Scoring Persuasive Essays Using Opinions and their Targets

Noura Farra
Columbia University
New York, NY 10027
noura@cs.columbia.edu

Swapna Somasundaran
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08540
ssomasundaran@ets.org

Jill Burstein
Educational Testing Service
660 Rosedale Road
Princeton, NJ 08540
jburstein@ets.org

Abstract

In this work, we investigate whether the analysis of opinion expressions can help in scoring persuasive essays. For this, we develop systems that predict holistic essay scores based on features extracted from opinion expressions, topical elements, and their combinations. Experiments on test taker essays show that essay scores produced using opinion features are indeed correlated with human scores. Moreover, we find that combining opinions with their targets (what the opinions are about) produces the best result when compared to using only opinions or only topics.

1 Introduction

In a persuasive essay, test takers are asked to take a stance on a given topic and to write an essay supporting their stance. Consider for example the following essay question, also known as the prompt:

"A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught."

Test takers have to write an essay describing whether they agree or disagree with the given prompt, using language expressing clear opinions. The scores for these essays are typically influenced by many factors, such as grammar, spelling errors, style and word usage, as well as the persuasiveness component: how well does the writer argue in favor of that writer's position on the subject? In this work, we try to tackle this last aspect, by studying how the expression of opinions influences the scores of expert human graders.

A number of essay scoring systems which rely on Natural Language Processing methods have been developed for automatically scoring persuasive essays, most notably (Page, 1966; Foltz et al., 1999; Burstein, 2003; Rudner and Liang, 2002; Attali and Burstein, 2006). The principal features for automatic essay-scoring have traditionally been based on grammar, usage, mechanics, and style, and have additionally included content-based features such as discourse and topic, as in Attali and Burstein (2006). These kind of features have been shown to have very strong performance in scoring holistic essay scores, and are very highly correlated with expert human scores (Bridgeman et al., 2012). However, in spite of their powerful predictive capability, these automated scoring systems have been criticized for limited coverage of the construct (Deane, 2013; Ben-Simon and Bennett, 2007; Williamson et al., 2012).

Our work addresses this concern by developing features specific to the persuasive construct. Incorporating knowledge of the persuasiveness factor into essay-scoring models can allow us to add features directly related to the scoring construct and to the writing task, which typically asks test takers to state and defend their opinion. Additionally, our linguistically motivated features encode intuitions which could allow for interpretable, useful and explicit feedback to students, test takers and educators regarding the persuasive aspect of the essays.

We build simple essay scoring systems which incorporate persuasiveness by engineering features based on the analysis of opinions expressed in the

essay and whether these opinions are being expressed about relevant topics. Specifically, the developed systems are based on simple features capturing (1) Opinion expressions, (2) Topics, and (3) Opinion-Target pairs which combine opinions with what they are about. We consider different methods for finding opinion-target pairs, and extract features which assess if the opinions in the essay are indeed relevant to the persuasion, and if the stance taken in the essay is consistently maintained. We find that our system predictions are indeed correlated with human scores, and the system using opinion-target information is the best.

The rest of the paper is organized as follows. Section 2 describes related work. In Section 3 we describe how we find opinions, topics and opinion-targets in essays, and Section 4 describes the features used accordingly to build the persuasive essay scoring systems. Section 5 describes our experiments, Section 6 presents analysis, and we conclude in Section 7.

2 Related Work

Automated essay scorers rely on a number of features based on grammar, usage, and content. Notable systems are Project Essay Grader (Page, 1966) which grades essays based on fluency and grammar; IEA (Foltz et al., 1999) which uses both content and mechanics-based features and relies on LSA word vector representations; e-rater (Burstein, 2003; Attali and Burstein, 2006) which combines syntactic, discourse, and topical components; and the Bayesian Essay Test Scoring System (Rudner and Liang, 2002). For a comprehensive description of these automatic essay scoring systems, the reader is referred to Dikli's survey (Dikli, 2006). Recently, there have been attempts to incorporate more non-traditional features for essay scoring; such as Beigman Klebanov and Flor (2013) who examined the relationship between the quality of essay writing and the use of word associations, and accordingly built a system to improve the prediction of holistic essay scores; and Somasundaran et al. (2014) who predicted discourse coherence quality of persuasive essays using lexical chaining techniques.

There has also been work on the study of argumentation in essays. Stab and Gurevych (2014a) propose an annotation scheme and a corpus for annotating different components of arguments and argumentative relations in persuasive essays. In addition, Stab and Gurevych (2014b) propose models for automatically recognizing arguing components in persuasive essays, and identifying whether the arguing components reflect support or non-support. Madnani et al. (2012) proposed a system for distinguishing the "shell" organizational elements of arguing expressions from actual argumentative content. Beigman Klebanov et al. (2013a) identify sentence-level sentiment in persuasive essays by considering the sentiment of multi-word expressions. In our work, we have used lexicons for identifying opinion expressions; however, our methods can be augmented by using such systems.

Opinion analysis has been applied to a number of natural language processing tasks and domains, such as sentiment in movie reviews (Turney, 2002; Pang and Lee, 2004), product reviews (Hu and Liu, 2004; Liu et al., 2005), social media (Go et al., 2009; Agarwal et al., 2011; Bollen et al., 2011), news, blogs, and political and online debates (Mullen and Malouf, 2006; Godbole et al., 2007; Somasundaran and Wiebe, 2009). The use of opinion and sentiment information to predict holistic essay scores, however, has remained unstudied.

Targets of sentiment have been studied in the form of finding features in product reviews (Qiu et al., 2011; Liu et al., 2014) and for classifying online debates (Somasundaran and Wiebe, 2010). The recent 2014 SemEval Task on aspect-based sentiment analysis (Pontiki et al., 2014) was concerned with identifying targets of sentiment in reviews of restaurants and laptops. Jiang et al. (2011) and (Dong et al., 2014) have explored target-dependent classification of sentiment in Twitter. In our work, we take a simple approach to finding targets of opinion expressions, since our focus is on determining whether opinion analysis is useful for persuasive essay scoring, even when using approximate opinion-targets.

3 Opinions and Topics in Persuasive Essays

Intuitively, well-written persuasive essays will clearly state the opinion of the writer and build support for their stance by evoking ideas and concepts

that are relevant for the argument. Thus, we investigate the role of opinions, topics, and their interactions in determining overall persuasive essay scores.

3.1 Opinion Expressions

We consider two distinct types of opinions important for persuasion: Sentiment and Arguing. Much work has been done on defining these two types of opinions (Wilson, 2008; Ruppenhofer et al., 2008). We focus on sentiment and arguing because we expect these types of expressions to be common in essays which require persuasion.

Sentiment Expressions Sentiment expressions reveal a writer's judgments, evaluations and feelings, and are likely to be employed to express a preference for a particular position, or to point out the shortcomings of an alternative position. In the following sentence, we see the sentiment expression in bold, and the target in brackets. The writer has a positive evaluation ("learning the most") of teachers' encouragement.

Example 1
At school, I always ***learned the most*** *from [teachers who encouraged me].*

Arguing Expressions Arguing expressions reveal the writer's beliefs and strong convictions, and is seen in the form of reasoning, justification, strong assertions, emphasis, and use of imperatives, necessities and conditionals (Wilson, 2008; Ruppenhofer et al., 2008). In the following sentence, we see the arguing expression in bold, and the target in brackets. Here, the writer clearly emphasizes the position taken with respect to the topic.

Example 2
For these reasons, I claim with confidence *that [excellent knowledge of the subject being taught is secondary to the teacher's ability to relate well with their students].*

We expect that persuasive essays where test takers clearly state their opinions will get better scores than the ones that do not.

3.2 Topical Elements

We define topical elements as words or concepts that are relevant to the topic of the essay, and which usually get invoked in the process of stance-taking. They essentially correspond to "common topics" that test takers are expected to write about when presented with a prompt. For example, given the prompt in Section 1, while words which appear in the prompt (*prompt words*), such as 'teacher', 'student', 'subject', and 'knowledge' are naturally expected, we also expect general topical words such as 'class' and 'school' to occur in response essays. Intuitively, we would expect essays containing sufficient topical elements to get higher scores.

3.3 Opinion Relevancy and Consistency

We expect that well-written persuasive essays will not only express opinions and evoke common topics, but in fact *express opinions about relevant topical elements*. Specifically, we hypothesize that the opinions should be about artifacts relevant to the theme of the essay, and not about irrelevant topics. For example, for the prompt described in Section 1, it is important that there be opinions expressed about topics such as teachers, school, learning, and so on. In addition, the essay also has to reflect a clear attempt at persuasion and stance-taking in relevance to the prompt statement and the underlying theme. We call this *opinion relevancy*.

We also expect that once a stance is taken, there should be sufficient elaboration and development such that the stance is consistently maintained. We hypothesize that essays where test takers support their stance will achieve higher scores than essays where they vacillate between options (for instance, in the example prompt in Section 1, the test taker is unable to decide whether the teachers' ability to relate well is more important or not). We call this *opinion consistency*.

These expectations are more stringent than those discussed in Sections 3.1 and 3.2, and we expect that a scoring system which captures these requirements will likely perform better.

4 Essay Scoring Systems

In order to test the intuitions described in Section 3, we build essay scoring systems based on features extracted from opinions, topics, and opinion-target pairs. We construct three separate systems:

1. **Opinion** This system uses features based on

opinion expressions only, and tests whether expressing opinions influences the essay score.

2. **Topic** This system uses features based on topical expressions alone, and tests whether evoking relevant topics associated with the prompt influences the essay score.

3. **Opinion-Target** This system uses features based on the combination of opinions and their targets, with the goal of measuring opinion relevancy and consistency. This system tests how well the essay score can be predicted based on the interactions of opinions with their targets.

4.1 Opinion System

4.1.1 Finding sentiment and arguing expressions

In order to find sentiment expressions in the essays, we used a combination of two lexicons: the MPQA subjectivity lexicon (Wilson et al., 2005) (Lexicon 1), and the sentiment lexicon developed by Beigman Klebanov et al. (2013b) (Lexicon 2). Each of these lexicons provides for each word, a sentiment polarity (positive, negative, or neutral), along with an indicator of sentiment intensity: strongly or weakly subjective (Lexicon 1) or a probability distribution over the polarity (Lexicon 2). For Lexicon 2, the sentiment polarity for a word is obtained by choosing the polarity corresponding to the highest probability score.

For identifying arguing expressions in the essays, we used an Arguing lexicon developed as part of a discourse lexicon (Burstein et al., 1998). The original lexicon has annotations for different types of expressions, including claim initializations and development, structure, rhetoric, among others. For this work, since we are concerned with arguing expressions that specifically reveal support for or against an idea, we used only lexicon entries which label an expression as *arguing-for* or *arguing-against*. For instance, in Example 2, the writer argues *for* teachers' ability to relate well with their students.

4.1.2 Features

We extract three (global) features based on opinion expressions:

1. The total count of sentiment words in the essay that are found in Lexicon 1 and Lexicon 2 respectively. These counts also include words with subjective neutral polarity.

2. The total count of words in the essay found in the arguing lexicon.

4.2 Topic System

4.2.1 Finding topical elements

In order to determine topical elements, we compute topic signatures (Lin and Hovy, 2000) over each prompt. Topic Signatures are defined as

$$TS = \{topic, signature\}$$

$$= \{prompt, < (t_1, w_1), (t_2, w_2)...(t_n, w_n) >\}$$

where topic in our case is the prompt. The signature comprises a vector of related terms, where each term t_i is highly correlated with the prompt with an association weight w_i.

For each prompt, we use a corpus of high-scoring essays (that was separate from our training and testing data) to find its topic signature[1]. The top 500 words with the highest signature scores are considered as topical elements for that prompt.

For a given essay, we annotate all prompt words and topic signature words. Note that our topical elements consist entirely of unigrams, but this need not necessarily be the general case (as seen in examples 1 and 2); extending the scope of topical elements to multi-word concepts is a direction for future work.

4.2.2 Features

Based on the prompt words and topical words we extract the following features:

1. The total count of topical words in the essay

2. The total count of actual prompt words

We distinguished between prompt words and topical words as the former measures whether the essay is clearly responding to the prompt, while the latter measures if thematic elements are indeed present in the essay and its arguments.

[1]We used the topic signatures code provided at http://homepages.inf.ed.ac.uk/alouis/topicS.html

4.3 Opinion-Target System

The opinion-target system relies on the extraction of features based on the opinion-target pairs found in the essay. The first step towards building this system is the identification of opinion-target pairs, after which we construct features which measure opinion relevancy and consistency. We investigated simple heuristic-based approaches for finding targets of opinions, described below.

4.3.1 Finding sentiment-target pairs

We explored three methods for finding targets of sentiment expressions. Our simplest approach, ***all-sentence***, finds all sentiment expressions in the sentence and assumes that all words are targets of each expression. This method introduces some noise as it results in some words becoming targets of multiple opinions with possibly conflicting polarities.

Our second approach, ***resolve-sentence***, resolves the sentiment at the sentence-level to a single polarity, as in (Somasundaran and Wiebe, 2010), and then assumes that all nouns, verbs, and adjectives in the sentence are targets. If we consider Example 1, suppose the sentiment of the sentence is resolved to positive, (due to the positive opinion words **learned**, **most**, and **encouraged**) then the words *school*, *teachers*, *learned* and *encouraged* would be considered as the targets. Ideally, we would like only the words *teachers* and *encouraged* to be targets. We note here that in our task a target can actually be a sentiment-containing word such as *encouraged*, which is why we don't disregard sentiment words when finding targets.

Our third method, ***resolve-constituent***, resolves sentiment at the syntactic constituent level instead of the sentence level, and assumes that all nouns, verbs and adjectives in the constituent phrase are targets. For obtaining the phrases, we used the regular expression parser from the Python NLTK toolkit (Bird, 2006) to define a custom grammar that describes noun, verb, and prepositional phrases. The parser uses regular expression rules for grouping words together based on their part of speech tags. Considering our example with this scenario, the phrases *"at school"*, *"I always"*, and *"learned the most from..."* will be considered separately in our grammar, so the word *school* will likely not end up as a target.

To resolve sentence-level or constituent-level polarity, we use a heuristic that aggregates polarity scores from both sentiment lexicons, and chooses the final polarity corresponding to the word with the maximum sentiment intensity.

While these methods are not exact and may lead to over-generating targets, for the purposes of this work (which is to determine whether a basic opinion system is effective in predicting essay scores), we are more interested in high recall of targets then high precision because they will be aggregated at the essay level.

4.3.2 Finding arguing-target pairs

For resolving arguing-target pairs, we use the ***all-sentence*** method. Resolving the dominant arguing polarity at the sentence level would be less straightforward than for sentiment, given that the argument lexicon does not provide us with scores for arguing intensity. Moreover, arguing targets are generally longer (Ruppenhofer et al., 2008); we would expect their spans to extend beyond constituent phrases. Finally, we observed that sentences generally do not contain multiple arguing expressions, thus alleviating the problem of spurious combinations.

4.3.3 Features

The features for the opinion-target system are based on measuring relevancy and consistency of opinions.

Relevancy Relevancy is measured by taking into account how many opinions (or proportion of opinions) are about prompt or topical elements. These include global engineered features as follows:

1. The number of times that topical elements (topic and prompt words) appear as a target in the essay's opinion-target pairs.

2. The ratio of topic targets (opinion-topic pairs) to all opinion-target pairs.

We distinguished between topic targets and prompt targets and also between sentiments which included subjective neutral versus only positive or negative sentiments. We had separate features for sentiment-target pairs and arguing-target pairs, resulting in 12 relevancy features.

Consistency Consistency is measured by determining how often the writer switches opinion polarity when referring to the same target. The consistency features included the following:

1. A binary feature indicating the presence of a reversal ('flip') of opinion towards any target.
2. The number of unique targets which get flipped.
3. The proportion of all flips where the target is a topical element.
4. The proportion of all topical elements which get flipped.
5. Statistics including max, mean, and median number of flips over all targets.

We also separated sentiment-target and arguing-target features, as well as prompt word targets and topic word targets, resulting in a total of 18 consistency features. We note that these features can only capture an approximate picture of consistency, because it is well-known (Aull and Lancaster, 2014) that mature writers tend to state and describe opposing arguments as well as their own.

5 Experiments

5.1 Data

The data used for this study consists of 58K essays, covering 19 different prompts, obtained from the TOEFL® (Test of English as a Foreign Language) persuasive writing task which pertains to essays written by undergraduate and graduate school applicants who are non-native English speakers. All essays are *holistically* scored by experts on an integer scale 0-5, with score point 5 assigned to excellent essays. Detailed studies of human-human agreement for this dataset can be found in Bridgeman et al. (2012). The holistic scores are assigned to essays based on English proficiency, and account for the quality of (and errors in) grammar, language use, mechanics, style, in addition to quality of the persuasive task. The scores for these essays are thus influenced by a number of factors other than the quality of persuasion (essays can get a low score if they use incorrect grammar, even if they make good persuasive arguments). However, we would like to test the extent to which our hypothesis holds when predicting such holistically graded essays.

We split this dataset randomly into a training and test set with proportions of 80% (46,404 essays) and 20% (11,603 essays) respectively. Table 1 shows the score distribution of essays for different score points, in the training and test set respectively. We note that the distribution of scores is unbalanced, with essays having scores 3, 4, and 2 occupying the majority in that order.

5.2 Setup

We modeled the system with a number of different regression learners, which have generally been shown to do well on the essay scoring task. We used a number of learners available from the Python Scikit-learn toolkit (Pedregosa et al., 2011) and the Scikit-learn-Laboratory (Blanchard et al., 2013): the Logistic Regression classifier (**LO**), which uses 6-way classification to predict integer essay scores in the range 0-5, the Linear Regression learner (**LR**), which predicts real-valued scores that are rounded to integers, and the Rescaled Linear Regression learner (**RR**), which rescales the predicted scores based on the training data distribution. Given an input essay, the learners predict essay scores in the range 0-5, based on the features described in Section 4.

We considered a number of evaluation metrics to test for the predictive ability of opinion, topic, and opinion-target information in scoring the essays. We tested if our proposed systems' score predictions are correlated with human scores, by computing the human score correlation *(HSC)* using Pearson's coefficient. As essay length is highly correlated with the human score (Attali and Burstein, 2006; Chodorow and Burstein, 2004), and as many of our features are based on counts, they can be influenced by essay length; so we also compute the partial correlations *(HSC-Part)* accounting for length, by partialing out the length of the essay in words. For measuring the performance of the system, we report Accuracy, F-measure – where we computed the weighted f-score (*F-w*) over the six score points – and Quadratic Weighted Kappa (QWK) (Cohen, 1968), which is the standard metric for essay scoring. Accuracy and F-measure are standard NLP metrics and provide a direct, interpretable measure of system performance which reflects the precision and recall of different

	Train		Test	
Score	**# Essays**	**Distribution(%)**	**# Essays**	**Distribution(%)**
0	278	0.6	65	0.6
1	1,177	2.5	304	2.6
2	6,812	14.7	1,668	14.4
3	27,073	58.3	6,714	57.9
4	8,902	19.2	2,305	19.9
5	2,162	4.7	546	4.7
Total	46,404	100	11,602	100

Table 1: Score distribution of essays in our dataset

score points. QWK corrects for chance agreement between the system prediction and the human prediction, and it also takes into account the extent of the disagreement between labels.

We compared all systems to a baseline *Length*, that predicts an essay score based solely on the length of the essay in words. Due to the strong correlation between length and essay scores, we consider this to be a strong (albeit simple) baseline. Another simple baseline was *Majority*, which always predicts the majority class (score point 3).

5.3 Results

We evaluate each of the Opinion, Topic, and Opinion-Target systems separately, to determine the effect of each and to test the hypotheses described in Section 3.

For the Opinion-Target system, we found that both the *resolve-sentence* and *resolve-constituent* methods (Section 4.3.1) consistently and significantly outperformed the *all-sentence* approach. The difference between *resolve-sentence* and *resolve-constituent* was not statistically significant. Thus we report results for the *resolve-sentence* approach, which had the best performance.

Table 2 shows the results of the correlation experiments for each system and for each of the three learners. We find that predictions based on opinions and topics are positively correlated with human scores. Furthermore, combining opinions with their targets produced the best correlation for all learners, with the Regression predictors achieving the best results (0.59). This result supports our hypothesis that the relevancy and consistency of opinions is more informative than simply measuring whether opinions are expressed or topics are invoked. Our results are particularly promising when considering the fact that the features only capture the persuasivenes component of the holistic score. As noted previously, the holistic score of this English proficiency test depends on a number of factors such as grammar, language usage, mechanics and style: effective persuasion is but one aspect of the score.

When partialing out the effect of length, we find that the partial correlation scores drop, but are still strong for the Opinion-Target system (0.23 for LR and RR). This drop is unsurprising, as human scores are influenced by the length of the essay, and so are the count-based features. We also note that the correlation results differ between the linear regression predictors (LR and RR) and the LO classifier. This is also expected because LR and RR report the correlation of real numbers while LO reports the correlation of an integer classification.

Next, Table 3 reports the performance for all systems in terms of Accuracy, F-measure, and QWK. For each system and for each metric, we present the results from all learners. For each learner, the results comparing the opinion-target system with the baselines are all statistically significant; we computed significance for each of the three metrics using the bootstrap sampling method described in (Berg-Kirkpatrick et al., 2012) with a subset size of $n = 11,000$ and $b = 10^4$ subset samples.

When considering the Linear Regression and Logistic Regression classifiers, we observe that the Opinion-Target system significantly outperforms the baselines and our other systems across all metrics. The Opinion-Target system also achieves the best QWK score over all systems. On the other hand,

	LR		RR		LO	
System	**HSC**	**HSC-*Part***	**HSC**	**HSC-*Part***	**HSC**	**HSC-*Part***
Opinion	0.29	0.10	0.28	0.10	0.33	0.07
Topic	0.29	0.081	0.30	0.086	0.33	0.15
Opinion-Target	**0.59**	**0.23**	**0.59**	**0.23**	**0.43**	**0.18**

Table 2: Correlation of System Predictions with Human Scores. The best system correlation is shown in **bold**.

it is outperformed by the majority and length baselines for Accuracy and F-measure when using the Rescaled Regression predictor. We suspect that the rescaling of the training data by the RR learner significantly alters the scores. We note for example that when using the LR predictor, all the predictions of the Length system fall in the range (3,3.5), and hence get rounded to score 3; thus it always predicts the majority class (3) and essentially functions as a majority predictor. This explains why it has a QWK of 0 and an F-measure equal to the majority baseline. On the other hand, when the data is rescaled to match the training data, the Length system predictions are stretched to match the distribution of scores observed in the training data, and the percentage of score 3 predictions drops to 56% of predictions, while the percentage of score 4 predictions jumps to 20%, and the recall of all other score points increases. This makes sense when considering that length predictions are highly correlated with human scores, and thus its linear regression predictions will be correlated with the human score irrespective of the training data distribution. On the other hand, the Opinion-Target system is able to produce more predictions across different score labels even when the test data is not rescaled.

6 Feature Analysis

To explore the impact of the different opinion-target features on essay scores, we tested the performance of individual features in predicting scores for our test set. We evaluated the features based on both accuracy and QWK. Table 4 shows the results, where we show the top 15 features ranked in order of QWK.

We observe that the best feature is the frequency of topic-relevant sentiment-target pairs, counting only positive and negative words (as opposed to neutral lexicon words). This indicates that expressing sentiment clearly in favor of or against the topical words is important for persuasion in this data.

We notice that most of the top-scoring features are sentiment rather than arguing features. This may be because our sentiment-target pairing system was more concise and precise than the arguing-target system. Additionally, our arguing features include strong modal words such as 'must', 'clearly' and 'obviously'. Previous research has shown that while writers with intermediate proficiency use such terms, they are used less often by the most proficient writers (Vázquez Orta and Giner, 2009; Aull and Lancaster, 2014). Thus it is possible that these features would not be found in essays with very high scores, whose writers would likely employ more subtle and sophisticated forms of argumentation.

We also observed that count-based features tend to perform better than their ratio-based counterparts, except in the case of the prompt word adherence feature (10), where the ratio feature actually outperforms the frequency feature (12). It is likely that the length effect is at play here. However, the fact that significant correlations exist, even after accounting for length (as seen in Table 2), indicates that these features are capturing meaningful information.

7 Conclusions

In this work, we investigated features for improving the persuasive construct coverage of automated scoring systems. Specifically, we explored the impact of using opinion and topic information for scoring persuasive essays. We hypothesized that essays with high scores will show evidence of clear and consistent stance-taking towards relevant topics. We built systems using features based on opinions, topics, and opinion-target pairs, and performed experiments with holistically scored data using different learners.

Our results are encouraging. We found that, in spite of the fact that the persuasive component is one of many factors influencing the holistic score, our system's predictions were positively correlated with the essay scores. Moreover, combining opin-

	LR			RR			LO		
System	**Acc%**	**F-*w*%**	**QWK**	**Acc%**	**F-*w*%**	**QWK**	**Acc%**	**F-*w*%**	**QWK**
Majority	57.87	42.43	_ _ _	**57.87**	42.43	_ _ _	57.87	42.43	_ _ _
Length	57.87	42.43	0	53.88	**54.11**	0.553	58.39	44.53	0.141
Opinion	57.85	43.84	0.032	41.33	42.47	0.275	58.38	44.71	0.169
Topic	58.39	43.83	0.0013	40.75	42.00	0.284	58.39	43.83	0.168
Opinion-Target	**[61.01]**	**[57.02]**	**0.45**	53.28	53.75	**[0.554]**	**60.15**	**48.26**	**0.28**

Table 3: Performance of different systems measured by accuracy, weighted F-score, and QWK. The best system for each learner is in **bold**. The best overall system for each metric is bracketed. For each learner, the results comparing the opinion-target system are statistically significant ($p < 0.0015$ for comparing with Length for RR, $p < 0.0001$ otherwise).

Feature Name	**Desc**	**QWK**	**Acc %**
(1) Freq of pos and neg sentiment-topic pairs	Rel	**0.418**	33.3
(2) Freq of all sentiment-topic pairs	Rel	0.411	32.1
(3) Freq of arguing-topic pairs	Rel	0.273	25.7
(4) Mean # of sentiment flips	Con	0.205	**37.3**
(5) Unique # of sentiment flips	Con	0.204	19.2
(6) Ratio of sentiment-topic flips to all topic words	Con	0.202	19.7
(7) Ratio of pos and neg sentiment-topic pairs to all sentiment-target pairs	Rel	0.197	22.9
(8) Freq of all sentiment-prompt pairs	Rel	0.185	21.8
(9) Median # of sentiment flips	Con	0.178	21.6
(10) Ratio of pos and neg sentiment-prompt pairs to all sentiment-target pairs	Rel	0.165	23.5
(11) Max # of sentiment flips	Con	0.162	19.8
(12) Freq of pos and neg sentiment-prompt pairs	Rel	0.160	24.4
(13) Freq of arguing-prompt pairs	Rel	0.159	20.9
(14) Flip presence	Con	0.155	21.2
(15) Ratio of sentiment-topic flips to all sentiment-target flips	Con	0.150	20.7

Table 4: Feature Analysis. A feature is described as 'Rel' if it assesses relevancy and 'Con' if it assesses consistency. Sentiment-topic, Arguing-topic, Sentiment-prompt, and Arguing-prompt refer to the opinion-target pairs where the target is a topic word or prompt word respectively. Ratios are all measured with respect to total number of sentiment-target pairs or arguing-target pairs, except for feature (6) where the ratio is measured against all topic words. This experiment was performed using the Logistic Regression (LO) classifier.

ions with their targets, and assessing their relevancy and consistency, resulted in a higher correlation than using only topics or only opinions. We also found that, for most learners, the opinion-target predictor performs better than a system which predicts essay scores based on the length of the essay.

Our initial feature analysis shows that opinion-target features seem to reasonably reflect the importance of persuasion information found in the essays, and that the co-occurrence of polar sentiment words with topic targets is particularly important.

Having demonstrated the viability of the approach using simple methods, our next step is to explore more precise ways of finding opinion-target pairs and topical elements, including resolving negations and co-references, exploring syntactic dependencies, as well as targets spanning multiple words. We also plan to validate our experiments with data from different writing exams. Future work will also involve exploring ways to combine our features with those of other automated scoring systems – such as grammar, usage and mechanics – in order to obtain more robust holistic scoring.

Acknowledgments

We would like to thank Binod Gyawali for running additional experiments. We thank all the reviewers for their valuable feedback and comments.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38. Association for Computational Linguistics.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Laura L Aull and Zak Lancaster. 2014. Linguistic markers of stance in early and advanced academic writing a corpus-based comparison. *Written Communication*, 31(2):151–183.

Beata Beigman Klebanov and Michael Flor. 2013. Word association profiles and their use for automated scoring of essays. In *ACL (1)*, pages 1148–1158.

Beata Beigman Klebanov, Jill Burstein, and Nitin Madnani. 2013a. Sentiment profiles of multiword expressions in test-taker essays: The case of noun-noun compounds. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(3):12.

Beata Beigman Klebanov, Nitin Madnani, and Jill Burstein. 2013b. Using pivot-based paraphrasing and sentiment profiles to improve a subjectivity lexicon for essay data. *TACL*, 1:99–110.

Anat Ben-Simon and Randy Elliot Bennett. 2007. Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment*, 6(1).

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005. Association for Computational Linguistics.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Daniel Blanchard, Michael Heilman, and Nitin Madnani. 2013. Scikit-learn laboratory.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *ICWSM*.

Brent Bridgeman, Catherine Trapani, and Yigal Attali. 2012. Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 25(1):27–40.

Jill Burstein, Karen Kukich, Susanne Wolff, Ji Lu, and Martin Chodorow. 1998. Enriching automated essay scoring using discourse marking. In *Workshop on Discourse Relations and Discourse Marking*. ERIC Clearinghouse.

Jill Burstein. 2003. The e-rater® scoring engine: Automated essay scoring with natural language processing.

Martin Chodorow and Jill Burstein. 2004. Beyond essay length: Evaluating e-rater®'s performance on toefl® essays. *ETS Research Report Series*, 2004(1):i–38.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

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

Semire Dikli. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 49–54.

Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2).

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. *ICWSM*, 7.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 151–160. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization.

In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 495–501. Association for Computational Linguistics.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

Kang Liu, Liheng Xu, and Jun Zhao. 2014. Extracting opinion targets and opinion words from online reviews with graph co-ranking.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28. Association for Computational Linguistics.

Tony Mullen and Robert Malouf. 2006. A preliminary investigation into sentiment analysis of informal political discourse. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 159–162.

Ellis B Page. 1966. The imminence of... grading essays by computer. *Phi Delta Kappan*, pages 238–243.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Maria Pontiki, Haris Papageorgiou, Dimitrios Galanis, Ion Androutsopoulos, John Pavlopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27.

Lawrence M Rudner and Tahung Liang. 2002. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2).

Josef Ruppenhofer, Swapna Somasundaran, and Janyce Wiebe. 2008. Finding the sources and targets of subjective expressions. In *LREC*.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 226–234. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays.

Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

Ignacio Vázquez Orta and Diana Giner. 2009. Writing with conviction: The use of boosters in modelling persuasion in academic discourses.

David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Theresa Ann Wilson. 2008. *Fine-grained subjectivity and sentiment analysis: recognizing the intensity, polarity, and attitudes of private states*. ProQuest.