Incorporating Coherence of Topics as a Criterion in Automatic Response-to-Text Assessment of the Organization of Writing

Zahra Rahimi[1], **Diane Litman**[1,2,3], **Elaine Wang**[3], **Richard Correnti**[3]
[1]Intelligent Systems Program, [2]Department of Computer Science
[3]Learning Research and Development Center
University of Pittsburgh
Pittsburgh, PA 15260
{zar10,dlitman,elw51,rcorrent}@pitt.edu

Abstract

This paper presents an investigation of score prediction for the Organization dimension of an assessment of analytical writing in response to text. With the long-term goal of producing feedback for students and teachers, we designed a task-dependent model that aligns with the scoring rubric and makes use of the source material. Our experimental results show that our rubric-based model performs as well as baselines on datasets from grades 6-8. On shorter and noisier essays from grades 5-6, the rubric-based model performs better than the baselines. Further, we show that the baseline model (lexical chaining) can be improved if we extend it with information from the source text for shorter and noisier data.

1 Introduction

As a construct, 'Organization' has figured in systems for scoring student writing for decades. On the NAEP (National Assessment of Educational Progress), the organization of the text, coherence, and focus are judged in relation to the writer's purpose and audience (National Assessment Governing Board, 2010) to determine a single holistic score. Alternatively, 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[1]. Assessments aligned to the Common Core State Standards (CCSS), the academic standards adopted 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[2]. The consideration of 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 Crossley 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 coher-

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

[2]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

Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015, pages 20–30,
Denver, Colorado, June 4, 2015.

ence (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; Burstein et al., 2003b).

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 of task-dependent scores is the ability to provide feedback that is better aligned with the task.

One of the types of writing emphasized in the CCSS is writing in response to text (Correnti et al., 2013). In as early as the fourth and fifth grades, students are expected to write analytical responses to text, which involves making claims and marshalling evidence from a source text to support a viewpoint.

The Response-to-Text Assessment (RTA) (Correnti et al., 2013; Correnti et al., 2012) was developed for research purposes to study upper-elementary students' text-based writing skills. The RTA is evaluated with a five-trait rubric. Efforts to automate the assessment of student responses have been underway to support scaling up the use of the RTA in research and also to explore the potential of providing feedback on student writing to teachers. Specifically, evaluation of the Evidence dimension is investigated in Rahimi et al. (2014). In the present study, we aim to design a model to evaluate the Organization dimension of the RTA.

Our study differs in three noteworthy ways from previous studies aiming to evaluate organization. Insofar as the Organization dimension of the RTA concerns the coherence of the essay, this is similar to previous investigations that operationalize this trait as adherence to the essay topic, sentence-to-sentence flow, and logical paragraph transitions. Specifically, however, Organization as conceived by the RTA also concerns how well the pieces of evidence provided from the text are organized to make a strong argument. In this sense, what matters is coherence around the ordering of pieces of evidence.

This additional aspect of Organization is important to the evaluation of the RTA and to text-based writing in general; yet, available models for assessing coherence do not capture this aspect, primarily because Organization has been treated largely as task-independent. As such, these models are insufficient for our purposes, even if they might perform well on score prediction. For our study, then, we set out to design a model that draws upon information from the source text as well as the scoring rubric to assess Organization in RTA.

Second, while past studies have focused on the writing of advanced students (i.e., in high school and beyond), we evaluate the writing of students in grades 5 through 8. An implication of this is that the pieces are typically very short, full of grammatical and spelling errors, and not very sophisticated in terms of organization. This difference in the population under study renders our task more complex than in previous studies.

Third, we sought to develop a model that is consistent with the rubric criteria and easily explainable. Such a model has greater potential to generate useful feedback to students and teachers.

In this paper, we first introduce the data (a set of responses written by 5th and 6th graders, and a set by students in grades 6-8). Next, we explain the two different structures we designed from which we extracted features. Then we explain the features, experiments, and results. We show that in general, our rubric-based task-dependent model performs as well as (if not better than) the rigorous baselines we used. Moreover, we show that different approaches to evaluating organization in student writing work differently on different populations. On shorter and noisier essays from grades 5-6, the rubric-based model performs better than the baselines. Meanwhile, for essays from grades 6-8, our rubric-based model does not perform significantly differently from the baselines; however, the combination of our new features with the baselines performs the best. Finally, we show that even a lexical chaining baseline can be improved with the use of topic information from the source text.

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 Organization dimension: Yes because Poverty should be beaten. Their are solutions to the Promblem that keep people impoverished. In 2004 two adults and three children was rushed to the hospital because of a disease. The disease was called Malaria. Mosquitoes carry Malaria. They pass it to people by bitting them. 20,000 kids die from malaria each day. A brighter future is a better life and better health. Poverty means to be Poor or have no money. People can end poverty. Ending poverty is easy. In 2004 Hannah Sachs visited the Millenium Villages Project in Kenya, a country in Africa. While they was there they saw people that were bare footed and had tattered clothing. The country that they went to had Poverty. She felt bad for the people. The Millennium Villages Project was created to help reach the Millennium Development Goals.
Essay with score of 4 on Organization dimension: 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 erent 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

Table 1: A small excerpt from the *Time for Kids* article, the prompt, and sample low and high-scoring essays from grades 5–6.

2 Data

Our dataset consists of student writing from the RTA introduced in Correnti et al. (2013). Specifically, we have two datasets from two different age groups (grades 5-6 and grades 6-8), which represent different levels of writing proficiency.

The administration of the RTA involves 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 two 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, have many spelling and grammatical errors, and the modal essays score at a basic level on Organization. Some statistics about the datasets are in Table 2.

The student responses have been assessed on five dimensions, each on a scale of 1-4 (Correnti et al., 2013). 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. The corpus from grades 5-6 consists of 1580 essays, with 602 of them double-scored for inter-rater reliability. The other corpus includes 812 essays, with almost all of them (802) double-scored. Inter-rater agreement (Quadratic Weighted Kappa) for Organization on the double-scored portion of the grades 5-6 and 6-8 corpora respectively are 0.68 and 0.69.

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 2: The two dataset's statistics

In this paper we focus only on predicting the score of the Organization dimension. The distribution of Organization scores is 398 (25%) ones, 714 (46%) twos, 353 (22%) threes, and 115 (7%) fours on the grades 5-6 dataset, and 128 (16%) ones, 316 (39%) twos, 246 (30%) threes, and 122 (15%) fours on the grades 6-8 dataset. Higher scores on the 6–8 corpus indicate that the essays in this dataset have better organization than the student essays in the 5–6 dataset. The rubric for this dimension is shown in Table 3.

1	2	3	4
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
Has little or no sense of beginning, middle, and end(2) (i.e., Lacks topic and concluding sentence, or has no identifiable middle)	Has a limited sense of beginning, middle, and end(2) (i.e., Lacks a topic or concluding sentence, or has short development in middle)	Has an adequate sense of beginning, middle, and end(2) (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 (2) (i.e., Must have topic sentence and concluding sentence that match up and relate closely to the same key idea, and well-developed middle)
Has little or no order; May feature a rambling collection of thoughts or list-like ideas with little or no flow(4)(5)	Attempts to address different ideas in turn+, in different parts of the response(3) (i.e., Some ideas may be repeated in different places)	Addresses different ideas in turn+, in different parts of the response(3), although multiple paragraphs may not be used(1)	Features multiple appropriate paragraphs (1), each addressing a different idea+
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)	Has some uneven or illogical flow from sentence to sentence or idea to idea (3)	Demonstrates logical flow from sentence to sentence and idea to idea(3)	Demonstrates logical and seamless flow from sentence to sentence and idea to idea(3)
*In implementation, when scoring the rubric 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. (4)(5)			

Table 3: Rubric for the Organization dimension of RTA. The numbers in the parentheses identify the corresponding feature group in section 4 that is aligned with that specific criteria.

3 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 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.

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 had experts of the RTA manually extract the exhaustive list of topics discussed in the article. Similarly, in other studies on evaluation of content (typically in short answer scoring), the identification of concepts and topics is manual (Liu et al., 2014). 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, we extract topics that provided evidence for the "before" and "after" states, respectively. That is, except for some general topics which are related to the conclusion of the text, 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 .

The resulting list of topics was used to generate the rows of the topic-grid. The experts defined 7 different topics; 4 of them have before and after states, resulting in 11 sub-topics in total. Each sub-topic is defined by an exhaustive list of related examples from the text. For instance, the topic "Hospitals_after" (extracted from part of the article mentioned in Table 1) includes 5 examples that are shown here by their domain words (we use the stemmed version of the words): "*1. Yala sub-district hospital medicine 2. medicine free charge 3. water connected hospital 4. hospital generator electricity 5. medicine common diseases*".

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.

	1	2	3	4	5	6	7	8	9	10
Hospitals.b	-	x	-	-	-	-	-	-	-	-
Hospitals.a	-	-	x	-	-	-	-	-	-	-
Education.b	-	-	-	x	-	-	-	-	-	-
Education.a	-	-	-	-	x	x	-	-	-	-
Farming.b	-	-	-	-	-	-	x	-	-	-
Farming.a	-	-	-	-	-	-	-	x	-	-
General	x	-	-	-	-	-	-	-	x	x

Topic	Chain
Hospitals	(b,2),(a,3)
Education	(b,4),(a,5),(a,6)
Farming	(b,7),(a,8)

Table 4: The topic-grid (on the top) and topic-chains (on the bottom) for the example essay with score=4 in Table 1. *a* and *b* indicate *after* and *before* respectively.

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. Otherwise, we break the sentence into multiple text units where each of them covers a different topic (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 4 for an example of a topic-grid for the essay with the score of four in Table 1. Consider the third column in the grid. It represents the bold text unit (the second part of the second sentence) in Table 1. The corresponding sentence has two text units since it covers two different topics "Hospitals_before" and "Hospitals_after". 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: the index of the text unit it appears in and whether it is a *before* or *after* state. We do not consider chains related to general topics that do not have a *before* or *after* state. Examples of topic-chains are presented in Table 4. Finally, we extract several features, explained in section 4, from the grid and the chains to represent some criteria from the rubric.

4 Features

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 one criterion 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, and it can be generalized to other contrasting prompts. Below, we explain each of the features and its relation to the rubric. Each group of features is indicated with a number that relates it to the corresponding criteria in the rubric in Table 3.

(1) Surface: 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: 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 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: 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.

(4) Topic development: 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: It is not just the number of topics and the amount of development on each topic that is important. More impor-

tant 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 3, little or no order in the discussion of topics in the essay means poor organization. In this section 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.

5 Experiments and Results

5.1 Experimental Setup

We configure a series of experiments to test the validity of three hypotheses: H1) the new features perform better than the baselines; H2) the topic-grid model performs better on shorter and noisier essays than longer and well-written essays; H3) the lexical chaining baseline can be improved with the use of topic information from the source document.

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 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. 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.

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 (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 using different configurations for the entity-grid baseline, and we find that the best model is an entity-grid model with history=2, salience=1, syntax=on and type/token ratios. 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 (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.

	Model	(5–6)	(6–8)
1	EntityGridTT	0.42	0.49
2	LEX1	0.45	0.53 (1)
3	EntityGridTT+LEX1	0.46 (1)	0.54 (1)
4	Rubric-based	**0.51** (1,2,3)	0.51
5	EntityGridTT+Rubric-based	0.49 (1,2,3)	0.53 (1)
6	LEX1+Rubric-based	**0.51** (1,2,3)	0.55 (1)
7	EntityGridTT+LEX1 +Rubric-based	0.50 (1,2,3)	**0.56** (1)

Table 5: Performance of our rubric-based model compared to the baselines on both datasets. The numbers in parenthesis show the model numbers which the current model performs significantly better than.

	Model	Train(5–6) Test(6–8)	Train(6–8) Test(5–6)
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

Table 6: Performance of our rubric-based model compared to the baselines. 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.

5.2 Results and Discussion

We first examine the hypothesis that the new features perform better than the baselines (H1). The results on the corpus of grades 5-6 (see Table 5) 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. The best result in all experiments is bolded.

We repeated the experiments on the corpus of grades 6-8. The results in Table 5 show that there is no significant difference between the rubric-based model and the baselines, except that in general, models that include lexical chaining features perform better than those with entity-grid features.

We configured another experiment to examine the generalizability of the models 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 6 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.

Altogether, our first and second hypotheses seem to hold. On the grade 5-6 data, the rubric-based model performs better than the baselines; for grades 6-8, the rubric-based features add value to the baselines. That is, with shorter and noisier essays, models based on coarse-grained topic information outperform state-of-the-art models based on syntactic and lexical information. Moreover, while the state of the art models perform better on better-written essays, to get an even better performing model for essays written by younger children, we need a model that examines more and different aspects of organization. Additionally, we believe that the rubric-based, task-dependent model yields more information about students' writing skills that could be fed back to teachers (and students) than the baselines.

Next, we repeated all of the experiments using each of the isolated groups of features. The results in Table 7 show that Topic-Development and Topic-Ordering are the most predictive set of features. 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.

We also conducted an ablation test to investigate how important each group of features is in the new model. In the first phase, we remove each group of features and select the one that decreases the performance most significantly. This group of features has the greatest influence after accounting for all other

	Model	(5–6) Cross-val	(6–8) Cross-val	Train(5–6) Test(6–8)	Train(6–8) Test(5–6)
1	TopicDevelopment	0.40	0.42	0.43	0.36
2	TopicOrdering	0.40	0.43	0.44	0.43
3	TopicDevelopment+TopicOrdering	0.42	0.45	0.46	0.40
4	Surface	0.32	0.40	0.42	0.35
5	LocalCoherence+ParagraphTransition	0.20	0.21	0.23	0.18
6	DiscourseStrucutre	0.25	0.19	0.26	0.22

Table 7: Performance of each group of features in isolation. The first two columns are for cross validation experiments. The last two column are the results for training on one corpus and testing on the other one.

features. In the second phase, we repeat the experiment, having already removed the most influential feature. We continue the experiment until we have reached a single group of features. The results show that the features in order of their importance are: $Surface > TopicOrdering > LocalCoherence + ParagraphTransitions > DiscourseStructure > TopicDevelopment$. In this test, surface features were more influential than topic ordering, despite the fact that topic-ordering in isolation is more predictive than surface features. One potential reason might be that the surface features may not be correlated with other task-dependent features such as topic-ordering and topic development. Examining the correlation between some of the features across feature groups is an area for future investigation.

As for Hypothesis 3, as we suggested in section 3, to measure the coherence in our text-based essays, we need to use the information from the source text. To reprise the example in section 3, we think there should be a chain containing both of the words "hospital" and "water", and a relation between them. To examine this claim, we modified the lexical chaining algorithm in such a way that it uses both external sources to measure semantic similarity and also our list of topics extracted from the source text. If we are adding a word w_1 from subtopic t_1 and there is a chain containing a word w_2 on the same subtopic t_1, there should be a relation in the chain between w_1 and w_2. If there is no Strong or Extra-Strong semantic relation between w_1 and w_2, we consider the relation as Medium-Strong. The relations are defined per Hirst and St-Onge (1998).

Table 8 presents the effect of this modification on the performance. As hypothesized, the modified version performs significantly better than the base lexical chains on essays from grades 5-6.

	Model	(5-6)	(6-8)
1	LEX1	0.45	0.53
2	LEX1+Topic	0.**48** (1)	**0.54**

Table 8: Performance of the baseline and the topic-extended lexical chaining model on the two datasets.

6 Conclusion and Future Work

We present the results for predicting the score of the Organization dimension 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 5-8. 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 either baseline on both datasets. On the shorter and noisier essays from grades 5-6, the rubric-based model performs better than the baselines. On the better-written essays from grades 6-8, the rubric-based features can add value to the baselines. We also show that the lexical chaining baseline can be improved on shorter and noisier data if we extend it using task-dependent information from the text.

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 based approaches.We need to tune all our parameters that were chosen intuitively or were set to the default value. We will test the generalizability of our model by using other texts and prompts from other response-to-text writing tasks. We would also like to extract topics and words automatically, 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).

Acknowledgments

This work was supported by the Learning Research and Development Center at the University of Pittsburgh. We thank the ITSPOKE group for their helpful feedback and suggestions.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 141–148.

Remco R Bouckaert and Eibe Frank. 2004. Evaluating the replicability of significance tests for comparing learning algorithms. In *Advances in knowledge discovery and data mining*, pages 3–12.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. 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*.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2003a. Criterion sm : Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003b. Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39, January.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 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, pages 681–684.

Jill Burstein, Joel Tetreault, and Martin Chodorow. 2013. Holistic discourse coherence annotation for noisy essay writing. *Dialogue & Discourse*, 4(2):34–52.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Robin Cohen. 1987. Analyzing the structure of argumentative discourse. *Comput. Linguist.*, 13(1-2):11–24, January.

Richard Correnti, Lindsay Clare Matsumura, Laura S Hamilton, and Elaine Wang. 2012. Combining multiple measures of students' opportunities to develop analytic, text-based writing skills. *Educational Assessment*, 17(2-3):132–161.

Richard Correnti, Lindsay Clare Matsumura, Laura S Hamilton, and Elaine Wang. 2013. Assessing students' skills at writing in response to texts. *Elementary School Journal*, 114(2):142–177.

Scott A Crossley and Danielle S McNamara. 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*, pages 1236–1241.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Michel Galley and Kathleen Mckeown. 2003. Improving word sense disambiguation in lexical chaining. In *In Proceedings of IJCAI*, pages 1486–1488.

Barbara J Grosz, Scott Weinstein, and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *HLT-NAACL*, pages 185–192.

Graeme Hirst and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: An electronic lexical database*, 305:305–332.

T.K. Landauer, P.W. Foltz, and D. Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25:259–284.

Ou Lydia Liu, Chris Brew, John Blackmore, Libby Gerard, Jacquie Madhok, and Marcia C Linn. 2014. Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, 33(2):19–28.

Eleni Miltsakaki and Karen Kukich. 2000. Automated evaluation of coherence in student essays. In *Proceedings of LREC 2000*.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Comput. Linguist.*, 17(1):21–48.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16.

Zahra Rahimi, Diane J Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa.

2014. Automatic scoring of an analytical response-to-text assessment. In *Intelligent Tutoring Systems*, pages 601–610. Springer.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 950–961. Dublin City University and Association for Computational Linguistics.