Automatic Essay Grading Using Extractive Summarization Techniques

Anand Natarajan, Joshua Wang, and Ivan Zhang

December 10, 2010

1 Introduction

Evaluation of student-written essays is a common task in education. In particular, the graders of standardized tests often have to grade large volumes of essays in a short period of time. In such situtations, it is also particularly important that all essays be evaluated on the same standard to ensure fairness. Naturally, many researchers have tried to automate this process, with some success. Indeed, the Educational Testing Service (ETS), which administers most standardized tests in the United States, uses an automated system in conjunction with one human grader on the GRE essay.

Essay grading is a problem that touches on several different areas of Natural Language Processing. For example, in order to evaluate grammatical correctness of essays one might apply techniques from parsing, and in order to evaluate the logical organization of ideas, one might use techniques from computational discourse analysis.

In this project, we will develop an essay evaluation system targeted towards SAT essays. We will focus on evaluating the logical structure of the essays using algorithms from extractive summarization.

2 SAT Essay Grading

The essay portion on the SAT requires the student to write a short (1-2 page) persuasive essay in response to an essay prompt. Below is a sample prompt from the November 2010 SAT [2].

> Think carefully about the issue presented in the following excerpt and the assignment below.
>
> We are very individually oriented. We see everything in terms of personal independence, personal pleasure, personal fulfillment. "Do your own thing," we say. The idea that people can actually do things for someone or something else – a community, a school, or any other group – is lost. It is important to realize, however, that all people are interconnected. We cannot survive without each other.
>
> Adapted from Willard Gaylin in Bill Moyers, A World of Ideas
>
> **Assignment:** Do people put too much emphasis on doing things by and for themselves? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.

As in the example above, most prompts contain a quotation from a famous author about an abstract theme, followed by a question asking the student to agree or disagree with the author's take on the theme. Students are given 25 minutes to write their essays.

The essays are evaluated on a discrete scale from 1-6, according to a standard grading rubric. The official grading rubric for the essay [3] focuses on the following five criteria:

1. Persuasiveness of the essay's argument

2. Logical structure/coherence of the essay
3. Varied vocabulary usage
4. Varied sentence structure
5. Correct "mechanics" (grammar, punctuation, spelling)

We attempted to capture these criteria with our choice of features, focusing on criteria 2 and 3.

3 Features

3.1 Basic Features

For our baseline model, we converted essays into vectors of word counts, as in multinomial Naive Bayes. As additional features, we used the total number of words, the number of unique words, and the average word length as features. Of these, the first seems intuitively correlated to a better-developed argument, and the latter two correlated with varied and sophisticated vocabulary usage.

3.2 Discourse Analysis

When we evaluate an essay for its logical structure, the quality we are looking for is "coherence." Most computational discourse techniques for measuring coherence are based on a few common-sense intuitions, which are described in [5] as follows:

1. The parts of the discourse should be ordered correctly, with transitional language to indicate "cohesion."
2. There are logical "relations" between parts of the discourse (examples: "evidence", "antithesis", "concession:").
3. Coherent discourses stay focused on particular entities.

SAT essays are especially interesting from the coherence standpoint, because they have a certain formulaic structure. Most good SAT essays have an introductory paragraph that states the author's opinion on the question raised in the prompt. Then there are two or three body paragraphs, each presenting an example of some kind in support of the thesis. The first sentence of each body paragraph is usually a topic sentence that identifies how the example addresses the thesis. Finally, there is a concluding paragraph that essentially restates the thesis.

Previous work on essay grading has shown that this structure can be a useful feature for grading essays. Indeed, the ETS autograder uses discourse parsing algorithms to identify sentences as introductory sentences, topic sentences, or supporting examples [1]. However, these algorithms required a large training set of over 1000 essays that had been hand-annotated with discourse structure. In this paper, however, we explore unsupervised techniques that do not require a particularly large dataset.

To simplify the problem, we decided to limit ourselves to finding good topic sentences, since they are a common type of cohesive language in the style SAT essays are written in. Our key intuition was that topic sentences are also likely to be considered salient in the context of extractive summarization. Thus, we can use a simpler unsupervised summarization algorithm to find and evaluate topic sentences rather than using more sophisticated supervised discourse segmentation algorithms. The particular algorithm we chose was Lexrank [4], which is described below.

3.2.1 Lexrank

Lexrank is a completely unsupervised graph-based algorithm for extractive summarization. It creates a graph of sentences using a similarity metric such as cosine similarity of tf-idf vectors, and then applies the PageRank algorithm [6] to score the sentences. As with PageRank, the score can be interpreted in terms of random walks: the score is the probability that a sufficiently long random walk on the graph using edge weights as transition probabilities will end up on that sentence.

Since Lexrank scores represent probabilities, they must sum up to 1 over all the sentences in an essay. Thus, the Lexrank score of a sentence is really only meaningful in comparsion with the other sentences in the essay. So in order to assign a numerical score to an essay, we took the variance of the lexrank scores of the sentences in the essay, after scaling the scores so that their mean would be 1. This makes sense intuitively: in a good essay, strong topic sentences will stand out from other sentences (high variance), whereas in a bad essay, there are no clear topic sentences (low variance).

4 Data

We trained and tested on a set of 80 graded SAT essays. These were obtained from sample materials published by the College Board for student and teacher training, as well as from preparation websites for the SAT. It seems that there are no significant publically-available corpora of graded essays - most previous work seems to use private corpora such as data from past administrations of the SAT.

5 Learning Algorithms

5.1 Constructing a Multiclass SVM

Since SAT essays are evaluated on a discrete scale from 1 to 6, our choice of learning algorithm needed to be able to distinguish between six different classes of essays. We opted to take a simple one-versus-one approach, where a binary classifier was trained between each pair of classes, for a total of fifteen classifiers. In order to make a classification of a test example, each trained classifier "votes" for one of the classes it was trained on when presented the example. The class with the most votes is the predicted score of the essay. The classifier algorithm we used was an SVM with a linear kernel, as implemented in liblinear 1.7.

5.2 Evaluation

Since we had obtained a limited number of essays (80), we decided to use leave-one-out cross-validation in order to leave out a minimum amount of data. For each set of features, we computed the error using two different metrics:

1. The fraction of essays that had predicted scores that were more than one away from their actual scores. We chose this metric because on the actual SAT, graders are allowed to give an essay scores that differ by one point.

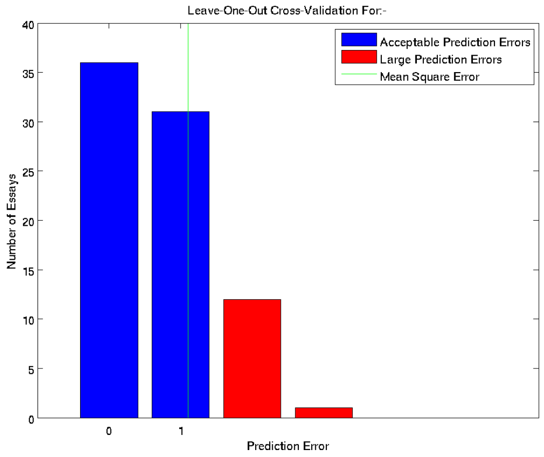
2. The mean square error. We chose this metric as a standard way to measure error; notice it appropiately punishes for predictions that are two or more points off their actual score, while punishing little for predictions that are within one of their actual score.

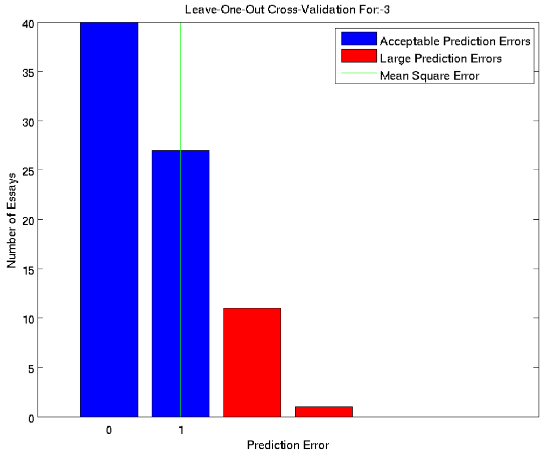
6 Results

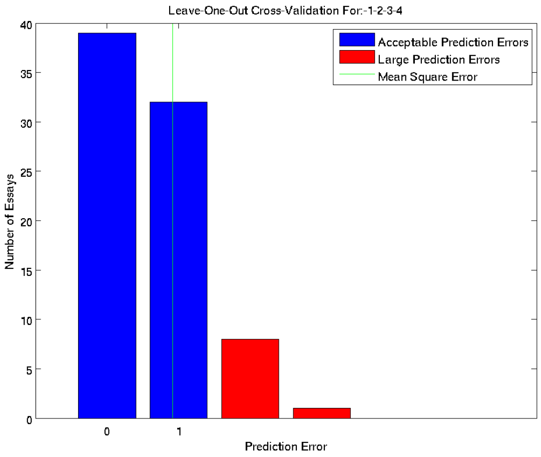
As described above, our algorithm had four additional features in addition to word counts: essay length (1), number of unique words in the essay (2), average word length (3), and lexrank score (4). Table 1 below contains the performance on both evaluation metrics for all 16 combinations of these features.

Features	Mean Squared Error	Num. Off by > 1 point
-	1.1	13
1	1.175	14
2	1.7625	15
12	1.1	13
3	1.0	12
13	1.025	12
23	1.0125	12
123	0.950	10
4	1.1125	14
14	1.2750	17
24	1.5250	15
124	1.0375	12
34	1.0250	13
134	1.0125	12
234	1.0750	14
1234	0.9125	9

We also plotted the number of essays misclassified by 0, 1, 2, 3, 4, or 5 points. Below are the plots for a few interesting configurations (in descending order: baseline, 3, and 1234).







7 Discussion & Conclusions

From table 1 above, we see that the best performance on both metrics was achieved with all four features. It is interesting to note that there appears to be a strong synergistic effect between total word count and number of unique words. Each feature on its own does relatively poorly, but when both used the performance is much better. This suggests that the relevant quantity may really be the ratio between them, which intuitively seems like it should measure how varied the vocabulary in an essay is. It is also interesting to note that average word length was the most helpful of all the single features. This suggests that sophistication of vocabulary is important in essay quality. Finally, the topic sentence feature seems to do better in combination with unique word counts. This could be because Lexrank calculates similarities by matching words, so essays with fewer distinct words also have less variance in lexrank scores. Adding the word count as a feature could help the SVM separate this from the actual topic sentence quality.

8 Acknowledgments

We would like to thank Prof. Ng and Prof. Jurafsky for their helpful advice.

References

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

[2] `http://professionals.collegeboard.com/testing/sat-reasoning/prep/essay-prompts`.

[3] `http://professionals.collegeboard.com/testing/sat-reasoning/scores/essay`.

[4] Güneş Erkan and Dragomir R. Radev. Lexrank: Graph-based centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22, 2004.

[5] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*, chapter 21. Pearson, 2009.

[6] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web, 1999.