Proceedings of the 52nd Annual Meeting of the Association for
Computational Linguistics (ACL), Baltimore, MD, June 2014, pp. 1534--1543.

Modeling Prompt Adherence in Student Essays

Isaac Persing and **Vincent Ng**
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688
{persingq,vince}@hlt.utdallas.edu

Abstract

Recently, researchers have begun exploring methods of scoring student essays with respect to particular dimensions of quality such as coherence, technical errors, and prompt adherence. The work on modeling prompt adherence, however, has been focused mainly on whether individual sentences adhere to the prompt. We present a new annotated corpus of essay-level prompt adherence scores and propose a feature-rich approach to scoring essays along the prompt adherence dimension. Our approach significantly outperforms a knowledge-lean baseline prompt adherence scoring system yielding improvements of up to 16.6%.

1 Introduction

Automated essay scoring, the task of employing computer technology to evaluate and score written text, is one of the most important educational applications of natural language processing (NLP) (see Shermis and Burstein (2003) and Shermis et al. (2010) for an overview of the state of the art in this task). A major weakness of many existing scoring engines such as the Intelligent Essay Assessor™(Landauer et al., 2003) is that they adopt a holistic scoring scheme, which summarizes the quality of an essay with a single score and thus provides very limited feedback to the writer. In particular, it is not clear which dimension of an essay (e.g., style, coherence, relevance) a score should be attributed to. Recent work addresses this problem by scoring a particular dimension of essay quality such as coherence (Miltsakaki and Kukich, 2004), technical errors, organization (Persing et al., 2010), and thesis clarity (Persing and Ng, 2013). Essay grading software that provides feedback along multiple dimensions of essay quality such as E-*rater*/Criterion (Attali and Burstein, 2006) has also begun to emerge.

Our goal in this paper is to develop a computational model for scoring an essay along an under-investigated dimension — *prompt adherence*. Prompt adherence refers to how related an essay's content is to the prompt for which it was written. An essay with a high prompt adherence score consistently remains on the topic introduced by the prompt and is free of irrelevant digressions.

To our knowledge, little work has been done on scoring the prompt adherence of student essays since Higgins et al. (2004). Nevertheless, there are major differences between Higgins et al.'s work and our work with respect to both the way the task is formulated and the approach. Regarding task formulation, while Higgins et al. focus on classifying each *sentence* as having either *good* or *bad* adherence to the prompt, we focus on assigning a prompt adherence score to the entire *essay*, allowing the score to range from one to four points at half-point increments. As far as the approach is concerned, Higgins et al. adopt a *knowledge-lean* approach to the task, where almost all of the features they employ are computed based on a word-based semantic similarity measure known as *Random Indexing* (Kanerva et al., 2000). On the other hand, we employ a large variety of features, including lexical and knowledge-based features that encode how well the concepts in an essay match those in the prompt, LDA-based features that provide semantic generalizations of lexical features, and "error type" features that encode different types of errors the writer made that are related to prompt adherence.

In sum, our contributions in this paper are two-fold. First, we develop a scoring model for the prompt adherence dimension on student essays using a feature-rich approach. Second, in order to stimulate further research on this task, we make our data set consisting of prompt adherence an-

Topic	Languages	Essays
Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.	13	131
The prison system is outdated. No civilized society should punish its criminals: it should rehabilitate them.	11	80
In his novel *Animal Farm*, George Orwell wrote "All men are equal but some are more equal than others." How true is this today?	10	64

Table 1: Some examples of writing topics.

notations of 830 essays publicly available. Since progress in prompt adherence modeling is hindered in part by the lack of a publicly annotated corpus, we believe that our data set will be a valuable resource to the NLP community.

2 Corpus Information

We use as our corpus the 4.5 million word International Corpus of Learner English (ICLE) (Granger et al., 2009), which consists of more than 6000 essays written by university undergraduates from 16 countries and 16 native languages who are learners of English as a Foreign Language. 91% of the ICLE texts are argumentative. We select a subset consisting of 830 argumentative essays from the ICLE to annotate for training and testing of our essay prompt adherence scoring system. Table 1 shows three of the 13 topics selected for annotation. Fifteen native languages are represented in the set of annotated essays.

3 Corpus Annotation

We ask human annotators to score each of the 830 argumentative essays along the prompt adherence dimension. Our annotators were selected from over 30 applicants who were familiarized with the scoring rubric and given sample essays to score. The six who were most consistent with the expected scores were given additional essays to annotate. Annotators evaluated how well each essay adheres to its prompt using a numerical score from one to four at half-point increments (see Table 2 for a description of each score). This contrasts with previous work on prompt adherence essay scoring, where the corpus is annotated with a binary decision (i.e., *good* or *bad*) (e.g., Higgins et al. (2004; 2006), Louis and Higgins (2010)). Hence, our annotation scheme not only provides

Score	Description of Prompt Adherence
4	essay fully addresses the prompt and **consistently stays on topic**
3	essay mostly addresses the prompt or **occasionally wanders off topic**
2	essay does not fully address the prompt or **consistently wanders off topic**
1	essay does not address the prompt at all or is **completely off topic**

Table 2: Descriptions of the meaning of scores.

a finer-grained distinction of prompt adherence (which can be important in practice), but also makes the prediction task more challenging.

To ensure consistency in annotation, we randomly select 707 essays to have graded by multiple annotators. Analysis reveals that the Pearson's correlation coefficient computed over these doubly annotated essays is 0.243. Though annotators exactly agree on the prompt adherence score of an essay only 38% of the time, the scores they apply fall within 0.5 points in 66% of essays and within 1.0 point in 89% of essays. For the sake of our experiments, whenever annotators disagree on an essay's prompt adherence score, we assign the essay the average of all annotations rounded to the nearest half point. Table 3 shows the number of essays that receive each of the seven scores for prompt adherence.

score	1.0	1.5	2.0	2.5	3.0	3.5	4.0
essays	0	0	8	44	105	230	443

Table 3: Distribution of prompt adherence scores.

4 Score Prediction

In this section, we describe in detail our system for predicting essays' prompt adherence scores.

4.1 Model Training and Application

We cast the problem of predicting an essay's prompt adherence score as 13 regression problems, one for each prompt. Each essay is represented as an instance whose label is the essay's true score (one of the values shown in Table 3) with up to seven types of features including baseline (Section 4.2) and six other feature types proposed by us (Section 4.3). Our regressors may assign an essay any score in the range of 1.0−4.0.

Using regression captures the fact that some pairs of scores are more similar than others (e.g., an essay with a prompt adherence score of 3.5 is more similar to an essay with a score of 4.0 than it is to one with a score of 1.0). A classification system, by contrast, may sometimes believe that the scores 1.0 and 4.0 are most likely for a particular essay, even though these scores are at opposite ends of the score range.

Using a different regressor for each prompt captures the fact that it may be easier for an essay to adhere to some prompts than to others, and common problems students have writing essays for one prompt may not apply to essays written in response to another prompt. For example, in essays written in response to the prompt "Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television," students sometimes write essays about all the evils of television, forgetting that their essay is only supposed to be about whether it is "the opium of the masses". Students are less likely to make an analogous mistake when writing for the prompt "Crime does not pay."

After creating training instances for prompt p_i, we train a linear regressor, r_i, with regularization parameter c_i for scoring test essays written in response to p_i using the linear SVM regressor implemented in the LIBSVM software package (Chang and Lin, 2001). All SVM-specific learning parameters are set to their default values except c_i, which we tune to maximize performance on held-out validation data.

After training the classifiers, we use them to classify the test set essays. The test instances are created in the same way as the training instances.

4.2 Baseline Features

Our baseline system for score prediction employs various features based on Random Indexing.

1. Random Indexing Random Indexing (RI) is "an efficient, scalable and incremental alternative" (Sahlgren, 2005) to Latent Semantic Indexing (Deerwester et al., 1990; Landauer and Dutnais, 1997) which allows us to automatically generate a semantic similarity measure between any two words. We train our RI model on over 30 million words of the English Gigaword corpus (Parker et al., 2009) using the S-Space package (Jurgens and Stevens, 2010). We expect that features based on RI will be useful for prompt adherence scoring because they may help us find text related to the prompt even if some of its concepts have have been rephrased (e.g., an essay may talk about "jail" rather than "prison", which is mentioned in one of the prompts), and because they have already proven useful for the related task of determining which sentences in an essay are related to the prompt (Higgins et al., 2004).

For each essay, we therefore attempt to adapt the RI features used by Higgins et al. (2004) to our problem of prompt adherence scoring. We do this by generating one feature encoding the entire essay's similarity to the prompt, another encoding the essay's highest individual sentence's similarity to the prompt, a third encoding the highest entire essay similarity to one of the prompt sentences, another encoding the highest individual sentence similarity to an individual prompt sentence, and finally one encoding the entire essay's similarity to a manually rewritten version of the prompt that excludes extraneous material (such as "In his novel Animal Farm, George Orwell wrote," which is introductory material from the third prompt in Table 1). Our RI feature set necessarily excludes those features from Higgins et al. that are not easily translatable to our problem since we are concerned with an entire essay's adherence to its prompt rather than with each of its sentences' relatedness to the prompt. Since RI does not provide a straightforward way to measure similarity between groups of words such as sentences or essays, we use Higgins and Burstein's (2007) method to generate these features.

4.3 Novel Features

Next, we introduce six types of novel features.

2. N-grams As our first novel feature, we use the 10,000 most important lemmatized unigram, bigram, and trigram features that occur in the essay. N-grams can be useful for prompt adherence scoring because they can capture useful words and phrases related to a prompt. For example, words and phrases like "university degree", "student", and "real world" are relevant to the first prompt in Table 1, so it is more likely that an essay adheres to the prompt if they appear in the essay.

We determine the "most important" n-gram features using information gain computed over the training data (Yang and Pedersen, 1997). Since the essays vary greatly in length, we normalize each essay's set of n-gram features to unit length.

3. Thesis Clarity Keywords Our next set of features consists of the keyword features we introduced in our previous work on essay thesis clarity scoring (Persing and Ng, 2013). Below we give an overview of these keyword features and motivate why they are potentially useful for prompt adherence scoring.

The keyword features were formed by first examining the 13 essay prompts, splitting each into its component pieces. As an example of what is meant by a "component piece", consider the first prompt in Table 1. The components of this prompt would be "Most university degrees are theoretical", "Most university degrees do not prepare students for the real world", and "Most university degrees are of very little value."

Then the most important (primary) and second most important (secondary) words were selected from each prompt component, where a word was considered "important" if it would be a good word for a student to use when stating her thesis about the prompt. So since the lemmatized version of the third component of the second prompt in Table 1 is "it should rehabilitate they", "rehabilitate" was selected as a primary keyword and "society" as a secondary keyword.

Features are then computed based on these keywords. For instance, one thesis clarity keyword feature is computed as follows. The RI similarity measure is first taken between the essay and each group of the prompt's primary keywords. The feature then gets assigned the lowest of these values. If this feature has a low value, that suggests that the student ignored the prompt component from which the value came when writing the essay.

To compute another of the thesis clarity keyword features, the numbers of combined primary and secondary keywords the essay contains from each component of its prompt are counted. These numbers are then divided by the total count of primary and secondary features in their respective components. The greatest of the fractions generated in this way is encoded as a feature because if it has a low value, that indicates the essay's thesis may not be very relevant to the prompt.[1]

4. Prompt Adherence Keywords The thesis clarity keyword features described above were intended for the task of determining how clear an essay's thesis is, but since our goal is instead to determine how well an essay adheres to its prompt, it makes sense to adapt keyword features to our task rather than to adopt keyword features exactly as they have been used before. For this reason, we construct a new list of keywords for each prompt component, though since prompt adherence is more concerned with what the student says about the topics than it is with whether or not what she says about them is stated clearly, our keyword lists look a little different than the ones discussed above. For an example, we earlier alluded to the problem of students merely discussing all the evils of television for the prompt "Marx once said that religion was the opium of the masses. If he was alive at the end of the 20th century, he would replace religion with television." Since the question suggests that students discuss whether television is analogous to religion in this way, our set of prompt adherence keywords for this prompt contains the word "religion" while the previously discussed keyword sets do not. This is because a thesis like "Television is bad" can be stated very clearly without making any reference to religion at all, and so an essay with a thesis like this can potentially have a very high thesis clarity score. It should not, however, have a very high prompt adherence score, as the prompt asked the student to discuss whether television is like religion in a particular way, so religion should be at least briefly addressed for an essay to be awarded a high prompt adherence score.

Additionally, our prompt adherence keyword sets do not adopt the notions of primary and secondary groups of keywords for each prompt component, instead collecting all the keywords for a component into one set because "secondary" keywords tend to be things that are important when we are concerned with what a student is saying about the topic rather than just how clearly she said it.

We form two types of features from prompt adherence keywords. While both types of features measure how much each prompt component was discussed in an essay, they differ in how they encode the information. To obtain feature values of the first type, we take the RI similarities between the whole essay and each set of prompt adherence keywords from the prompt's components. This results in one to three features, as some prompts have one component while others have up to three.

We obtain feature values of the second type as follows. For each component, we count the number of prompt adherence keywords the essay contains. We divide this number by the number of prompt adherence keywords we identified from

[1] Space limitations preclude a complete listing of the thesis clarity keyword features. See our website at `http://www.hlt.utdallas.edu/~persingq/ICLE/` for the complete list.

the component. This results in one to three features since a prompt has one to three components.

5. LDA Topics A problem with the features we have introduced up to this point is that they have trouble identifying topics that are not mentioned in the prompt, but are nevertheless related to the prompt. These topics should not diminish the essay's prompt adherence score because they are at least related to prompt concepts. For example, consider the prompt "All armies should consist entirely of professional soldiers: there is no value in a system of military service." An essay containing words like "peace", "patriotism", or "training" are probably not digressions from the prompt, and therefore should not be penalized for discussing these topics. But the various measures of keyword similarities described above will at best not notice that anything related to the prompt is being discussed, and at worst, this might have effects like lowering some of the RI similarity scores, thereby probably lowering the prompt adherence score the regressor assigns to the essay. While n-gram features do not have exactly the same problem, they would still only notice that these example words are related to the prompt if multiple essays use the same words to discuss these concepts. For this reason, we introduce Latent Dirichlet Allocation (LDA) (Blei et al., 2003) features.

In order to construct our LDA features, we first collect all essays written in response to each prompt into its own set. Note that this feature type exploits unlabeled data: it includes all essays in the ICLE responding to our prompts, not just those in our smaller annotated 830 essay dataset. We then use the MALLET (McCallum, 2002) implementation of LDA to build a topic model of 1,000 topics around each of these sets of essays. This results in what we can think of as a soft clustering of words into 1,000 sets for each prompt, where each set of words represents one of the topics LDA identified being discussed in the essays for that prompt. So for example, the five most important words in the most frequently discussed topic for the military prompt we mentioned above are "man", "military", "service", "pay", and "war".

We also use the MALLET-generated topic model to tell us how much of each essay is spent discussing each of the 1,000 topics. The model might tell us, for example, that a particular essay written on the military prompt spends 35% of the time discussing the "man", "military", "service", "pay", and "war" topic and 65% of the time discussing a topic whose most important words are "fully", "count", "ordinary", "czech", and "day". Since the latter topic is discussed so much in the essay and does not appear to have much to do with the military prompt, this essay should probably get a bad prompt adherence score. We construct 1,000 features from this topic model, one for each topic. Each feature's value is obtained by using the topic model to tell us how much of the essay was spent discussing the feature's corresponding topic. From these features, our regressor should be able to learn which topics are important to a good prompt adherent essay.

6. Manually Annotated LDA Topics A weakness of the LDA topics feature type is that it may result in a regressor that has trouble distinguishing between an infrequent topic that is adherent to the prompt and one that just represents an irrelevant digression. This is because an infrequent topic may not appear in the training set often enough for the regressor to make this judgment. We introduce the manually annotated LDA topics feature type to address this problem.

In order to construct manually annotated LDA topic features, we first build 13 topic models, one for each prompt, just as described in the section on LDA topic features. Rather than requesting models of 1,000 topics, however, we request models of only 100 topics[2]. We then go through all 13 lists of 100 topics as represented by their top ten words, manually annotating each topic with a number from 0 to 5 representing how likely it is that the topic is adherent to the prompt. A topic labeled 5 is very likely to be related to the prompt, where a topic labeled 0 appears totally unrelated.

Using these annotations alongside the topic distribution for each essay that the topic models provide us, we construct ten features. The first five features encode the sum of the contributions to an essay of topics annotated with a number ≥ 1, the sum of the contributions to an essay of topics annotated with a number ≥ 2, and so on up to 5.

The next five features are similar to the last, with one feature taking on the sum of the contributions to an essay of topics annotated with the number 0, another feature taking on the sum of the

[2] We use 100 topics for each prompt in the manually annotated version of LDA features rather than the 1,000 topics we use in the regular version of LDA features because 1,300 topics are not too costly to annotate, but manually annotating 13,000 topics would take too much time.

contributions to an essay of topics annotated with the number 1, and so on up to 4. We do not include a feature for topics annotated with the number 5 because it would always have the same value as the feature for topics ≥ 5.

Features like these should give the regressor a better idea how much of an essay is composed of prompt-related arguments and discussion and how much of it is irrelevant to the prompt, even if some of the topics occurring in it are too infrequent to judge just from training data.

7. Predicted Thesis Clarity Errors In our previous work on essay thesis clarity scoring (Persing and Ng, 2013), we identified five classes of errors that detract from the clarity of an essay's thesis:
Confusing Phrasing. The thesis is phrased oddly, making it hard to understand the writer's point.
Incomplete Prompt Response. The thesis leaves some part of a multi-part prompt unaddressed.
Relevance to Prompt. The apparent thesis's weak relation to the prompt causes confusion.
Missing Details. The thesis leaves out an important detail needed to understand the writer's point.
Writer Position. The thesis describes a position on the topic without making it clear that this is the position the writer supports.

We hypothesize that these errors, though originally intended for thesis clarity scoring, could be useful for prompt adherence scoring as well. For instance, an essay that has a Relevance to Prompt error or an Incomplete Prompt Response error should intuitively receive a low prompt adherence score. For this reason, we introduce features based on these errors to our feature set for prompt adherence scoring[3].

While each of the essays in our data set was previously annotated with these thesis clarity errors, in a realistic setting a prompt adherence scoring system will not have access to these manual error labels. As a result, we first need to predict which of these errors is present in each essay. To do this, we train five maximum entropy classifiers for each prompt, one for each of the five thesis clarity errors, using MALLET's (McCallum, 2002) implementation of maximum entropy classification. Instances are presented to classifier for prompt p for error e in the following way. If a training essay is written in response to p, it will be used to generate a training instance whose label is 1 if e was annotated for it or 0 otherwise. Since error prediction and prompt adherence scoring are related problems, the features we associate with this instance are features $1-6$ which we have described earlier in this section. The classifier is then used to generate probabilities telling us how likely it is that each test essay has error e.

Then, when training our regressor for prompt adherence scoring, we add the following features to our instances. We add a binary feature indicating the presence or absence of each error. Or in the case of test essays, the feature takes on a real value from 0 to 1 indicating how likely the classifier thought it was that the essay had each of the errors. This results in five additional features, one for each error.

5 Evaluation

In this section, we evaluate our system for prompt adherence scoring. All the results we report are obtained via five-fold cross-validation experiments. In each experiment, we use $\frac{3}{5}$ of our labeled essays for model training, another $\frac{1}{5}$ for parameter tuning, and the final $\frac{1}{5}$ for testing.

5.1 Experimental Setup

5.1.1 Scoring Metrics

We employ four evaluation metrics. As we will see below, $S1$, $S2$, and $S3$ are *error* metrics, so lower scores imply better performance. In contrast, PC is a *correlation* metric, so higher correlation implies better performance.

The simplest metric, $S1$, measures the frequency at which a system predicts the wrong score out of the seven possible scores. Hence, a system that predicts the right score only 25% of the time would receive an $S1$ score of 0.75.

The $S2$ metric measures the average distance between a system's score and the actual score. This metric reflects the idea that a system that predicts scores close to the annotator-assigned scores should be preferred over a system whose predictions are further off, even if both systems estimate the correct score at the same frequency.

The $S3$ metric measures the average square of the distance between a system's score predictions and the annotator-assigned scores. The intuition behind this system is that not only should we prefer a system whose predictions are close to the annotator scores, but we should also prefer

[3] See our website at `http://www.hlt.utdallas.edu/~persingq/ICLE/` for the complete list of error annotations.

one whose predictions are not too frequently very far away from the annotator scores. These three scores are given by:

$$\frac{1}{N}\sum_{A_j \neq E'_j} 1, \quad \frac{1}{N}\sum_{i=1}^{N} |A_j - E_j|, \quad \frac{1}{N}\sum_{i=1}^{N} (A_j - E_j)^2$$

where A_j, E_j, and E'_j are the annotator assigned, system predicted, and rounded system predicted scores[4] respectively for essay j, and N is the number of essays.

The last metric, PC, computes Pearson's correlation coefficient between a system's predicted scores and the annotator-assigned scores. PC ranges from -1 to 1. A positive (negative) PC implies that the two sets of predictions are positively (negatively) correlated.

5.1.2 Parameter Tuning

As mentioned earlier, for each prompt p_i, we train a linear regressor r_i using LIBSVM with regularization parameter c_i. To optimize our system's performance on the three error measures described previously, we use held-out validation data to independently tune each of the c_i values[5]. Note that each of the c_i values can be tuned independently because a c_i value that is optimal for predicting scores for p_i essays with respect to any of the error performance measures is necessarily also the optimal c_i when measuring that error on essays from all prompts. However, this is not case with Pearson's correlation coefficient, as the PC value for essays from all 13 prompts cannot be simplified as a weighted sum of the PC values obtained on each individual prompt. In order to obtain an optimal result as measured by PC, we jointly tune the c_i parameters to optimize the PC value achieved by our system on the same held-out validation data. However, an exact solution to this optimization problem is computationally expensive, as there are too many (7^{13}) possible combinations of c values to exhaustively search. Consequently, we find a local maximum by employing the simulated annealing algorithm (Kirkpatrick et al., 1983), altering one c_i value at a time to optimize PC while holding the remaining parameters fixed.

System	$S1$	$S2$	$S3$	PC
Baseline	.517	.368	.234	.233
Our System	.488	.348	.197	.360

Table 4: Five-fold cross-validation results for prompt adherence scoring.

5.2 Results and Discussion

Five-fold cross-validation results on prompt adherence score prediction are shown in Table 4. On the first line, this table shows that our baseline system, which recall uses only various RI features, predicts the wrong score 51.7% of the time. Its predictions are off by an average of .368 points, and the average squared distance between its predicted score and the actual score is .234. In addition, its predicted scores and the actual scores have a Pearson correlation coefficient of 0.233.

The results from our system, which uses all seven feature types described in Section 4, are shown in row 2 of the table. Our system obtains $S1$, $S2$, $S3$, and PC scores of .488, .348, .197, and .360 respectively, yielding a significant improvement over the baseline with respect to $S2$, $S3$, and PC with $p < 0.05$, $p < 0.01$, and $p < 0.06$ respectively[6]. While our system yields improvements by all four measures, its improvement over the baseline $S1$ score is not significant. These results mean that the greatest improvements our system makes are that it ensures that our score predictions are not too often very far away from an essay's actual score, as making such predictions would tend to drive up $S3$, yielding a relative error reduction in $S3$ of 15.8%, and it also ensures a better correlation between predicted and actual scores, thus yielding the 16.6% improvement in PC.[7] It also gives more modest improvements in how frequently exactly the right score is predicted ($S1$) and is better at predicting scores closer to the actual scores ($S2$).

5.3 Feature Ablation

To gain insight into how much impact each of the feature types has on our system, we perform fea-

[4]Since our regressor assigns each essay a real value rather than an actual valid score, it would be difficult to obtain a reasonable $S1$ score without rounding the system estimated score to one of the possible values. For that reason, we round the estimated score to the nearest of the seven scores the human annotators were permitted to assign (1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0) only when calculating $S1$. For other scoring metrics, we only round the predictions to 1.0 or 4.0 if they fall outside the 1.0−4.0 range.

[5]For parameter tuning, we employ the following values. c_i may be assigned any of the values 10^0 10^1, 10^2, 10^3, 10^4, 10^5, or 10^6.

[6]All significance tests are paired t-tests.

[7]These numbers are calculated $\frac{B-O}{B-P}$ where B is the baseline system's score, O is our system's score, and P is a perfect score. Perfect scores for error measures and PC are 0 and 1 respectively.

(a) Results using the $S1$ metric

3	5	1	7	2	6	4
.527	.502	.512	.502	.511	.500	.488
.527	.502	.512	.501	.513	.500	
.525	.508	.505	.505	.504		
.513	.527	.520	.513			
.523	.520	.506				
.541	.527					

(b) Results using the $S2$ metric

2	6	3	1	4	5	7
.356	.350	.348	.350	.349	.348	.348
.351	.349	.348	.348	.348	.347	
.351	.349	.348	.348	.347		
.350	.349	.348	.348			
.358	.351	.349				
.362	.352					

(c) Results using the $S3$ metric

2	6	1	5	4	7	3
.221	.201	.197	.197	.197	.197	.196
.215	.201	.197	.196	.196	.196	
.212	.203	.199	.197	.196		
.212	.203	.199	.197			
.212	.203	.199				
.223	.204					

(d) Results using the PC metric

6	3	2	1	7	5	4
.326	.332	.303	.344	.348	.348	.361
.326	.332	.304	.343	.348	.348	
.324	.337	.292	.345	.352		
.322	.337	.297	.346			
.316	.321	.323				
.218	.325					

Table 5: Feature ablation results. In each subtable, the first row shows how our system would perform if each feature type was removed. We remove the least important feature type, and show in the next row how the adjusted system would perform without each remaining type. For brevity, a feature type is referred to by its feature number: (1) RI; (2) n-grams; (3) thesis clarity keywords; (4) prompt adherence keywords; (5) LDA topics; (6) manually annotated LDA topics; and (7) predicted thesis clarity errors.

ture ablation experiments in which we remove the feature types from our system one-by-one.

Results of the ablation experiments when performed using the four scoring metrics are shown in Table 5. The top line of each subtable shows what our system's score would be if we removed just one of the feature types from our system. So to see how our system performs by the $S1$ metric if we remove only predicted thesis clarity error features, we would look at the first row of results of Table 5(a) under the column headed by the number 7 since predicted thesis clarity errors are the seventh feature type introduced in Section 4. The number here tells us that our system's $S1$ score without this feature type is .502. Since Table 4 shows that when our system includes this feature type (along with all the other feature types), it obtains an $S1$ score of .488, this feature type's removal costs our system .014 $S1$ points, and thus its inclusion has a beneficial effect on the $S1$ score.

From row 1 of Table 5(a), we can see that removing feature 4 yields a system with the best $S1$ score in the presence of the other feature types in this row. For this reason, we permanently remove feature 4 from the system before we generate the results on line 2. Thus, we can see what happens when we remove both feature 4 and feature 5 by looking at the second entry in row 2. And since removing feature 6 harms performance least in the presence of row 2's other feature types, we permanently remove both 4 and 6 from our feature set when we generate the third row of results. We iteratively remove the feature type that yields a system with the best performance in this way until we get to the last line, where only one feature type is used to generate each result.

Since the feature type whose removal yields the best system is always the rightmost entry in a line, the order of column headings indicates the relative importance of the feature types, with the leftmost feature types being most important to performance and the rightmost feature types being least important in the presence of the other feature types. This being the case, it is interesting to note that while the relative importance of different feature types does not remain exactly the same if we measure performance in different ways, we can see that some feature types tend to be more important than others in a majority of the four scoring metrics. Features 2 (n-grams), 3 (thesis clarity keywords), and 6 (manually annotated LDA topics) tend to be the most important feature types, as they tend to be the last feature types removed in the ablation subtables. Features 1 (RI) and 5 (LDA topics) are of middling importance, with neither ever being removed first or last, and each tending to have a moderate effect on performance. Finally, while features 4 (prompt adherence keywords) and 7 (predicted thesis clarity errors) may by themselves provide useful information to our system, in the presence of the other feature types they tend to be the least important to performance as they are often the first feature types removed.

While there is a tendency for some feature types to always be important (or unimportant) regardless of which scoring metric is used to measure per-

Gold	$S1$ .25	.50	.75	$S2$ .25	.50	.75	$S3$ .25	.50	.75	PC .25	.50	.75
2.0	3.35	3.56	3.79	3.40	3.52	3.73	3.06	3.37	3.64	3.06	3.37	3.64
2.5	3.43	3.63	3.80	3.25	3.52	3.79	3.24	3.45	3.67	3.24	3.46	3.73
3.0	3.64	3.78	3.85	3.56	3.70	3.90	3.52	3.65	3.74	3.52	3.66	3.79
3.5	3.73	3.81	3.88	3.63	3.78	3.90	3.59	3.70	3.81	3.60	3.74	3.85
4.0	3.76	3.84	3.88	3.70	3.83	3.90	3.63	3.75	3.84	3.66	3.78	3.88

Table 6: Regressor scores for our system.

formance, the relative importance of different feature types does not always remain consistent if we measure performance in different ways. For example, while we identified feature 3 (thesis clarity keywords) as one of the most important feature types generally due to its tendency to have a large beneficial impact on performance, when we are measuring performance using $S3$, it is the least useful feature type. Furthermore, its removal increases the $S3$ score by a small amount, meaning that its inclusion actually makes our system perform worse with respect to $S3$. Though feature 3 is an extreme example, all feature types fluctuate in importance, as we see when we compare their orders of removal among the four ablation subtables. Hence, it is important to know how performance is measured when building a system for scoring prompt adherence.

Feature 3 is not the only feature type whose removal sometimes has a beneficial impact on performance. As we can see in Table 5(b), the removal of features 4, 5, and 7 improves our system's $S2$ score by .001 points. The same effect occurs in Table 5(c) when we remove features 4, 7, and 3. These examples illustrate that under some scoring metrics, the inclusion of some feature types is actively harmful to performance. Fortunately, this effect does not occur in any other cases than the two listed above, as most feature types usually have a beneficial or at least neutral impact on our system's performance.

For those feature types whose effect on performance is neutral in the first lines of ablation results (feature 4 in $S1$, features 3, 5, and 7 in $S2$, and features 1, 4, 5, and 7 in $S3$), it is important to note that their neutrality does not mean that they are unimportant. It merely means that they do not improve performance in the presence of other feature types. We can see this is the case by noting that they are not all the least important feature types in their respective subtables as indicated by column order. For example, by the time feature 1 gets permanently removed in Table 5(c), its removal harms performance by .002 $S3$ points.

5.4 Analysis of Predicted Scores

To more closely examine the behavior of our system, in Table 6 we chart the distributions of scores it predicts for essays having each gold standard score. As an example of how to read this table, consider the number 3.06 appearing in row 2.0 in the .25 column of the $S3$ region. This means that 25% of the time, when our system with parameters tuned for optimizing $S3$ is presented with a test essay having a gold standard score of 2.0, it predicts that the essay has a score less than or equal to 3.06.

From this table, we see that our system has a strong bias toward predicting more frequent scores as there are no numbers less than 3.0 in the table, and about 93.7% of all essays have gold standard scores of 3.0 or above. Nevertheless, our system does not rely entirely on bias, as evidenced by the fact that each column in the table has a tendency for its scores to ascend as the gold standard score increases, implying that our system has some success at predicting lower scores for essays with lower gold standard prompt adherence scores.

Another interesting point to note about this table is that the difference in error weighting between the $S2$ and $S3$ scoring metrics appears to be having its desired effect, as every entry in the $S3$ subtable is less than its corresponding entry in the $S2$ subtable due to the greater penalty the $S3$ metric imposes for predictions that are very far away from the gold standard scores.

6 Conclusion

We proposed a feature-rich approach to the under-investigated problem of predicting essay-level prompt adherence scores on student essays. In an evaluation on 830 argumentative essays selected from the ICLE corpus, our system significantly outperformed a Random Indexing based baseline by several evaluation metrics. To stimulate further research on this task, we make all our annotations, including our prompt adherence scores, the LDA topic annotations, and the error annotations publicly available.

References

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with E-rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3).

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: A library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent smeantic analysis. *Journal of American Society of Information Science*, 41(6):391–407.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. *International Corpus of Learner English (Version 2)*. Presses universitaires de Louvain.

Derrick Higgins and Jill Burstein. 2007. Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics*.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Human Language Technologies: The 2004 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 185–192.

Derrick Higgins, Jill Burstein, and Yigal Attali. 2006. Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, 12(2):145–159.

David Jurgens and Keith Stevens. 2010. The S-Space package: An open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations*, pages 30–35.

Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, pages 103–106.

Scott Kirkpatrick, C. D. Gelatt, and Mario P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220(4598):671–680.

Thomas K. Landauer and Susan T. Dutnais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

Thomas K. Landauer, Darrell Laham, and Peter W. Foltz. 2003. Automated scoring and annotation of essays with the Intelligent Essay Assessor$^{\text{TM}}$. In *Automated Essay Scoring: A Cross-Disciplinary Perspective*, pages 87–112. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Annie Louis and Derrick Higgins. 2010. Off-topic essay detection using short prompt texts. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–95.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. `http://mallet.cs.umass.edu`.

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.

Robert Parker, David Graf, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2009. *English Gigaword Fourth Edition*. Linguistic Data Consortium, Philadelphia.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239.

Magnus Sahlgren. 2005. An introduction to random indexing. In *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*.

Mark D. Shermis and Jill C. Burstein. 2003. *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Mark D. Shermis, Jill Burstein, Derrick Higgins, and Klaus Zechner. 2010. Automated essay scoring: Writing assessment and instruction. In *International Encyclopedia of Education (3rd edition)*. Elsevier, Oxford, UK.

Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pages 412–420.