for Kids* article, the prompt, and sample low and high-scoring essays with supporting evidence in bold from grades 5–6

Excerpt from the article: The people of Sauri have made amazing progress in just four years. The Yala Sub-District Hospital has medicine, free of charge, for all of the most common diseases. Water is connected to the hospital, which also has a generator for electricity.

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 both Evidence and Organization dimensions: Yes because **ending poverty is achievable in my lifetime** because you can tell that our nations is helping the homeless by offering them food shelther and by putting out things or stands that help donate to people who are poverty. & in other countries do help to, like for example our country will sometimes help other countries if they have poverty & if adults or **kids are dieing every day** by offering them clothes food and sometimes some shelter. Poverty can be stopped in my lifetime if we help or if we try help people or atleast help and so if we do helpful we today can help stop proverty just by doing & putting 1 step in.

Essay with Evidence score of 4 : 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.

Essay with Organization score of 4: This story convinced me that **"winning the fight against poverty is achievable** because they showed many example in the beginning and showed how it changed at the end. One example they sued show a great amount oF change when they stated at first most people thall were ill just stayed in the hospital Not even getting treated either because of the cost or the hospital didnt have it, **but at the end it stated they now give free medicine to most common deseases**.

Anotehr amazing change is in the beginning majority of the **childrenw eren't going to school because the parents couldn't affford the school fee**, and the kdis didnt like school because tehre was **No midday meal**, and **Not a lot of book, pencils, and paper**. Then in 2008 the **perceNtage of kids going to school increased** a lot because **they Now have food to be served aNd they Now have more supplies.** So Now theres a better chance of the childreN getting a better life

The last example is Now they dont have to worry about their families starving because **Now they have more water and fertalizer**. They have made some excellent changes in sauri. Those chaNges have saved many lives and I think it will continue to change of course in positive ways

Some statistics about the datasets are in Table 2. On average the essays in the 6–8 dataset are longer than essays in the 5–6 dataset. They have more unique words and longer sentences.

The student responses have been assessed on five dimensions (Analysis, Evidence, Organization, Style/vocabulary and MUGS (mechanics/usage/grammar/syntax)) , each on a scale of 1-4 (Correnti et al. 2013). The Analysis dimension is about addressing the prompt, understanding the text and insightful and clear conclusions. The Evidence dimension is related to demonstrating integral use of selected details from the text to support the claim. The Organization rubric is about clear structure of the essay and logical flow of the ideas. The Style rubric addresses use of sophisticated language and vocabulary. Finally, Mugs is about errors in mechanics, usage, grammar, and syntax. The standards stay fixed across grade levels (and thus across the datasets). Half of the assessments are scored by an expert. The rest are scored by undergraduate students trained to evaluate the essays based on the criteria. All the raters were blind to the grades to which the essays belonged. The corpus from grades 5–6 consists of 1569 essays, with 602 of them double-scored for inter-rater reliability. The other corpus includes 809 essays, with almost all of them (802) double-scored (9 of these essays do not have score for Evidence dimension). Inter-rater agreement (Quadratic Weighted Kappa) on the double-scored portion of the grades 5-6 and 6-8 corpora respectively are 0.67 and 0.73 for Evidence and 0.68 and 0.69 for Organization.

The correlation between the scores of Organization and Evidence dimensions (for rater 1) for 5-6 and 6-8 corpora respectively are (pearson = 0.55 , spearman = 0.54) and (pearson = 0.50, spearman = 0.48) with all p-values $\leq$ 0.0001. It is possible to have an essay that scores well on one dimension but poor on the other one although it is more common to have a good Organization score but a poor Evidence score than vice versa. As shown in Table 3, there are 48 essays that have poor organization score but good Evidence score on 5–6 dataset and 65 essays vice versa (the upper right and the middle left triangles respectively). There are only 8 essays that have poor Organization score but good Evidence score on 6–8 dataset and 79 essays vice versa (the middle right and the lower left triangles respectively).

Table 2 The two datasets' statistics

Dataset		Mean	SD
5–6 grades	# words	161.25	92.24
	# unique words	93.27	40.57
	# sentences	9.01	6.39
	# paragraphs	2.04	1.83
6–8 grades	# words	207.99	104.98
	# unique words	113.14	44.14
	# sentences	12.51	7.53
	# paragraphs	2.71	1.74

Table 3 The distribution of the Evidence and the Organization scores with respect to each other on the two datasets

DataSet	Organization/Evidence	1	2	3	4
5–6	1	251	122	**21**	**2**
	2	174	349	160	**25**
	3	**33**	104	128	85
	4	**13**	**19**	25	58
6–8	1	77	46	**4**	**0**
	2	84	169	59	**4**
	3	**30**	97	82	36
	4	**15**	**34**	24	49

Bold indicates the number of essays that have poor score on one dimension but good on the other one

In this paper we focus only on predicting the score of the Evidence and the Organization dimensions,[7] which are the two dimensions most related to argumentation. The distributions of the Evidence and the Organization scores are in Table 4. Higher scores on the 6–8 corpus indicate that the essays in this dataset have better quality in terms of Evidence and Organization than the student essays in the 5–6 dataset. The rubric for the Evidence and the Organization dimensions are shown respectively in Tables 5 and 7.

Modeling the Source Article

To build both Evidence and Organization models, we use the information in the source *"Time For Kids"* text, where exhaustive list of topics, important topic words and examples are provided manually by experts (see Tables 17, 18, and 19 in the Appendix). Similarly, in other studies on evaluation of content (typically in short answer scoring), the identification of concepts and topics is often manual (Liu et al. 2014). First, experts provide a list of important words for each of the main topics in the article (Table 17). Second, the experts provide a comprehensive list of topics which includes every specific example from the text related to each topic (Table 19). Since the source text explicitly addresses the conditions in a Kenyan village before and after the United Nations intervention, and since the prompt leads students to discuss the contrasting conditions at these different time points, topics provide evidence for the "before" and "after" states, respectively. That is, except for some topics which do not have a temporal aspect, for each major topic t the experts define two sub-topics t_{before} and t_{after} by listing specific examples related to each sub-topic. Finally, the experts remove the temporal aspect of topics from the comprehensive list of examples by merging the "after" states to a single "Topic7" which is about progress made in the village as it originally was represented in the article (Table 18). This is because

[7]The other three dimensions of RTA are Analysis, Style, and MUGS (Mechanics, Usage, Grammar, Spelling).

Table 4 The distribution of the Evidence and the Organization scores on the two datasets

Dimension	DataSet/Score	1	2	3	4	total
Evidence	5–6	471	594	334	170	1569
		(30 %)	(38 %)	(21 %)	(11 %)	
	6–8	206	341	165	88	800
		(26 %)	(42 %)	(21 %)	(11 %)	
Organization	5–6	396	708	350	115	1569
		(25 %)	(46 %)	(22 %)	(7 %)	
	6–8	127	316	244	122	809
		(16 %)	(39 %)	(30 %)	(15 %)	

in modeling the Evidence dimension, we do not care about the temporal aspect of the topics.

Modeling the Evidence Dimension

Introduction to Evidence Rubric

The Evidence rubric (see Table 5) takes into account four criteria related to the quality of text evidence provided in the response. First, we consider the number of pieces of evidence used. More evidence (i.e., above 3) is scored higher. Second, we consider the relevance of the evidence to the central idea. Writing that includes cogent evidence is scored high, while writing that provides irrelevant details is scored low. The third criterion is the specificity of the evidence provided. Writing that features detailed, specific evidence is scored high, while responses that feature cursory, general references is scored low. Finally, the extent to which the evidence is elaborated upon is considered. Strong responses feature evidence that help support and develop the main idea. Evidence is weak when it is just presented as a short phrase or listed in a sentence. The rubric also notes that when the response features a summary of the whole text or directly copies from the source text, it automatically scores 1.

Features to Model the Rubric

As discussed above, one goal of our research in predicting scores is to design a small set of rubric-based meaningful features that performs acceptably and also models what is actually important in the rubric. To this end, we designed several groups of features, primarily addressing one criterion in the rubric. Below, we explain each of the features and its relation to the rubric. Each group of features is indicated with an abbreviation that relates it to the corresponding criteria in the rubric in Table 5.

(1) Number of Pieces of Evidence (NPE) addresses the first row of the rubric, e.g., if there are fewer than 2 pieces of evidence, score the essay as 1. For calculating

Table 5 Rubric for the Evidence dimension of RTA

	1	2	3	4
Number of pieces of evidence	Features one or no pieces of evidence (NPE)	Features at least 2 pieces of evidence (NPE)	Features at least 3 pieces of evidence (NPE)	Features at least 3 pieces of evidence (NPE)
Relevance of evidence	Selects inappropriate or irrelevant details from the text to support key idea (SPC); references to text feature serious factual errors or omissions	Selects some appropriate and relevant evidence to support key idea, or evidence is provided for some ideas, but not actually the key idea (SPC); evidence may contain a factual error or omission	Selects pieces of evidence from the text that are appropriate and relevant to key idea (SPC)	Selects evidence from the text that clearly and effectively supports key idea
Specificity of evidence	Provides general or cursory evidence from the text (SPC)	Provides general or cursory evidence from the text (SPC)	Provides specific evidence from the text (SPC)	Provides pieces of evidence that are detailed and specific (SPC)
Elaboration of Evidence	evidence may be listed in a sentence (CON)	Evidence provided may be listed in a sentence, not expanded upon (CON)	Attempts to elaborate upon Evidence (CON)	Evidence must be used to support key idea / inference(s)
Plagiarism	Summarize entire text or copies heavily from text (in these cases, the response automatically receives a 1)			

The abbreviations in the parentheses identify the corresponding feature group discussed in the Evidence modeling section of this paper that is aligned with that specific criteria

NPE, we use manually provided topics in Table 17. 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[8] 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.[9] Each topic is only counted as a piece of evidence once to avoid redundancy.

(2) Concentration (CON) 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 1 or 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.

(3) Specificity (SPC) High quality evidence includes specific examples from different parts of the text, or an explanation of why the evidence is important. We use the manually provided list of topics and examples in Table 18. For each of the examples we need to answer the 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.

(4) Word Count (WOC) is used as a fallback feature because our features do not yet completely cover all rubric cells, and in prior work and in our own data, longer essays tend to receive higher scores.

The value of the features for the example essays with Evidence score of 4 and 1 in Table 1 are shown in Table 6. For the high-scoring essay, the value of the NPE feature is four because all four topics in Table 17 are mentioned in the essay. The essay is not concentrated (CON=0) because there are more than three sentences with topic words in the essay. The value of the Specificity for topics three and eight are 1 because of these two pieces of evidence respectively: *couldn't afford medicine* and *winning the fight of poverty is achievable* which are bolded in Table 1.

For the low-scoring essay, the value of the NPE feature is one because only one topic from Table 17 is mentioned in the essay. The essay is concentrated (CON=1)

[8]For all window-based features, we set the window size to 6 by trying some different values on our training data of corpus 5–6 and choosing the best one. We use this same value on all other experiments on two other datasets.

[9]Two words overlap can cause some errors. But since the size of the lists are very short, the problem is reduced.

Table 6 Feature vector representation of the high and low-scoring Evidence essays from Table 1

	NPE	CON	WOC	SPC (for each topic)							
				1	2	3	4	5	6	7	8
High-scoring	4	0	178	0	0	1	4	3	3	5	1
Low-scoring	1	1	118	0	0	1	1	0	0	0	1

because there are less than three sentences with topic words in the essay. The value of the Specificity for topics three and four are 1 because of *adults or kids* and *kids are dieing every day* respectively. The value of the topic eight is 1 because of *winning the fight of poverty is achievable* which are bolded in Table 1.

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 a condition of poverty that was improved, but there is other relevant evidence in the text that you also need to focus on, such as the lack of fertilizer for crops." For teachers, we envision providing summary information such as students' weakness in elaborating on the evidence they provided.

Modeling the Organization Dimension

Introduction to Organization Rubric

The Organization dimension rubric in Table 7 is made of four main criteria that relates to how and to what extent students present their ideas in an organized and logical way. The first criterion is Adherence to Main Idea. This concerns the extent to which the written response focuses clearly on a key idea. Weakly organized responses often stray from the intended main idea. The second criterion is Sense of Beginning-Middle-End. Here, the expectation is that strong writing would have easily identifiable sections, often signalled by introductory and concluding paragraphs and sentences. Such elements are lacking in weak writing. Third, organization concerns the clarity with which ideas are presented. One idea should be addressed before another is brought up. Ideally too, different ideas should be treated in different paragraphs. Weakly organized writing treats ideas in little or no discernible order. The fourth criterion concerns sentence-to-sentence flow. In strong writing, this is logical and seamless. In contrast, weak writing may sound rambling. Finally, the rubric makes note of a special rule, that when the response consists mostly of a summary or word-for-word copying of the text, it automatically receives a score of 1 because the organization of the response is necessarily the organization of the original text, and does not reflect the student's own efforts at organization.

Table 7 Rubric for the Organization dimension of RTA

	1	2	3	4
Adherence to main idea	Strays frequently or significantly from main idea*	Attempts to adhere to the main idea*	Adheres to the main idea* (i.e., The main idea is evident throughout the response)	Focuses clearly on the main idea throughout piece* and within paragraph
Sense of beginning-middle-end	Has little or no sense of beginning, middle, and end (DIS) (i.e., Lacks topic and concluding sentence, or has no identifiable middle)	Has a limited sense of beginning, middle, and end (DIS) (i.e., Lacks a topic or concluding sentence, or has short development in middle)	Has an adequate sense of beginning, middle, and end (DIS) (topic and concluding sentences may not quite match up. Or, may be missing a beginning or ending, but organization is very clear and strong)	Has a strong sense of beginning, middle, and end (DIS) (i.e., Must have topic sentence and concluding sentence that match up and relate closely to the same key idea, and well-developed middle)
Paragraphing / Idea chunking	Has little or no order (TD)(TOP)	Attempts to address different ideas in turn+ (TD)(TOP), in different parts of the response (LCPT) (i.e., Some ideas may be repeated in different places)	Addresses different ideas in turn+ (TD)(TOP), in different parts of the response (LCPT), although multiple paragraphs may not be used (SUR)	Features multiple appropriate paragraphs (SUR), each addressing a different idea+ (TD)(TOP)
Sentence flow	May feature a rambling collection of thoughts or list-like ideas with little or no flow (LCPT)(TD)(TOP)	Has some uneven or illogical flow from sentence to sentence (LCPT) or idea to idea (TD)(TOP)	Demonstrates logical flow from sentence to sentence (LCPT) and idea to idea (TD)(TOP)	Demonstrates logical and seamless flow from sentence to sentence (LCPT) and idea to idea (TD)(TOP)

Table 7 (continued)

	1	2	3	4
Plagiarism	Consists mostly of a summary or copy of the whole text or large sections of the text (The organization of the response is necessarily the organization of the original text). In these cases, the response automatically receives a 1			

*In implementation, when annotating the score, experts and trained coders considered the coherence of the evidence in support of the author's main claim for the text. Thus, in implementation, coders placed pre-eminence on whether the evidence contributing support to the original claim formed a coherent body of evidence.

+When scoring the rubric, experts and trained coders considered whether the different ideas were presented in a logical order to evaluate how well they worked together to form coherent evidence for the main claim. The sequence of the evidence as well as how well the author elaborated different pieces of evidence, in turn, were both considered when coding.

The abbreviations in the parentheses identify the corresponding feature group discussed in the Organization modeling section of this paper that is aligned with that specific criteria

Topic-Grid and Topic-Chains

Lexical chains (Somasundaran et al. 2014) and entity grids (Burstein et al. 2010) have been used to measure lexical cohesion. In other words, these models measure the continuity of lexical meaning. Lexical chains are sequences of related words characterized by the relation between the words, as well as by their distance and density within a given span. Entity grids capture how the same word appears in a syntactic role (Subject, Object, Other) across adjacent sentences.

Intuitively, we hypothesize that these models will not perform as well on short, noisy, and low quality essays as on longer, better written essays. When the essays are short, noisy, and of low quality (i.e., limited writing proficiency), the syntactic information produced automatically by the parser may not be reliable. Moreover, even when there is elaboration on a single topic (continuation of meaning), there may not be repetition of identical or similar words. This is because words that relate to a given topic in the context of the article may not be deemed similar according to external similarity sources such as WordNet. Take, for example, the following two sentences:

"The hospitals were in bad situation. There was no electricity or water."

In the entity grid model, there would be no transition between these two sentences because there are no identical words. The semantic similarity of the nouns "hospitals" and "water" is very low and there would not be any chain including a relation between the words "hospitals", "water", and "electricity". But if we look at the source document and the topics within it, these two sentences are actually addressing a very specific sub-topic. Therefore, we think there should be a chain containing both of these words and a relation between them. Zhang et al. (2015) addresses a similar issue of capturing information from semantically related entities by leveraging world knowledge such as "**Gates** is the person who created **Microsoft**".

More importantly, what we are really interested in evaluating in this study is the organization and cohesion of pieces of evidence, not the lexical cohesion. These reasons, altogether, motivated us to design new topic-grid and topic chain models (inspired by entity-grids and lexical chains), which are more related to our rubric and may be able to overcome the issues we mentioned above.

A topic-grid is a grid that shows the presence or absence of each topic addressed in the source text (i.e., the article about poverty) in each text unit of a written response. The rows are analogous to the words in an entity-grid, except here they represent topics instead of individual words. The columns are text units. We consider the unit as a sentence or a sub-sentence (since long sentences can include more than one topic and we don't want to lose the ordering and transition information from one topic to the next). We explain how we extract the units later in this section.

To build the grids, we use the information in the source text. That is, we use the manually extracted exhaustive list of topics in Table 19 which considers the temporal aspect discussed in the article. Following this, each text unit of the essay is automatically labeled with topics using a simple window-based algorithm (with a fixed

window size = 10), which relies on the presence and absence of topic-words in a sliding window and chooses the most similar topic to the window. (Several equally similar topics might be chosen). If there are fewer than two words in common with the most similar topic, the window is annotated with no topic. We did not use spelling correction to handle topic words with spelling errors, although it is in our future plan.

The rule is that each column in the grid represents a text unit. A text unit is a sentence if it has no disjoint windows annotated with different topics or different examples from a topic. Otherwise, we break the sentence into multiple text units where each of them covers a different topic or example (the exact boundaries of the units are not important). Finally, if the labeling process annotates a single window with multiple topics, we add a column to the grid with multiple topics present in it.

See Table 8 for an example of a topic-grid for the essay with the Organization score of four in Table 1. Consider the following sentence from the essay:

> "One example they sued show a great amount oF change when they stated at first most people thall were ill just stayed in the hospital Not even getting treated either because of the cost or the hospital didnt have it, but at the end it stated they now give free medicine to most common deseases"

This sentence has two disjoint windows annotated with different topics. So, we break the sentence into two text units where they cover two different topics "Hospitals_before" and "Hospitals_after". The first part of the sentence is a unit that covers "Hospitals_before" because of a window including *"Not even getting treated"*. The second text unit covers "Hospitals_after" because of a window including *"free medicine to most common deseases"*. The third column in the grid represents the second unit of this sentence underlined which is underlined. The "x" in the third column indicates the presence of the topic "Hospital_after" which is mentioned above. The topics that are not mentioned in the essay are not included in the grid.

Then, chains are extracted from the grid. We have one chain for each topic t including both t_{before} and t_{after}. Each node in a chain carries two pieces of information:

Table 8 The topic-grid (on the left) and topic-chains (on the right) for the example essay with Organization score of 4 in Table 1

	1	2	3	4	5	6	7	8	9	10	Topic	Chain
Hospitals.b	-	x	-	-	-	-	-	-	-	-	Hospitals	(b,2),(a,3)
Hospitals.a	-	-	x	-	-	-	-	-	-	-	Education	(b,4),(a,5),(a,6)
Education.b	-	-	-	x	-	-	-	-	-	-	Farming	(b,7),(a,8)
Education.a	-	-	-	-	x	x	-	-	-	-		
Farming.b	-	-	-	-	-	-	x	-	-	-		
Farming.a	-	-	-	-	-	-	-	x	-	-		
General	x	-	-	-	-	-	-	-	x	x		

a and *b* indicate *after* and *before* respectively

the index of the text unit it appears in and whether it is a *before* or *after* state. Because transition of temporally-oriented topics are the point of interest in designing topic-chains, we ignore the topics that do not have any temporal aspect (*before* or *after* state). Examples of topic-chains are presented in Table 8. Finally, we extract several features, explained below, from the grid and the chains to represent some criteria from the rubric.

Features to Model the Rubric

As indicated above, one goal of this research in predicting Organization scores is to design a small set of rubric-based features that performs acceptably and also models what is actually important in the rubric. To this end, we designed 5 groups of features, each addressing criteria in the rubric. Some of these features are not new and have been used before to evaluate the organization and coherence of the essay; however, the features based on the topic-grid and topic-chains (inspired by entity-grids and lexical chains) are new and designed for this study. The use of *before* and *after* information to extract features is based on the rubric and the nature of the prompt.[10] Below, we explain each of the features and its relation to the rubric. Each group of features is indicated with an abbreviation that relates it to the corresponding criteria in the rubric in Table 7.

(1) Surface (SUR) captures the surface aspect of organization; it includes two features: *number of paragraphs* and *average sentence length*. Multiple paragraphs and medium-length sentences help readers follow the essays more easily.

(2) Discourse Structure (DIS) investigates the discourse elements in the essays. We cannot expect the essays written by students in grades 5-8 to have all the discourse elements mentioned in Burstein et al. (2003a), as might be expected of more sophisticated writers. Indeed, most of the essays in our corpora are short and single-paragraph (the median of # paragraphs is one). In terms of the structure, then, taking cues from the rubric, we are interested in the extent to which it has a clear beginning idea, concluding sentence, and well-developed middle. We define two binary features, *beginning* and *ending*. In the Topic-list, there is a general topic that represents general statements from the text and the prompt. If this topic is present at the beginning or at the end of the grid, the corresponding feature gets a value of 1. A third feature measures if the beginning and the ending match. We measure LSA-similarity (Landauer et al. 1998) of 1 to 3 sentences from the beginning and ending of the essay with respect to the length of the essay. The LSA is trained by the source document

[10]We hypothesize our approach can be generalized to other contrasting prompts, however, which we are about to investigate in a new dataset containing responses to a different text, but graded according to the same RTA rubric.

and the essays in the training corpus. The number of sentences are chosen based on the average essay length.

(3) Local Coherence and Paragraph Transitions (LCPT) Local coherence addresses the rubric criterion related to logical sentence-to-sentence flow. It is measured by the average LSA (Foltz et al. 1998) similarity of adjacent sentences. Paragraph transitions capture the rubric criterion of discussing different topics in different paragraphs. It is measured by the average LSA similarity of all paragraphs (Foltz et al. 1998). For an essay where each paragraph addresses a different topic, the LSA similarity of paragraphs should be less than for an essay in which the same topic appears in different paragraphs. For one paragraph essays, we divide the essays into 3 equal parts and calculate the similarity of 3 parts.

The average LSA similarity of text units (sentences or paragraphs) are calculated as follows: A semantic space was constructed based on the essays in the training set. The vector for each text unit is computed (as the weighted sum of its weighted terms) and then is compared to the vector for the adjoining text unit by cosine similarity measure. The average LSA similarity is then calculated for each text by averaging these cosines between the vectors for all pairs of adjoining text units.

(4) Topic Development (TD) Good essays should have a developed middle relevant to the assigned prompt. The following features are designed to capture how well-developed an essay is:

- *Topic-Density*: Number of topics covered in the essay divided by the length of the essay. Higher Density means less development on each topic.
- *Before-only, After-only* (i.e., Before and after the UN-led intervention referenced in the source text): These are two binary features. It measures if all the sentences in the essay are labeled only with "before" or only with "after" topics. A weak essay might, for example, discuss at length the condition of Kenya before the intervention (i.e., address several "before" topics) without referencing the result of the intervention (i.e., "after" topics).
- *Discourse markers*: Four features that count the discourse markers from each of the four groups: contingency, expansion, comparison, and temporal, extracted by "AddDiscourse" connective tagger (Pitler and Nenkova 2009). Eight additional features represent count and percentage of discourse markers from each of the four groups that appear in sentences that are labeled with a topic.
- *Average chain size*: Average number of nodes in chains. Longer chains indicate more development on each topic.
- *Number and percentage of chains with variety*: A chain on a topic has variety if it discusses both aspects ('before' and 'after') of that topic.

(5) Topic Ordering and Patterns (TOP) It is not just the number of topics and the amount of development on each topic that is important. More important is how students organized these topics in their essays. Logical and strategic organization of

topics helps to strengthen arguments. Meanwhile, as reflected in the rubric in Table 7, little or no order in the discussion of topics in the essay means poor organization. Here we present the features we designed to assess the quality of the essays in terms of organization of topics.

- *Levenshtein edit-distance* of the topic vector representations for "befores" and "afters", normalized by the number of topics in the essay. If the essay has a good organization of topics, it should cover both the *before* and the *after* examples on each discussed topic. It is also important that they come in a similar order. For example, suppose the following two vectors represent the order of topics in an essay: befores=[3,4,4,5] , afters=[3,6,5]. First we compress the vectors by combining the adjacent similar topics. In this example topic number 4 will be compressed. So the final vectors are: befores=[3,4,5] , afters=[3,6,5]. The normalized Levensthein between these two vectors is 1/4, which shows the number of edits required to change one number string into the other normalized by total number of topics in the two vectors. The greater the value, the worse the pattern of discussed topics.
- *Max distance between chain's nodes*: Large distance can be a sign of repetition. The distance between two nodes is the number of text units between those nodes in the grid.
- *Number of chains starting and ending inside another chain*: There should be fewer in well-organized essays.
- *Average chain length (Normalized)*: The length of the chain is the sum of the distances between each pair of adjacent nodes. The normalized feature is divided by the length of the essay.
- *Average chain density*: Equal to average chain size divided by average chain length.
- *Topic transition probability*: Transition probabilities are the proportions of topic transition types within a text. Transition types include {- -, -X, X-, XX}.

The value of the features for the example essay with Organization score of 4 in Table 1 are shown in Table 9.

Based on the defined features, we imagine generating feedback that helps students address criteria that received low scores. For example, if the value of the discourse structure feature *beginning* is false, the system could remind students to write a clear introductory sentence where they tell the reader whether or not they believe that the author provides a convincing argument that "wining the fight against poverty is achievable in our lifetime."

Experiments and Results

Experimental Setup

We configure a series of experiments to test the validity of three hypotheses for the two dimensions. These hypotheses are designed to validate the usefulness of the

Table 9 Feature vector representation of the high-scoring Organization essay from Table 1

Feature Set	Specific Feature	Value
Surface (SUR)	Number of paragraphs	3
	Average sentence length	15.75
Discourse Structure (DIS)	Beginning	1
	Ending	1
	Similarity of beginning and ending	0.41
Local Coherence and Paragraph Transitions (LCPT)	LSA sentence similarity	0.30
	LSA paragraph similarity	0.40
Topic Development (TD)	Topic density	0.01
	Before-only	0
	After-only	0
	Contingency	5
	Expansion	2
	Comparison	10
	Temporal	2
	ContigencyWithEvidence	5
	ContigencyWithEvidence%	1
	ExpansionWithEvidence	2
	ExpansionWithEvidence%	1
	ComparisonWithEvidence	10
	ComparisonWithEvidence%	1
	TemporalWithEvidence	2
	TemporalWithEvidence%	1
	Average topic chain size	2.5
	Chains with variety	1
	Chains with variety%	0.5
Topic Ordering and Patterns (TOP)	Levenshtein before/after distance	0.33
	Max chain distance	0.004
	Starting inside chain	0.25
	Ending inside chain	0
	Average chain length	1.5
	Average chain length normalized	0.01
	Average chain density	0.83
	- -	0.63
	-X	0.13
	X-	0.17
	XX	0.07

model in terms of performance, generalizability of the model across different grades, and the utility of the rubrics for designing predictive features:

H1: the rubric-based models can match or even outperform competitive baselines,[11]
H2: the rubric-based models generalize better across students from different grades[12] (i.e., across our two datasets), and
H3: the more that cells in the rubric are covered by a feature group, the more predictive utility the feature group will have in isolation.

We use our two datasets in three different ways: 1) cross validation on each dataset (to test H1 and H3), 2) combining the two datasets to a one big dataset of grades 5–8 and performing cross validation (also to test H1 and H3), 3) training the model on one dataset and testing on the other one (to test H2). The motivation behind combining the two corpora to one bigger dataset is the fact that in a small pilot study (as part of an additional experiment on analyzing the impact of training sample size on the reliability of AES), we found that each doubling of the RTA training sample size increased Quadratic Weighted Kappa by .03.

For all experiments we use 10 runs of 10 fold cross validation using Random Forest as a classifier (max-depth=5). We also tried some other classification and regression methods, such as Naive Bayes, logistic regression and gradient boosting regression, and all the conclusions remained the same. Since our dataset is imbalanced, we use SMOTE (Chawla et al. 2002) oversampling method. This method involves creating synthetic minority class examples. This algorithm generates synthetic examples by operating in feature space. The minority class is oversampled by taking each minority class sample, randomly choosing neighbors from the k nearest neighbors of it, and introducing synthetic examples along the line segments joining any/all of these nearest neighbors. We only oversampled the training data, not the testing data.

All performance measures are calculated by comparing the 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. We report the performance as Quadratic Weighted Kappa, which is a standard evaluation measure for essay assessment systems. We use corrected paired t-test (Bouckaert and Frank 2004) to measure the significance of any difference in performance.

Baselines for Evidence

As a baseline we choose a unigram model. Unigrams are extracted and filtered down to the top 500 features by the chi-squared statistic, then a Random Forest model

[11] There is no overlap between Evidence rubric-based features and the baseline. For Organization, some of the features have similar definitions as the baseline features but they are derived from different sources (topic-grid and topic-chain versus entity-grid and lexical chain).

[12] We expect the rubric-based features generalize better across grades because they represent the rubric which is the same for all grades. But the language of the students might vary as they develop.

is trained on the resulting feature set. We choose this baseline based on the results represented in Rahimi et al. (2014) which shows unigram model is a well-performing baseline.

Baselines for Organization

We use two well-performing baselines from recent methods to evaluate organization and coherence of the essays. The first baseline (EntityGridTT) is based on the entity-grid coherence model introduced by Barzilay and Lapata (Barzilay and Lapata 2005). This method has been used to measure the coherence of student essays (Burstein et al. 2010). It includes transition probabilities and type/token ratios for each syntactic role as features. We perform a set of experiments to find the best configuration.[13] We therefore use this best configuration in all experiments. It should be noted that this works to the advantage of the entity-grid baseline since we do not have parameter tuning for the other models.

The second baseline (LEX1) is a set of features extracted from Lexical Chaining (Morris and Hirst 1991). We use Galley and McKeown (Galley and Mckeown 2003) lexical chaining and extract the first set of features (LEX1) introduced in Somasundaran et al. (2014). We do not implement the second set because we do not have the annotation or the tagger to tag discourse cues.

Results and Discussion

Evidence

We first examine the hypothesis that our new features will outperform or at least perform equally well as the baselines (H1). Our results support this hypothesis. Run 2 in Table 10 shows that the rubric-based model yields higher performance than the unigram baseline on all three datasets although it is not significantly higher on (5–6) dataset. Comparing Run 4 with Run 1 and 2 shows that adding unigrams to our rubric-based model does not improve our results but adding the rubric-based features to the unigram model improves the performance.This shows that the rubric based model has information which is not captured in the unigram model and also the rubric-based model captures much of what is already captured by the unigram baseline. The reason is that the NPE and Specificity features are designed to look for existence and co-occurrence of the important unigrams. Runs 3 and 5 investigate the performance of our model without the fallback Word-count feature. The results show that our model still outperforms the unigram baseline although not significantly and adding the rubric-based features (except word count) improves the unigram baseline significantly on two of the datasets.

Looking at the features that were selected in our feature selection phase in Runs 4 and 5 shows that the rubric-based features: NPE, CON, and most of the Specificity

[13] We find that the best model is an entity-grid model with history=2, salience=1, syntax=on and type/token ratios.

Table 10 Cross-validated performance of our rubric-based Evidence model compared to the baseline on both datasets and a combination of the two datasets (5–8)

	Model	5–6 [n=1569]	6–8 [n=800]	5–8 [n=2369]
1	Unigram	0.62	0.56	0.59
2	Rubric-based	**0.64**	**0.62** (1)	**0.64** (1,3)
3	Rubric-based -WOC	0.62	0.59	0.62
4	Unigram+Rubric-based	**0.64**	0.61 (1)	**0.64** (1,3)
5	Unigram+Rubric-based-WOC	**0.64**	0.60 (1)	0.62 (1)

The numbers in parenthesis show the model numbers which the current model performs significantly better than. The numbers in brackets show the size of the datasets in use. Bolded numbers are the best results in each column

features are always among the best 500 selected features for all three datasets. Looking at the confusion matrix of the rubric-based model for grades 5–8, we notice that our model performs the best on score 1 (F1 = 0.68) and the worst on score 3 (F1 = 0.38). The F1 is equal to 0.52 and 0.47 for scores 2 and 4 respectively. The F1 values are similar for all three datasets.

We configured another experiment to examine the generalizability of the models across different grades (H2). In this experiment, we used one dataset for model training and the other for testing. We divided the test data into 10 disjoint sets to be able to perform significance tests on the performance measure.[14] The results in Table 11 show that for both experiments, the rubric-based model performs significantly better than the baseline, which supports the findings from the cross-validation experiment and hypothesis (H2). Comparing the Quadratic Weighted Kappa figures across columns, the Models 1 and 3 which include the unigram features perform better when the training size is bigger. The rubric-based model performs comparably even when we train on the smaller 6–8 dataset and test on the noisier 5–6 corpus. These results suggest that our features are more robust to both lack of training data and training/test set differences.

Finally, our last hypothesis is that although each rubric-based feature group should be capturing useful information, the feature group designed to capture information about specific pieces of evidence that covers more cells in the rubric is the most important one. To test this hypothesis, we performed an experiment using each of the isolated groups of features. The results in Table 12 show that Specificity is the most predictive feature group in isolation. Specificity alone also almost matches or outperforms the unigram baseline, and approaches the performance of the full rubric-based model. The NPE rubric-based feature is also consistently more predictive than either the CON feature or the word count feature which is not based on the rubric at all.

[14]This experiment does not include cross-validation. There is only one training set. The 10 disjoint sets of test data is to be able to perform significance testing.

Table 11 Performance of our rubric-based Evidence model compared to the baselines

	Model	Train(5–6) [n=1569] Test(6–8)[n=800]	Train(6–8) [n=800] Test(5–6) [n=1569]
1	Unigram	0.58	0.46
2	Rubric-based	**0.61** (1)	**0.62** (1,3)
3	Unigram+Rubric-based	**0.61** (1)	0.56 (1)

Each time, we train the models on one dataset and test on the other. The numbers in parenthesis show the model numbers which the current model performs significantly better than. The numbers in brackets show the size of the datasets in use

To further investigate the effect of the word count feature, we perform an experiment in which we compare the performance of the rubric-based model with the same model after removing the word count to predict the scores for 3 different data subsets defined by Evidence scores: 1) essays rated as 1 and 2; 2) essays rated as 1, 2 or 3; and 3) essays rated as 3 and 4. The results are in Table 13. As can be seen, including word count only significantly improves performance for the data subset of [1,2,3] and [1,2,3,4], meaning it is useful to discriminate essays with score 3 and 4 from essays with score 1 and 2. Recall that our rubric-based features most sparsely cover the rows in the score 4 column of the Evidence rubric, where we would thus expect word count to play a fallback role.

In sum, our rubric-based model is advantageous to the unigram baseline because the rubric-based model yields higher performance than the unigram baseline on all three datasets although it is not significantly higher on (5–6) dataset; the rubric based model has information which is not captured in the unigram model; and our features are more robust to both lack of training data and training/test set differences. As we hypothesized, the Specificity feature designed to capture information about specific pieces of evidence that covers more cells in the rubric is the most important feature. Word count is useful to discriminate essays with score 3 and 4 from essays with score 1 and 2 (recall that we most sparsely cover the rows in the score 4 column of the Evidence rubric, where we would thus expect word count to play a fallback role).

Table 12 Cross-validated performance evaluation of Evidence feature groups in isolation on the two datasets and their combination

Method	5–6 [n=1569]	6–8 [n=800]	5–8 [n=2369]
Rubric-based	**0.64**	**0.62**	**0.64**
NPE	0.51	0.48	0.53
CON	0.36	0.39	0.37
SPC	0.61	0.60	0.61
WOC	0.39	0.31	0.38

The numbers in brackets show the size of the dataset in use

Table 13 Cross-validated performance evaluation of the word count feature

Scores	Features	5–6	6–8	5–8
1,2	Rubric-based	0.49	0.49	0.50
	Rubric-based minus WOC	0.49	0.49	0.50
1,2,3	Rubric-based	0.60	0.54	0.59*
	Rubric-based minus WOC	0.58	0.54	0.57
3,4	Rubric-based	0.26	0.27	0.30
	Rubric-based minus WOC	0.24	0.21	0.26
1,2,3,4	Rubric-based	0.64*	0.61	0.64*
	Rubric-based minus WOC	0.62	0.60	0.62

Significant improvements when including word count are marked by * ($p < 0.05$)

Organization

We first examine the hypothesis that the new features perform comparably or even better than the baselines (H1). The results on the corpus of grades 5–6 (see Table 14) show that the new features (Model 4) yield significantly higher performance than either baseline (Models 1 and 2) or the combination of the baselines (Model 3). The results of Models 5, 6, and 7 show that our new features capture information that is not in the baseline models (since each of these three models is significantly better than models 1, 2, and 3 respectively), but that the baseline features provide no value when added to the rubric-based features (since none of these three models is better than model 4). The best result in all experiments is bolded.

We repeated the experiments on the corpus of grades 6-8. The results in Table 14 show that there is no significant difference between the rubric-based model and the

Table 14 Cross-validated performance of our rubric-based Organization model compared to the baselines on both datasets and their combination

	Model	5–6 [n=1569]	6–8 [n=809]	5–8 [n=2378]
1	EntityGridTT	0.42	0.49	0.48
2	LEX1	0.45	0.53 (1)	0.51 (1)
3	EntityGridTT+LEX1	0.46 (1)	0.54 (1)	0.52 (1,2)
4	Rubric-based	**0.51** (1,2,3)	0.51	0.54 (1,2)
5	EntityGridTT+Rubric-based	0.49 (1,2,3)	0.53 (1)	0.53 (1,2)
6	LEX1+Rubric-based	**0.51** (1,2,3)	0.55 (1)	**0.55** (1,2,3,5)
7	EntityGridTT+LEX1 +Rubric-based	0.50 (1,2,3)	**0.56** (1)	**0.55** (1,2,3)

The numbers in parenthesis show the model numbers which the current model performs significantly better than. The numbers in brackets show the size of the dataset in use

baselines, except that in general, models that include lexical chaining features perform better than those with entity-grid features. Although not significant, the best result comes from adding the rubric-based features to the baseline features (Model 7).

The experiments on the combination of the two datasets show that our rubric-based model yields significantly higher performance than either baseline (Models 1 and 2) and is comparable to the combination of the baselines (Model 3). The results of Models 5, 6, and 7 show that our new features capture information that is not in the baseline models since each of these three models is significantly better than models 1, 2, and 3 respectively. The final conclusion is that the first hypothesis that the new features perform comparably or even better than the baselines is supported by the results.

Comparing the columns for Models 1, 2, 3, and 4 shows that the baseline models perform better on 6–8 dataset that has higher quality essays compared to 5–6 corpus even though the size of 6–8 dataset is smaller. But, the rubric-based model performs the same on both 6–8 and noisier 5–6 datasets and the performance increases when we combine the two corpora.

Looking at the confusion matrices of all three datasets, the performance of our rubric-based model is the worst on score 3. The F1 is equal to 0.497, 0.504, 0.369, and 0.487 for scores 1 to 4 respectively on 5–8 dataset. The F1 values are similar for all three datasets.

We configured another experiment to examine the generalizability of the models (Hypothesis H2) across different grades. In this experiment, we used one dataset for model training and the other for testing. We divided the test data into 10 disjoint sets to be able to perform significance tests on the performance measure. The results in Table 15 show that for both experiments, the rubric-based model performs at least as well as the baselines. Where the training is on grades 6-8 and we test the model on the shorter and noisier set of 5-6, the rubric-based model performs significantly better than the baselines. Where we test on the 6-8 corpus, the rubric-based model performs better than the baselines (although not always significantly), and adding it to the baselines (Model 5) adds value to them significantly. Comparing the columns, all the models perform better when the training size is bigger.

Table 15 Performance of our rubric-based Organization model compared to the baselines

	Model	Train(5–6) [n=1569] Test(6–8) [n=809]	Train(6–8) [n=809] Test(5–6) [n=1569]
1	EntityGridTT	0.51 (2)	0.43
2	LEX1	0.43	0.41
3	EntityGridTT+LEX1	0.52 (2)	0.42
4	Rubric-based	0.56 (2)	**0.47** (1,2,3)
5	EntityGridTT+LEX1 +Rubric-based	**0.58** (2,3,1)	0.45

Each time, we train the models on one dataset and test on the other. The numbers in parenthesis show the model numbers which the current model performs significantly better than. The numbers in brackets show the size of the dataset in use

Table 16 Cross-validated performance of Organization feature groups in isolation. The numbers in brackets show the size of the dataset in use

	Model	5–6 [n=1569]	6–8 [n=809]	5–8 [n=2378]
1	TopicDevelopment	0.40	0.42	0.43
2	TopicOrdering	0.40	0.43	0.44
3	TopicDevelopment+TopicOrdering	**0.42**	**0.45**	**0.46**
4	Surface	0.32	0.40	0.39
5	LocalCoherence+ParagraphTransition	0.20	0.21	0.21
6	DiscourseStrucutre	0.25	0.19	0.27

As for our last hypothesis, we investigate the effect of rubric-based features in isolation. To do so, we repeated the cross-validated experiments using each of the isolated groups of features. The results in Table 16 show that Topic-Development and Topic-Ordering are the most predictive set of features. This result supports the hypothesis since these two feature groups cover more cells in the rubric. While the topic-based features may not be better than the baselines, they can be improved. One potential improvement is to enhance the alignment of the sentences with their corresponding topics (since we currently use a very simple model for alignment). Moreover, we believe that the topic ordering features are more substantive and potentially provide more useful information for students and teachers in downstream applications such as providing feedback and analytics.

In sum, our rubric-based model has some advantages to the baseline models. First, it yields either significantly higher performance than baselines or comparable to them. Second, our new features capture information that is not in the baseline models. Third, the rubric-based model performs the same on both 6–8 and noisier 5–6 datasets while the baseline models perform better on 6–8 dataset that has higher quality essays compared to 5–6 corpus. Finally, the rubric-based model is tied to the rubric of the construct. Moreover, in cross-dataset experiments, the rubric-based model performs at least as well as the baselines but all the models perform better when the training size is bigger. Topic-Development and Topic-Ordering, which cover more cells in the rubric, are the most predictive sets of features.

Conclusion and Future Work

In this study, we attempt to measure two targeted constructs within analytic text-based writing: 1) students' effective use of evidence and, 2) their organization of ideas and evidence in support of their claim. We present the results for predicting the score of the Evidence and Organization dimensions of a response-to-text assessment in a way that aligns with the scoring rubric. We used two datasets of essays written by students in grades 5–6 and 6–8. We designed a set of features aligned with the rubric that we believe will be meaningful and easy to interpret given the writing task. Our experimental results show that our task-dependent model (consistent with the rubric)

performs as well as if not outperforms the baselines. We also show the potential generalizability of the rubric-based model by performing cross-corpus experiments. Finally, we show that the more a designed feature group covers criteria in the rubric, the more predictive utility the feature group generally has. In sum, our set of results thus provides support for all three of the hypotheses motivating our experiments.

There are several ways to improve our work. First, we plan to use a more sophisticated method to annotate text units, such as information retrieval or sentence similarity based approaches. Currently we are using a simple window-based algorithm that looks for word overlaps. In the future, we will incorporate more sophisticated methods such as text similarity approaches based on word-embedding representations. Second, we are working towards replacing manually extracted topics and examples by automatically extracted ones, as our current approach requires these to be manually defined by experts (although this task needs to be only done once for each new text and prompt). We proposed (Rahimi and Litman 2016) to use a data-driven model enabled by LDA topic modeling to automatically extract the topical components (i.e., topic words and significant N-grams ($N \geq 1$) as examples for each topic) needed for our scoring approach. Our preliminary results are promising. Third, we will design and validate a system for providing automated formative feedback to students on their responses. We will investigate the extent to which our automated essay scoring system serves this purpose. Specifically, we will build on our research to study the influence of the formative feedback generated by the AES system on the quality of students' writing and teachers' instruction. Fourth, we need to develop additional features to fully operationalize both the Evidence and Organization rubrics. Next, we have a new dataset from a second prompt which we will use to further test the generalizability of our model. We hypothesize the same approach works on the data for the new prompt. Finally, we need to tune all our parameters that were chosen intuitively or were set to the default value.

Acknowledgments This work was supported by the Learning Research and Development Center at the University of Pittsburgh.

Appendix

Table 17 The main topics (and associated words) used to calculate NPE for the Time for Kids source text

1.	**Hospitals**: care, health, hospital, treatment, doctor, electricity, disease, water, sick, medicine, generator, no, die, kid, bed, patient, clinical, officer, running
2.	**Malaria**: bed, net, malaria, infect, bednet, mosquito, bug, sleeping, die, cheap, infect, biting
3.	**Farming**: farmer, fertilizer, irrigation, dying, crop, seed, water, harvest, hungry, feed, food, irrigation
4.	**School**: school, supplies, fee, student, midday, meal, lunch, supply, book, paper, pencil, energy, free, children, kid, go, attend

These are the four topics among the eight (see next table) that are considered major and are expected to be mentioned in the essays. The word "no" is correlated with topic "Health" since it is used in many examples of it

Table 18 Word lists for the specific examples associated with each topic used to calculate SPC for the Time for Kids source text

Topic1	**Topic2**	**Topic3**	**Topic4**
• unpaved roads	• United Nations intervention	• Yala sub district Hospital	• Malaria common disease preventable
• tattered clothing	• safer healthier better life	• three kids bed two adults rooms packed patients	• mosquitoes carry malaria infect people biting
• bare feet	• out poverty stabilize economy quality life communities	•not medicine treatment could afford	• kids die malaria adults sick 20 000 day
• less than 1 dollar day	• Africa Kenya Sauri	• no doctor only clinical officer running hospital	• bed nets mosquitoes away people save millions lives
	• goals met 2015 2025	• no running water electricity	• bed nets cost 5 $ dollar
	• 80 villages across sub-Saharan Africa	•sad people dying near death preventable	•cheap medicines treat malaria
Topic 5	**Topic 6**	**Topic 7**	**Topic 8 (General)**
• crops dying	•kids not attend go school	• progress just four years	• progress encouraging supporters
• not afford fertilizer irrigation	• not afford school fees	• Yala sub district hospital has medicine	• solutions problems keep people impoverished
• outcome poor crops	• kids help chores fetching water wood	• medicine free charge	• change poverty stricken areas good
• lack fertilizer water	• schools minimal supplies books paper pencils	• medicine most common diseases	• poverty history not easy task hard
• enough food crops harvest feed whole family hungry sick	• concentrate not energy	• water connected hospital	• winning against poverty possible achievable lifetime
	• no midday meal lunch	• hospital generator electricity	
		• bed nets used every sleeping site	
		• hunger crisis addressed fertilizer seeds	
		• tools needed maintain food supply	
		• kids go school now	
		• no school fees	
		• now serves lunch students	
		• school attendance rate way up	

Table 19 Word lists for the specific examples associated with each topic from the Time for Kids source text used to calculate Topic-grids and Topic-chains

Topic1	**Topic2**	**Topic3_before (Hospital)**	**Topic4_before (Malaria)**
• unpaved roads	• United Nations intervention	• Yala sub district Hospital	•Malaria common disease preventable treatable
• tattered clothing	• safer healthier better life	• three kids bed two adults rooms packed patients	•mosquitoes carry malaria infect people biting
• bare feet	• out poverty stabilize economy quality life communities	•not medicine treatment could afford	• kids die malaria adults sick 20 000 day
• less than 1 dollar day	• Africa Kenya Sauri	• no doctor only clinical officer running hospital	• bed nets mosquitoes away people save millions lives
	• goals met 2015 2025	• no running water electricity	• bed nets cost 5 $ dollar
	• 80 villages across sub-Saharan Africa	•sad people dying near death preventable	• cheap medicines treat malaria
Topic5_before (Farming)	**Topic6_before (Education)**	**Topic3_after (Hospital)**	**Topic8 (General)**
• crops dying	•kids not attend go school	• Yala sub district hospital has medicine	• progress encouraging supporters
• not afford fertilizer irrigation	• not afford school fees	• medicine free charge	•solutions problems keep people impoverished
• outcome poor crops	• kids help chores fetching water wood	• medicine most common diseases	• change poverty stricken areas good
• lack fertilizer water	• schools minimal supplies books paper pencils	• water connected hospital	• poverty history not easy task hard

Table 19 (continued)

• enough food crops harvest feed whole family hungry sick	• concentrate not energy	• hospital generator electricity	• winning against poverty possible achievable life-time
	• no midday meal lunch		
Topic4_after (Malaria)	**Topic5_after (Farming)**	**Topic6_after (Education)**	
• bed nets used every sleeping site	• hunger crisis addressed fertilizer seeds	• kids go school now	
	• tools needed maintain food supply	• no school fees	
		• now serves lunch students	
		• school attendance rate way up	

Some of the topics can be given intuitive names based on the including examples. The intuitive names are mentioned in parenthesis

References

Attali, Y. (2011). A differential word use measure for content analysis in automated essay scoring. *ETS Research Report Series*, *2011*(2), i–19.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2. *The Journal of Technology, Learning and Assessment*, *4*(3), 1–29.

Attali, Y., & Powers, D. (2008). A developmental writing scale. Wiley Online Library. *ETS Research Report Series RR-08-19*, *2008*(1). http://www.ets.org/Media/Research/pdf/RR-08-19.pdf.

Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, *30*(1), 125–141. Sage Publications Sage UK: London, England.

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, *29*(3), 371–383.

Barzilay, R., & Lapata, M. (2005). Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05* (pp. 141–148).

Bouckaert, R.R., & Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In *Advances in knowledge discovery and data mining* (pp. 3–12).

Burstein, J., Kukich, K., Wolff, S., Chi, L., & Chodorow, M. (1998). Enriching automated essay scoring using discourse marking. In *Proceedings of the Workshop on Discourse Relations and Discourse Marking Annual Meeting of the Association of Computational Linguistics*.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Chi, L., Nolan, J., Rock, D., & Wolff, S. (1999). Computer analysis of essay content for automated score prediction. TOEFL Monograph Series Report No. 13.

Burstein, J., Chodorow, M., & Leacock, C. (2003a). Criterion sm : Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the 15th Annual Conference on Innovative Applications of Artificial Intelligence*.

Burstein, J., Marcu, D., & Knight, K. (2003b). Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, *18*(1), 32–39.

Burstein, J., Tetreault, J., & Andreyev, S. (2010). Using entity-based features to model coherence in student essays. In *Human Language Technologies The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10* (pp. 681–684).

Burstein, J., Tetreault, J., & Chodorow, M. (2013). Holistic discourse coherence annotation for noisy essay writing. *Dialogue & Discourse*, *4*(2), 34–52.

Chawla, N.V., Bowyer, K.W., Hall, L.O., & Kegelmeyer, P.W. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Cohen, R. (1987). Analyzing the structure of argumentative discourse. *Computational linguistics*, *13*(1-2), 11–24.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*(1), 100–108.

Correnti, R., Matsumura, L.C., Hamilton, L., & Wang, E. (2013). Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal*, *114*(2), 142–177. JSTOR.

Crossley, S.A., Varner, L.K., Roscoe, R.D., & McNamara, D.S. (2013). Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In *International Conference on Artificial Intelligence in Education* (pp. 269–278) Springer.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. Elsevier. *Assessing Writing*, *18*(1), 7–24.

Deane, P., Williams, F., Weng, V., & Trapani, C.S. (2013). Automated essay scoring in innovative assessments of writing from sources. *Writing Assessment*, *6*(1), 40–56.

Elliot, S. (2003). Intellimetric: from here to validity. In M.D. Shermis & J.Burstein (Eds.) Automated Essay Scoring: A Cross Disciplinary Perspective, (pp. 71–86). Hillsdale, NJ: Lawrence Erlbaum Associates.

Foltz, P.W., Kintsch, W., & Landauer, T.K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, *25*(2-3), 285–307.

Galley, M., & Mckeown, K. (2003). Improving word sense disambiguation in lexical chaining. In *Proceedings of IJCAI* (pp. 1486–1488).

Grosz, B.J., Weinstein, S., & Joshi, A.K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, *21*(2), 203–225.

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). Evaluating multiple aspects of coherence in student essays. In *HLT-NAACL* (pp. 185–192).

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*(02), 145–159.

Kakkonen, T., Myller, N., Timonen, J., & Sutinen, E. (2005). Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the 2nd workshop on Building Educational Applications Using NLP* (pp. 29–36). Association for Computational Linguistics.

Klebanov, B.B., & Higgins, D. (2012). Measuring the use of factual information in test-taker essays. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP* (pp. 63–72): Association for Computational Linguistics.

Klebanov, B.B., Madnani, N., Burstein, J., & Somasundaran, S. (2014). Content importance models for scoring writing from sources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 247–252). Baltimore, Maryland: Association for Computational Linguistics.

Landauer, T.K., Foltz, P.W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284.

Lee, Y.-W., Gentile, C., & Kantor, R. (2008). Analytic scoring of toefl® cbt essays: Scores from humans and e-rater®. *ETS Research Report Series*, *2008*(1), i–71.

Lemaire, B., & Dessus, P. (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research*, *24*(3), 305–320.

Liu, L.O., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M.C. (2014). Automated scoring of constructed-response science items Prospects and obstacles. *Educational Measurement: Issues and Practice*, *33*(2), 19–28.

Louis, A., & Higgins, D. (2010). Off-topic essay detection using short prompt texts. In *Proceedings of the NAACL HLT 5th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 92–95). Association for Computational Linguistics.

Loukina, A., Zechner, K., Chen, L., & Heilman, M. (2015). Feature selection for automated speech scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 12–19).

Miltsakaki, E., & Kukich, K. (2000). Automated evaluation of coherence in student essays. In *Proceedings of LREC 2000*.

Morris, J., & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, *17*(1), 21–48.

Ong, N., Litman, D., & Brusilovsky, A. (2014). Ontology-based argument mining and automatic essay scoring. In *Proceedings of the 1st Workshop on Argumentation Mining* (pp. 24–28).

Page, E.B. (2003). Project Essay Grade: PEG In M.D. Shermis & J. Burstein (Eds.) *Automated Essay Scoring: A Cross Disciplinary Perspective.* (pp. 43–54). Hillsdale, NJ: Lawrence Erlbaum Associates.

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In Bazerman, C., Dean, C., Early, J., Lunsford, K., Null, S., Rogers, P., & Stansell, A. (Eds.) *International advances in writing research: Cultures, places, measures* (pp. 121–131). WAC Clearinghouse/Parlor Press Fort Collins, Colorado/Anderson, SC.

Perelman, L. (2013). Critique of Mark D. Shermis & Ben Hammer, Contrasting State-of-the-Art Automated Scoring of Essays: Analysis. *Journal of Writing Assessment*, *6*(1). http://journalofwritingassessment.org/article.php?article=69.

Persing, I., & Ng, V. (2014). Modeling prompt adherence in student essays. In *ACL (1)* (pp. 1534–1543).

Persing, I., & Ng, V. (2015). Modeling argument strength in student essays. In *Proceedings of ACL*.

Pitler, E., & Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers* (pp. 13–16).

Rahimi, Z., & Litman, D. (2016). Automatically extracting topical components for a response-to-text writing assessment. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*.

Rahimi, Z., Litman, D., Correnti, R., Matsumura, L.C., Wang, E., & Kisa, Z. (2014). Automatic Scoring of an Analytical Response-To-Text Assessment. In *Intelligent Tutoring Systems*, Springer (pp. 601–610). doi:10.1007/978-3-319-07221-0_76.

Rahimi, Z., Litman, D., Wang, E., & Correnti, R. (2015). Incorporating coherence of topics as a criterion in automatic response-to-text assessment of the organization of writing. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 20–30).

Scott, C.A., & McNamara, D.S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1236–1241).

Shermis, M.D., & Burstein, J. (2003). Automated essay scoring: A cross-disciplinary perspective, Routledge.

Shermis, M.D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. In *Annual national council on measurement in education meeting* (pp. 14–16).

Somasundaran, S., Burstein, J., & Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *COLING* (pp. 950–961).

Song, Y., Heilman, M., Beigman, B., & Deane, K.P. (2014). Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining, ACL 2014* (pp. 69–78): Citeseer.

Stab, C., & Gurevych, I. (2014a). Annotating argument components and relations in persuasive essays. In *COLING* (pp. 1501–1510).

Stab, C., & Gurevych, I. (2014b). Identifying argumentative discourse structures in persuasive essays. In *EMNLP* (pp. 46–56).

Weigle, S.C. (2002). *Assessing writing*. New York: Cambridge University Press.

Xie, S., Evanini, K., & Zechner, K. (2012). Exploring content features for automated speech scoring. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (p. 2012): Association for Computational Linguistics.

Zhang, M., Feng, V.W., Qin, B., Hirst, G., Liu, T., & Huang, J. (2015). Encoding world knowledge in the evaluation of local coherence. In *NAACL*.