Evaluating Multiple Aspects of Coherence in Student Essays

Derrick Higgins
Educational
Testing Service

Jill Burstein
Educational
Testing Service

Daniel Marcu
University of Southern
California
/ Information Sciences
Institute

Claudia Gentile
Educational
Testing Service

Abstract

Criterion$^{\mathrm{SM}}$ Online Essay Evaluation Service includes a capability that labels sentences in student writing with essay-based discourse elements (e.g., thesis statements). We describe a new system that enhances *Criterion*'s capability, by evaluating multiple aspects of coherence in essays. This system identifies features of sentences based on semantic similarity measures and discourse structure. A support vector machine uses these features to capture breakdowns in coherence due to relatedness to the essay question and relatedness between discourse elements. Intra-sentential quality is evaluated with rule-based heuristics. Results indicate that the system yields higher performance than a baseline on all three aspects.

1 Overview

This work is motivated by a need for advanced discourse analysis capabilities for writing instruction applications. *Criterion*$^{\mathrm{SM}}$ Online Essay Evaluation Service is an application for writing instruction which includes a capability to annotate sentences in student essays with discourse element labels. These labels include the categories Thesis Statement, Main Idea, Supporting Idea, and Conclusion (Burstein et al., 2003b). Though it accurately annotates sentences with essay-based discourse labels, *Criterion* does not provide an evaluation of the expressive quality of the sentences that comprise a discourse segment. The system might accurately label a student's essay as having all of the typically expected discourse elements: thesis statement, 3 main ideas, supporting evidence linked to each main idea, and a conclusion. As teachers have pointed out, however, an essay may have all of these organizational elements, but the quality of individual elements may need improvement.

In this paper, we present a capability that captures expressive quality of sentences in the discourse segments of an essay. For this work, we have defined expressive quality in terms of four aspects related to global and local essay coherence. The first two dimensions capture global coherence, and the latter two relate to local coherence: a) relatedness to the essay question (topic), b) relatedness between discourse elements, c) intra-sentential quality, and d) sentence-relatedness within a discourse segment. Each dimension represents a different aspect of coherence.

Essentially, the goal of the system is to be able to predict whether a sentence in a discourse segment has high or low expressive quality with regard to a particular coherence dimension. We have deliberately developed an approach to essay coherence that is comprised of multiple dimensions, so that an instructional application may provide appropriate feedback to student writers, based on the system's prediction of high or low for each dimension. For instance, sentences in the student's thesis statement may have a strong relationship to the essay topic, but may have a number of serious grammatical errors that make it hard to follow. For this student, we may want to point out that on the one hand, the sentences in the thesis address the topic, but the thesis statement as a discourse segment might be more clearly stated if the grammar errors were fixed. By contrast, the sentences that comprise the student's thesis statement may be grammatically correct, but only loosely related to the essay topic. For this student, we would also want the system to provide appropriate feedback to, so that the student could revise the thesis statement text appropriately.

In earlier work, Foltz, Kintsch & Landauer (1998), and Wiemer-Hastings & Graesser (2000) have developed systems that also examine coherence in student writing. Their systems measure lexical relatedness between text segments by using vector-based similarity between adjacent sentences. This linear approach to similarity scoring is in line with the TextTiling scheme (Hearst and Plaunt, 1993; Hearst, 1997), which may be used to identify the subtopic structure of a text. Miltsakaki and Kukich (2000) have also addressed the issue of establishing the coherence of student essays, using the Rough Shift element of Centering Theory. Again, this previous work looks at the relatedness of adjacent text segments, and does not explore global aspects of text coherence.

Hierarchical models of discourse have been applied to the question of coherence (Mann and Thompson, 1986), but so far these have been more useful in language generation than in determining how coherent a given text is, or in identifying the specific problem, such as the breakdown of coherence in a document.

Our approach differs in fundamental ways from this earlier work that deals with student writing. First, Foltz et al. (1998), Wiemer-Hastings and Graesser (2000), and Miltsakaki and Kukich (2000) assume that text coherence is linear. They calculate the similarity between adjacent segments of text. By contrast, our approach considers the discourse structure in the text, following Burstein et al. (2003b). Our method considers sentences with regard to their discourse segments, and how the sentences relate to other text segments both inside (such as the essay thesis) and outside (such as the essay topic) of a document. This allows us to identify cases in which there may be a breakdown in coherence due to more global aspects of essay-based discourse structure. Second, previous work has used Latent Semantic Analysis as a semantic similarity measure (Landauer and Dumais, 1997). We have adapted another vector-based method of semantic representation: Random Indexing (Kanerva et al., 2000; Sahlgren, 2001). Another difference between our system and earlier systems is that we use essays manually annotated on the four coherence dimensions to train our system.

The final system employs a hybrid approach to classify the first two of the four coherence dimensions with a high or low quality rank. For these dimensions, a support vector machine is used to model features derived from Random Indexing and from essay-based discourse structure information. A third local coherence dimension component is driven by rule-based heuristics. A fourth dimension related to coherence within a discourse segment cannot be classified due to a lack of data characterizing low expressive quality. This is fully explained later in the paper.

2 Protocol Development and Human Annotation

2.1 Protocol Development

The development of this system required a corpus of human annotated essay data for modeling purposes. In the end, the goal is to have the system make judgments similar to those made by a human with regard to ranking the coherence of an essay on four dimensions. Therefore, we created a detailed protocol for annotating the expressive quality of essay-based discourse elements in essays with regard to four aspects related to global and local essay coherence. This protocol was designed for the following purposes:

1. To yield annotations that are useful for the purpose of providing students with feedback about the expressive relatedness of discourse elements in their essays, given four relatedness dimensions;
2. To permit human annotators to achieve high levels of consistency during the annotation process;
3. To produce annotations that have the potential of being derivable by computer programs through training on corpora annotated by humans.

2.1.1 Expressive Quality of Discourse Segments: Protocol Description

According to writing experts who collaborated in this work, the expressive relatedness of a sentence discourse element may be characterized in terms of four dimensions: a) relationship to prompt (essay question topic), b) relationship to other discourse elements, c) relevance with discourse segment, and d) errors in grammar, usage, and mechanics. For the sake of brevity, we refer to these four dimensions as Dim_P (relatedness to *prompt*), Dim_T (typically, relatedness to *thesis*), Dim_S (relatedness within a discourse *segment*), and Dim_{ERR}.

The two annotators were required to label each sentence of an essay for expressive quality on the four dimensions (above). For the 989 essays used in this study, each sentence had already been manually annotated with these discourse labels: background material, thesis, main idea, supporting idea, and conclusion (Burstein et al., 2003b).[1] An assignment of high (1) or low (0) was given to each sentence, on the dimensions relevant to the discourse element. Not all dimensions apply to all discourse elements. The protocol is extremely specific as to how annotators should label the expressive quality for each sentence in a discourse element with regard to the four dimensions. In this paper, we provide a brief description of the labeling protocol, so that the purpose of each dimension is clear.

Figure 1 shows a sample essay and prompt. A human judge has assigned a label to each sentence in the essay, resulting in the illustrated division into discourse segments. In addition, the figure indicates human annotators' ratings for two of our coherence dimensions (Dim_P and Dim_T, discussed below). By and large, the essay consistently follows up on the ideas of the essay thesis, and so most sentences get a high relatedness score on Dim_T. However, much of the essay fails to directly address the question posed in the essay prompt, and so many sentences are assigned low relatedness on Dim_P.

Dimension 1: $\mathbf{Dim_P}$ (Relatedness to Prompt)

The text of the discourse element and the prompt (text of the essay question) must be related. Specifically, the thesis statement, main ideas, and conclusion statement should all contain text that is strongly related to the essay topic. If this relationship does not exist, this is perhaps evidence that the student has written an off-topic essay. For this dimension, a high rank is assigned to each sentence from background material, thesis, main idea and conclusion statement that is related to the prompt text; otherwise a low rank is assigned.

[1] The annotated data from the Burstein et al. (2003b) study were used to develop a commercial application that automatically assigns these discourse labels to student essays.

Discourse Segment	Sentence	Dim_P	Dim_T
Prompt	**Images of beauty–both male and female–are promoted in magazines, in movies, on billboards, and on television. Explain the extent to which you think these images can be beneficial or harmful.**		
Background	A lot of people really care about how they look or how other people look.	Low	High
	A lot of people like reading magazines or watch t.v about how you can fix your looks if you don't like the way your looks are.	High	High
Thesis	People that care about how they look is because they have problems at home, their parents don't pay attention to them or even that they have a high self-steem which that is not good.	Low	N/A
	A lot of people get to the extent of killing themselfs just because they're not happy with there looks.	Low	N/A
Support	Many people go thru make-overs to experiment how they will look but, some people still don't like themself.	N/A	High
Main Point	The people that don't like themselfs need some helps and they probably feel like that because they have told them oh! your ugly , you look like Blank! or maybe a guy never ask a her out.	Low	Low
Support	In case of a guy probably the same comments but he won't dare to ask a girl out because he feels that the girl is going to say no because of the way he looks.	N/A	High
	Things like this make people don't like each other.	N/A	High
Conclusion	I suggest that a those people out here that are not happy with their looks get some help.	Low	High
	Theirs alot of programs that you can get help.	Low	Low

Figure 1: Student essay with discourse segments and two coherence dimensions as annotated by human judge

Dimension 2: $\mathrm{Dim_T}$ (Relatedness to Thesis)

The relationship between a discourse element and other discourse elements in the text governs the global coherence of the essay text. For a text to hold together, certain discourse elements must be related or the text will appear choppy and will be difficult to follow. Specifically, a high rank is assigned to each sentence in the background material, main ideas and conclusion that is related to the thesis, and supporting idea sentences that relate to the relevant main idea. A conclusion sentence may also be given a high rank if it is related to a main idea or background information. Low ranks are assigned to sentences that do not have these relationships.

Dimension 3: $\mathrm{Dim_S}$ (Relatedness within Segment)

This dimension indicates the cohesiveness of the multiple sentences in a discourse segment of a text. This dimension distinguishes a text segment that may go off task within a discourse segment. For this dimension, a high rank was assigned to each sentence in a discourse segment that related to at least one other sentence in the segment; otherwise the sentence received a low rank. If the discourse segment contained only one sentence, then the Dim_T label was assigned as the default.

Dimension 4: $\mathrm{Dim_{ERR}}$ (Technical Errors)

Dimension 4 measures a sentence's relatedness of expression with regard to grammar, mechanics and word usage. More specifically, a sentence is considered to be low on this dimension if it contains frequent patterns of error, defined as follows: (a) contains 2 errors in grammar, word usage or mechanics (i.e., spelling, capitalization or punctuation), (b) is an incomplete sentence, or (c) is a run-on sentence (i.e., 4 or more independent clauses within a sentence).

2.2 Topics, Human Annotation, and Human Agreement

2.2.1 Topics & Writing Genre

Essays written to two genres were used: five of the topics were *persuasive*, and one was *expository*. Persuasive writing requires the reader to state an opinion on a particular topic, support the stated opinion, and convince the reader that the perspective is valid and well-supported. An expository topic requires the writer only to state an opinion on a topic. This typically elicits more personal and descriptive writing. Four of the five sets of persuasive essay responses were written by college freshman, and the fifth by 12th graders. The set of expository responses were also written by 12th graders.

2.2.2 Human Annotation

Two human judges participated in this study. The judges were instructed to assign relevant dimension labels to each sentence. Pre-training of the judges was done using a set of approximately 50 essays across the six topics in the study. During this phase, the authors and the judges discussed and labeled the essays together. During the next training phase, the judges labeled a total of 292 essays across six topics. They labeled the identical set of essays, and were allowed to discuss their decisions. In the next annotation phase, the judges did not discuss their annotations. In this post-training phase (annotation phase), each judge labeled an average of about 278 unique essays for each of four prompts (556 essays together). Each judge also labeled an additional set of 141 essays that was overlapping. So, about 20 percent of the data annotated by each judge in the annotation phase was overlapping,

	Agreement	κ
Dim_P (N=779)	99%	.99
Dim_T (N=1890)	100%	.99
Dim_S (N=2119)	100%	.99
Dim_{ERR} (N=2170)	99%	.98

Table 1: Annotator agreement across coherence dimensions—data from annotation phase

and 80 percent was unique. The 20 percent is used to obtain human agreement.[2] During both the training and annotation phases, Kappa statistics were run on their judgments regularly, and if the Kappa for any particular category fell below 0.8, then the judges were asked to review the protocol until their agreement was acceptable. At the end of the annotation phase, we had a total of 989 labeled essays: 292 (training phase) + 278 × 2 (unique essays from annotator 1 + annotator 2, annotation phase) + 141 (overlapping set, annotation phase).

Human Judge Agreement

It is critical that the annotation process yields agreement that is high enough between human judges, such that it suggests that people can agree on how to categorize the discourse elements. As is stated in the above section, during the training of the judges for this study, Kappa statistics were computed on a regular basis. Kappa between the judges for each category had to be maintained at least 0.8, since this is believed to represent strong agreement (Krippendorff, 1980). In Table 1 we report human agreement for overlapping data from the four topics on all four dimensions. Clearly, the level of human agreement is quite high across all four coherence dimensions. In addition, if we look at kappas of sentences based on discourse category, no kappa falls below 0.9.

3 Method

Our final system uses a hybrid approach to label three of the four coherence dimensions. For Dim_P and Dim_T, assigning coherence judgments to sentences in an essay proceeds in three stages 1) identifying the discourse label associated with each sentence in an essay, 2) computing features that quantify the semantic similarity between different discourse segments of the essay, and 3) applying a classifier to make a coherence judgment on a dimension. Consistent with the human annotated data, a coherence judgment on any dimension is either “high” or “low.” The method for Dim_{ERR} is rule-based, and is discussed later.

3.1 Discourse element feature identification

As noted earlier, the two human judges in this study annotated the four coherence dimensions according to the human discourse label assignments. Accordingly, we also used the human assigned discourse labels as features for predicting coherence judgments. In a deployed system, however, we would use discourse element labels generated from *Criterion*’s discourse analysis system (Burstein et al., 2003b). Further evaluation is, of course, necessary in order to determine the effect of using these automatically assigned labels in place of the gold standard discourse labels.

3.2 Semantic similarity features

Given the partition of an essay into discourse segments, we then derive a set of features from the essay in order to predict how closely related each sentence is to various important text segments, such as the essay topic, and discourse elements, such as thesis statement. As described in Section 4, the features that are most useful for classifying sentences according to coherence are *semantic similarity* features derived from Random Indexing (Kanerva et al., 2000; Sahlgren, 2001). Random Indexing is a vector-based semantic representation system similar to Latent Semantic Analysis. Our Random Indexing (RI) semantic space is trained on about 30 million words of newswire text.

When we extract a feature such as “RI similarity to prompt” for a sentence, this essentially measures to what extent the sentence contains terms in the same semantic domain as compared to those found in the prompt. Within any discourse segment, any semantic information that is word-order dependent is lost.

3.3 Support vector classification

Finally, for each sentence in the essay we use the features derived from the essay to make a determination as to whether it meets our criteria for coherence in these dimensions (Dim_P and Dim_T). To make this determination, we use a support vector machine (SVM) classifier (Vapnik, 1995; Christianini and Shawe-Taylor, 2000). Specifically, we use an SVM with a radial basis function kernel, which exhibited good performance on a subset of about 30 essays from the pre-training data.

4 Results

In each of the experiments below, the results are reported for the entire set of 989 essays annotated for this project. We performed ten-fold cross-validation, training our SVM classifier on $\frac{9}{10}$ of the data at a time, and testing on the remaining $\frac{1}{10}$. We report the results on the cross-validation set for all runs combined.

For each dimension, we also report the performance of a simple baseline measure, which assumes that all of our essay coherence criteria are satisfied. That is, **our baseline assigns category 1 (high relevance) to every sentence, on every dimension.**

These essays were written in response to six different prompts, and had an average (human-assigned) score of

[2] For the annotation phase, we were unable to collect data for two essay prompts because of our annotators’ availability. This means that we only have inter-annotator agreement statistics on 4 prompts, although some data from all six prompts was available for training and testing our models (with the extra two prompts being represented in the training phase of annotation).

Score	Dim_P	Dim_T	Dim_S	Dim_{ERR}
1–2	64.1%	71.2%	94.8%	61.1%
5–6	72.0%	70.9%	97.2%	92.9%

Table 2: Baseline performance on each coherence dimension, broken down by essay score point

4.0 on a six-point scale. Therefore, *a priori*, it seems possible that we could build a better baseline model by conditioning its predictions on the overall score of the essay (assigning 1's to sentences from better-scoring essays, and 0's to sentences from lower-scoring essays). However, the coherence requirements of each of our dimensions are usually met even in the lowest-scoring essays, as shown in Table 2, which lists the percentage of sentences in different essay score ranges which our human annotators assigned category 1. Looking at the highest and lowest score points on our six-point scale, it is clear that higher-scoring essays do tend to have fewer problems with coherence, but this effect is not overwhelming. (The largest gap between the highest- and lowest-scoring essays is on Dim_{ERR}, which deals with errors in grammar, usage, and mechanics.)

4.1 $\mathbf{Dim}_P$

According to the protocol, there are four discourse elements for which Dim_P, the degree of relatedness to the essay prompt, is relevant: Background, Conclusion, Main Point, and Thesis. The Supporting Idea category of sentence is not required to be related to the prompt, because it may express an elaboration of one of the main points of the essay, and has a more tenuous and mediated logical connection to the essay prompt text.

The features which we provide to the SVM for predicting a sentence's relatedness to the prompt are:

1. The RI similarity score of the target sentence with the entire essay prompt,
2. The maximum RI similarity score of the target sentence with any sentence in the essay prompt,
3. The RI similarity score of the target sentence with the required task sentence (a designated portion of the prompt text which contains an explicit directive to the student to write about a specific topic),
4. The RI similarity score of the target sentence with the entire thesis of the essay,
5. The maximum RI similarity score of the target sentence with any sentence in the thesis,
6. The maximum RI similarity score of the target sentence with any sentence in the preceding discourse chunk,
7. The number of sentences in the current chunk,
8. The offset of the target sentence (sentence number) from the beginning of the current discourse chunk,
9. The number of sentences in the current chunk whose similarity with the prompt is greater than .2,
10. The number of sentences in the current chunk whose similarity with the required task sentence is greater than .2,
11. The number of sentences in the current chunk whose similarity with the essay thesis is greater than .2,
12. The number of sentences in the current chunk whose similarity with the prompt is greater than .4,
13. The number of sentences in the current chunk whose similarity with the required task sentence is greater than .4,
14. The number of sentences in the current chunk whose similarity with the essay thesis is greater than .4,
15. The length of the target sentence in words,
16. A Boolean feature indicating whether the target sentence contains a *transition word*, such as "however", or "although",
17. A Boolean feature indicating whether the target sentence contains an anaphoric element, and
18. The category of the current chunk. (This is encoded as five Boolean features: one bit for each of "Background", "Conclusion", "Main Point", "Supporting Idea", and "Thesis".)

In calculating features 2, 5, and 6, we use the *maximum* similarity score of the sentence with any other sentence in the relevant discourse segment, rather than simply using the similarity score of the sentence with the entire text chunk. We add this feature based on the intuition that for a sentence to be relevant to another discourse segment, it need only be connected to some *part* of that segment.

It is perhaps surprising that we include features which measure the degree of similarity between the sentence and the thesis, since we are trying to predict its relatedness to the prompt, rather than the thesis. However, there are two reasons we believe this is fruitful. First, since we are dealing with a relatively small amount of text, comparing a single sentence to a short essay prompt, looking at the thesis as well helps to overcome data sparsity issues. Second, it may be that the relevance of the current sentence to the prompt is mediated by the student's thesis statement. For example, the prompt may ask the student to take a position on some topic. They may state this position in the thesis, and provide an example to support it as one of their Main Points. In such a case, the example would be more clearly linked to the Thesis, but this would suffice for it to be related to the prompt.

Considering the similarity scores of sentences in the current discourse segment is also, in part, an attempt to overcome data sparsity issues, but is also motivated by the idea that it may be an entire discourse segment which can properly be said to be (ir)relevant to the essay prompt.

The sentence length and transition word features do not directly reflect the relatedness of a sentence to the prompt, but they are likely to be useful correlates.

Finally, the feature (#17) indicating the presence of a pronoun is to help the system deal with cases in which a sentence contains very few content words, but is still linked to other material in the essay by means of anaphoric elements, such as "*This* is shown by my argument." In such as case, the sentence would normally get a low similarity score with the prompt (and other parts of the essay), but the information that it contains a pronoun might still allow the system to classify it correctly.

Table 3 shows results using the baseline algorithm to classify sentences according to their relatedness to the prompt. Table 4 presents the results using the SVM classifier. We provide precision, recall, and f-measure for the assignment of the labels 1 and 0, and an overall accuracy measure in the far right column. (The accuracy measure is the value for precision and recall when 1 and 0 ranks are collapsed. Precision and recall will be the same, since the number of labels assigned by the model is equal to the number of labels in the target assignment.)

The SVM model outperforms the baseline on every subcategory, with the largest gains on Background sentences, most of which are, in fact, unrelated to the prompt according to our human judges. This low baseline result on Background sentences could indicate that many students have a problem with providing unnecessary and irrelevant prefaces to the important points in their essays.

Note that the trained SVM has around .9 recall on the class of sentences which according to our human annotators have high relevance to the prompt. This means that our system is less likely to incorrectly assign a low rank to a sentence that is high. So, the system will tend to err on the side of the student, which is a preferable trade-off. In part, this is due to the nature of the semantic similarity measure we are using, which does not take word order into account. While RI does allow us to capture a richer meaning component than simply matching words which co-occur in the target sentence and prompt, it still does not encompass all that goes into determining whether a sentence "relates" to another chunk of text. Students often write something which bears a loose topical connection with the essay prompt, but does not directly address the question. This sort of problem is hard to address with a tool such as LSA or RI; the vocabulary of the sentence on its own will not provide a clue to the sentence's failure to address the task.

4.2 $\mathbf{Dim}_T$

The annotation protocol states that these four discourse elements come into play for Dim_T: Background, Conclusion, Main Point, and Supporting Idea. Because this dimension indicates the degree of relatedness to the thesis of the essay (and also other discourse segments in the case of Supporting Idea and Conclusion sentences; see Section 2.1.1 above), we do not consider thesis sentences with regard to this aspect of coherence.

The features which we provide to the SVM for predicting whether or not a given sentence is related to the thesis are almost the same ones used for Dim_P. The only difference is that we omit features #12 and #13 in our model of Dim_T. These are the features which evaluate how many sentences in the current chunk have a similarity score with the prompt and required task sentence greater than 0.4. While Dim_P is to some degree sensitive to the similarity of a sentence to the thesis, and Dim_T can likewise benefit from the information about a sentence's similarity to the prompt, it seems that the latter link is less important, so a single cutoff suffices for this model.

Tables 5–6 present the results for our SVM model and for a baseline which assigns all sentences "high" relevance. The improvements on Dim_T are smaller than the ones reported for Dim_P, but we still record an overall gain of four percentage points in accuracy. Only on conclusion sentences were we unable to produce an improvement over the baseline; we need to investigate this further.

Again, the system achieves high recall on sentences with high relatedness. It outperforms the baseline by correctly identifying a modest percentage of the sentences labeled as having low relatedness with the thesis.

4.3 $\mathbf{Dim}_S$

Dim_S, which concerns whether the target sentence relates to another sentence within the same discourse segment, seems another good candidate for applying our semantic similarity score to the task of establishing coherence. At present, however we have not made substantial progress on this task. The baselines for Dim_S are substantially higher than those for dimensions Dim_P and Dim_T — 98.1% of all sentences in our data were annotated as "highly related" with respect to this dimension. This indicates that it is relatively rare to find a sentence which is not related to anything in the same discourse segment. This makes our task, to characterize those sentences which are not related to the discourse segment, much more difficult, since there are so few examples of sentences with low-ranking coherence.

4.4 $\mathbf{Dim}_{ERR}$

Dim_{ERR} is clearly a different kind of problem. Here, we are looking for clarity of expression, or coherence within a sentence. We base this solely on technical correctness. We are able to automatically assign high and low ranks to essay sentences using a set of rules based on the number of grammar, usage and mechanics errors. The rules used for Dim_{ERR} are as follows: a) assign a low label if the sentence is a fragment, if the sentence contains 2 or more grammar, usage, and mechanics errors, or if the sentence is a run-on, b) assign a high label if no criteria in (a) apply.

Criterion's discourse analysis system also provides an essay score with *e-rater*®, and qualitative feedback about grammar, usage, mechanics, and style (Leacock

	High			**Low**			**Total**
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	**Accuracy**
Background ($N = 1077$)	0.486	1.000	0.654	0.000	0.000	0.000	**0.486**
Conclusion ($N = 1830$)	0.757	1.000	0.862	0.000	0.000	0.000	**0.757**
Main Point ($N = 1566$)	0.663	1.000	0.797	0.000	0.000	0.000	**0.663**
Thesis ($N = 1899$)	0.712	1.000	0.832	0.000	0.000	0.000	**0.712**
All sentence types ($N = 6372$)	0.675	1.000	0.806	0.000	0.000	0.000	**0.675**

Table 3: Baseline performance on Dim_P

	High			**Low**			**Total**
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	**Accuracy**
Background ($N = 1077$)	0.714	0.702	0.708	0.723	0.735	0.729	**0.719**
Conclusion ($N = 1830$)	0.784	0.959	0.863	0.578	0.175	0.269	**0.768**
Main Point ($N = 1566$)	0.729	0.888	0.801	0.616	0.352	0.448	**0.708**
Thesis ($N = 1899$)	0.771	0.929	0.843	0.644	0.318	0.426	**0.753**
All sentence types ($N = 6372$)	0.759	0.901	0.824	0.665	0.407	0.505	**0.740**

Table 4: SVM performance on Dim_P

	High			**Low**			**Total**
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	**Accuracy**
Background ($N = 1060$)	0.793	1.000	0.885	0.000	0.000	0.000	**0.793**
Conclusion ($N = 1829$)	0.834	1.000	0.909	0.000	0.000	0.000	**0.834**
Main Point ($N = 1556$)	0.742	1.000	0.852	0.000	0.000	0.000	**0.742**
Support ($N = 10332$)	0.664	1.000	0.798	0.000	0.000	0.000	**0.664**
All sentence types ($N = 14777$)	0.702	1.000	0.825	0.000	0.000	0.000	**0.702**

Table 5: Baseline performance on Dim_T

	High			**Low**			**Total**
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**	**Accuracy**
Background ($N = 1060$)	0.856	0.980	0.914	0.827	0.368	0.509	**0.853**
Conclusion ($N = 1829$)	0.834	1.000	0.910	0.000	0.000	0.000	**0.834**
Main Point ($N = 1556$)	0.776	0.997	0.873	0.958	0.172	0.292	**0.785**
Support ($N = 10332$)	0.709	0.945	0.810	0.684	0.237	0.352	**0.706**
All sentence types ($N = 14777$)	0.744	0.962	0.839	0.709	0.221	0.337	**0.741**

Table 6: SVM performance on Dim_T

and Chodorow, 2000; Burstein et al., 2003a). We can easily use *Criterion*'s outputs about grammar, usage, and mechanics errors to assign high and low ranks to essay sentences, using the rules described in the previous section.

The performance of the module that does the Dim_{ERR} assignments is in Table 7. We used half of the 292 essays from the training phase of annotation for development, and the remaining data from the training and post-training phases of annotation for cross-validation. Results are reported for the cross-validation set. Text labeled as titles, or opening or closing salutations, are not included in the results. The baselines were computed by assigning all sentences a high rank label. The baseline is high; however, the algorithm outperforms the baseline.

	Sentence N	**Precision**	**Recall**	**F-measure**
Baseline				
High	11789	0.83	1.00	0.91
Low	2351	0.00	0.00	0.00
Overall	14140	0.83	0.83	**0.83**
Algorithm				
High	11789	0.88	0.96	0.92
Low	2351	0.63	0.34	0.44
Overall	14140	0.86	0.86	**0.86**

Table 7: Performance on Dim_{ERR}

5 Discussion and Conclusions

There were multiple goals in this work. We wanted to introduce a concept of essay coherence comprising multiple aspects, and investigate what linguistic features drive each aspect in student essay writing. Further, we wanted to build a system to automatically evaluate these multiple aspects of coherence, so that appropriate feedback can be provided through a writing instruction application.

To accomplish these goals, we have worked with writing experts to develop a comprehensive protocol that details how coherence in writing can be evaluated, either manually or automatically. Using this protocol, human annotators labeled a corpus of student essays, using the coherence dimensions. These annotations built on a previous set of annotations for these data, whereby discourse element labels were assigned. The result is a richly annotated data set with information about discourse elements, as well as their coherence in the context of the discourse structure. Using this data set, we were able to learn what linguistic features can be used to evaluate various aspects of coherence in student writing. We then developed a prototype system that ranks global and local aspects of coherence in an essay. This capability shows promise in ranking three aspects of coherence in essays: a) relationship to essay topic, b) relationship between discourse elements, and c) intra-sentential technical quality. More low ranking data on a fourth dimension, coherence within a discourse segment, needs to be identified and annotated before this dimension can be modeled.

The approach used is innovative, since it moves beyond earlier methods of evaluating coherence in student writing that capture only local information between adjacent sentences. Two methods are used to model the aspects of coherence handled by the system. For the two global coherence dimensions, Dim_P and Dim_T, a support vector machine provides a coherence ranking of sentences based on features related to essay-based discourse information, and semantic similarity values derived from the RI algorithm. Using this classification method, we are able to rank the expressive quality of sentences in essay-based discourse segments, with regard to relatedness to the text of the prompt, and also as they relate to the thesis statement. With regard to the local coherence dimension, Dim_{ERR}, we use a rule-based heuristic to rank intra-sentential quality. This addresses the issue of sentences in essays that have serious grammatical problems that may interfere with a reader's comprehension. We take advantage of *Criterion*'s identification of grammar, usage, and mechanics errors to design the rules for ranking this local coherence dimension.

We hope that in further investigation of this richly annotated data set, we will be able to build on the current prototype and develop a full-scale writing instruction capability that provides feedback on the coherence dimensions described in this paper.

Acknowledgements

We would like to thank Irma Lorenz and Shauna Cooper for advice on protocol development and for the annotation work, and Martin Chodorow for discussions about Random Indexing. We thank the anonymous reviewers for their helpful feedback.

Any opinions expressed here are those of the authors and not necessarily of the Educational Testing Service.

References

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2003a. *Criterion*$^{\mathrm{SM}}$: Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco, Mexico.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003b. Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Transactions on Intelligent Systems: Special Issue on Advances in Natural Language Processing*, 181:32–39.

Nello Christianini and John Shawe-Taylor. 2000. *Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, UK.

Peter Foltz, Walter Kintsch, and Thomas K. Landauer. 1998. The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25(2&3):285–307.

Marti A. Hearst and Christian Plaunt. 1993. Subtopic structuring for full-length document access. In *Proceedings of ACM SIGIR*, pages 59–68.

Marti A. Hearst. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

P. Kanerva, J. Kristoferson, and A. Holst. 2000. Random indexing of text samples for Latent Semantic Analysis. In L. R. Gleitman and A. K. Josh, editors, *Proc. 22nd Annual Conference of the Cognitive Science Society*.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Sage Publications.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Claudia Leacock and Martin Chodorow. 2000. An unsupervised method for detecting grammatical errors. In *Proceedings of NAACL 2000*, pages 140–147.

William Mann and Sandra Thompson. 1986. Relational processes in discourse. *Discourse Processes*, 9:57–90.

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

Magnus Sahlgren. 2001. Vector based semantic analysis: Representing word meanings based on random labels. In *Proceedings of the ESSLLI 2001 Workshop on Semantic Knowledge Acquisition and Categorisation*. Helsinki, Finland.

Vladimir Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer Verlag, New York.

Peter Wiemer-Hastings and Arthur Graesser. 2000. Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, 8(2):149–169.