Automated Essay Scoring Using Machine Learning

Shihui Song SHIHUI@STANFORD.EDU
Jason Zhao JLZHAO@STANFORD.EDU

Abstract

We built an automated essay scoring system to score approximately 13,000 essay from an online Machine Learning competition Kaggle.com. There are 8 different essay topics and as such, the essays were divided into 8 sets which differed significantly in their responses to the our features and evaluation. Our focus for this essay grading was the style of the essay, which is an extension on the studies conducted determining the quality of scientific articles by adding maturity to the feature set (Louis and Nenkova, 2013). An aspect of this project was to recognize the difference between the advanced nature of scientific articles to the coherency of middle to high school test essays. We evaluated Linear Regression, Regression Tree, Linear Discriminant Analysis, and Support Vector Machines on our features and discovered that Regression Trees achieved the best results with $\kappa = 0.52$.

NOTE: I (Shihui Song) had mentioned that I was working with Jason Zhao from Machine Learning, but we divided the project into two tracks: one for Machine Learning and one for NLP. The NLP track resulted in this project but the Machine Learning track was focusing on other Machine Learning techniques that we did not get the chance to test on this project. I worked on the NLP aspect and the Machine Learning part as well, but as he's not in the class, he was only consulting for the NLP project. This I would humbly request that you treat this as a single person project if possible.

1. Introduction

The automated essay scoring model is a topic of interest in both linguistics and Machine Learning. The model systematically classifies our varying degrees of speech and can be applied in both academia and large industrial organizations to improve operational efficiency.

1.1. Motivation

Each year, thousands of students take standardize tests with the same essay topics. Hand grading these essays is tedious and subjective. Instead, many organizations have already turned to automated essay grading to improve consistency and efficiency. Accurate models will not only reduce the amount of human error/variance in essay grading but could also save school boards and teachers many precious hours that could be used to improve the educational system.

2. Data Set

The training and test data were acquired from a past competition from Kaggle.com [1] sponsored by Hewitt-Packard. We had approximately 13,000 number of essays ranging from 150-550 words each provided for us. We split the essays into a 70-30 training and validation scheme, which results in a size of 9,100 essays for the training set and 3,900 essay for the test set. This divides further into around 1,200 training essays and 500 test sets per essay set.

3. Feature Generation

Python was utilized for the pre-processing of the data into matrices that were then fed to Matlab for supervised learning. The Natural Language Toolkit (NLTK) was the sole library used for the assignment and only for Parts of Speech Tagging.

As mentioned above, there were five categories of features that were considered and generated for this project.

3.1. Its visual nature

A more descriptive and visual description awes the reader, lingering with them. The source of imagery

[1] http://www.kaggle.com/c/asap-aes

and meaningfulness used for this dataset is derived from the British Natural Corpus (Kilgarriff, 1995) where each word has a imagery score between 0 and 999 where a higher number would be more visual. The features derived from this set include the proportion of words that are visual, proportion of unique visual words, average imagery scores for those words, the average imagery score for the essay, and all of the above for every third of the essay by dividing the essay into an introduction, a middle, and a conclusion.

3.2. Its use of people

Just as scientific articles which explicitly reference people would most likely be more respectable and concrete, as we had imagined allusions to people would also be in these essays. The Kaggle essays already contained name entity tags from the Name Entity Recognizer from the Stanford NLP group. Name entities for PERSON, ORGANIZATION, and LOCATION were categorized as *proper pronouns*, there were also counts of *personal pronouns* such as "us", "myself", etc, and the last pronoun count was for *relative pronouns*, which were noun phrases (tagged by the python NLTK tagger) followed by "who", "which", and "where".

We then divided all people pronouns into *animate* and organizations and location into *inanimate* under the assumption that the use of people would be more engaging than locations and organizations.

3.3. Its beautiful words

Beautiful word choices are thought to increase an essay's elegance, thereby its score. Only words above 5 characters in length are considered beautiful for this project. Two factors are considered for individual word beauty:

- **High perplexity-letter model** - how unlikely the combination of characters in the word is. For word, find the product of its character frequencies as according to the Cornell Math Cryptography [2]. The lower the product the more complex the word.

- **High perplexity-phoneme model** - we downloaded the corpus from the CMU pronunciation dictionary [3] to create a 4-gram frequency model for syllables. For every word in the essay, check to see if there exists a pronunciation for it, and if so, then also find the likeliness of its 4-gram combination.

- **Low word frequency** - this was included in the original paper by Louis and Nenkova, this was deemed too fundamental as that's often included for basic Machine Learning essay automations.

The ultimate features for this category were the average letter and phoneme frequencies per beautiful word and per essay, as well as the top 10, 20, and 30 average phoneme and letter frequencies where the top is the lowest frequency.

3.4. Its emotive effectiveness

A very dry and emotionless essay is not powerful. Twitter capitalizes on the burst of emotions with its short tweets, and millions of people follow one another because they are captivated by the emotions. The Subjectivity Lexicon from MPQA provides a list of words and their sentiments (positive, negative, neutral, or both) and the strength of those sentiments. The resulting features are proportions of sentiments and strength individually and combined for the entire essay or for given sentiments and strength. In addition, we also calculated the proportions of different emotions to one another.

In the end however, we realized that exhausting every combination of sentiment and strength proportional to another was actually noise that hurt the Kappa score. Therefore, we removed all proportions of sentiment and strength to other sentiments and strengths.

3.5. Its maturity

Our vocabulary expands as we grow older. Therefore, in a sense, as we mature, so does our vocabulary. This is particularly poignant for this set of essay, as they are written by students. Although this wasn't part of the Louis and Nenkova paper, we thought this would be a good addition to style. The Age of Acquisition is the average age when a person learns the certain word (Kuperman, Gonzalez, and Brysbaert, 2012).

4. Learning Algorithms

We mapped these features individually as categories against the various Machines Learning algorithms for individual Kappa scores which we then dissected and compiled an ultimate list with through the use of feature selection.

For the project, we evaluated several different classes of learning algorithms which will be described below.

[2] http://www.math.cornell.edu/mec/2003-2004/cryptography/subs/frequencies.html

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Most of the algorithms we evaluated are regression based where we treat the essay scores as a range of values and predict a floating point value within that range.

4.1. Generalized Linear Models

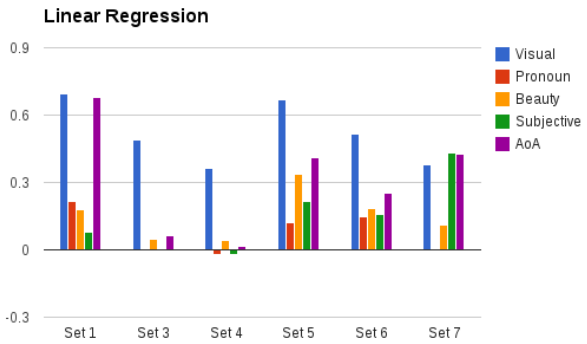

Figure 1. The Linear Regression Kappa score measurements for per essay set.

For the milestone, we used a simple linear regression model implemented by the statistics packet in Matlab(The MathWorks, 2013). The LinearModel class fits a linear function $h_\theta(x) = \theta^T x + c$ to a design matrix X in order to minimize the least square error as discussed in class. In the future, we are considering using Softmax Regression for multi-class labels.

4.2. SVM

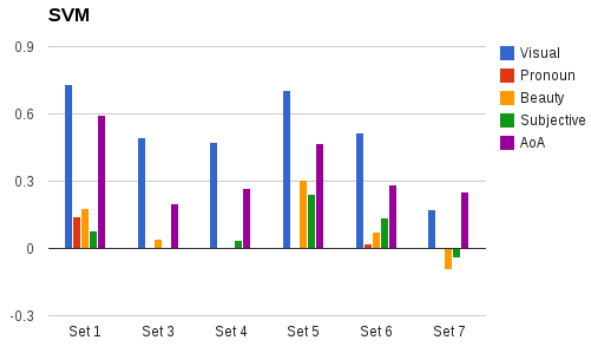

Figure 2. The SVM Kappa score measurements for per essay set.

We used the ν-SVM algorithm presented by (Schölkopf et al., 2000). It is a regression SVM algorithm based on the ϵ-SVM which introduces the ξ_i slack variables for capturing error (Vapnik, 1995). Specifically, the ν-SVM attempts to solve the following problem:

$$\min \tau(\mathbf{w}, \xi^{(*)}, \epsilon) = \frac{1}{2}||\mathbf{w}||^2 + C\left(\nu\epsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)\right)$$

$$s.t.\ ((\mathbf{w} \times x_i) + b) - y_i \leq \epsilon - \xi_i$$
$$y_i - ((\mathbf{w} \times x_i) + b) \leq \epsilon + \xi_i^*$$
$$\xi_i^{(*)} \geq 0, \epsilon \geq 0$$

We chose the ν-SVM because it reparameterizes the loss sensitivity term ϵ in the tradition C-SVM. This is desirable because ϵ is very hard to tune in practice whereas ν is simply an upper bound between the training error and the number of support vectors.

4.3. Multiclass Linear Discriminate Analysis

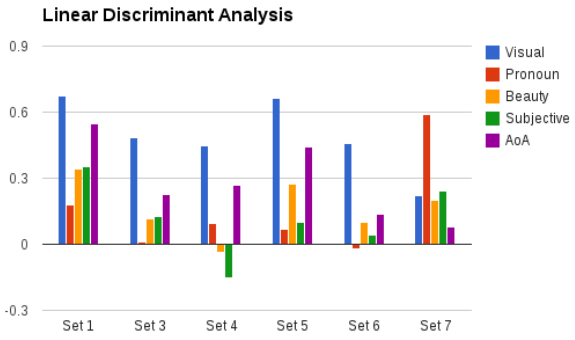

Figure 3. The Linear Discriminant Kappa score measurements for per essay set.

We used Matlab's ClassificationDiscriminant class to train a multi-class linear discriminant classifier(The MathWorks, 2013). The classifier predicts new examples using the following rule:

$$\hat{y} = arg \min_{y=1,\ldots K} \sum_{k=1}^{K} \hat{P}(k|x)C(y|k)$$

where K is the number of classes, $\hat{P}(k|x)$ is the posterior probability of k given x and $C(y|k)$ is the cost of classifying y when true class is k. We choose to use the default cost matrix for the milestone but may investigate other cost matrices for the final report. The posterior probability is estimated using a multivariate Gaussian distribution where the mean μ_k and covariance matrix Σ_k are approximated from the training

set.

$$P(x|k) = \frac{1}{2\pi|\Sigma_k|^{1/2}} exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right)$$

4.4. Regression Trees

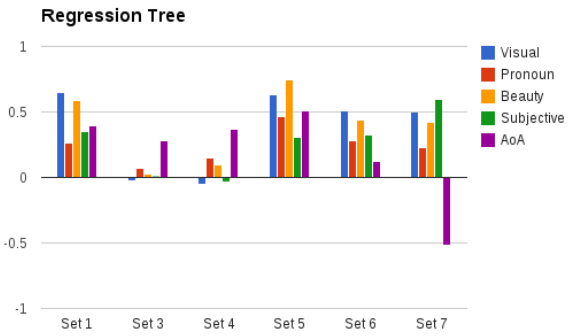

Figure 4. The Regression Tree Kappa score measurements for per essay set.

Matlab also supports a binary splitting decision tree that can fit to some response variable Y. This is an interesting algorithm because it uses a recursive partitioning model to divide the feature space into simple buckets. We used the mean squared error as the splitting criterion with a minimum leaf size of a single observation. The tree is post-pruned to generate the optimal sequence of subtrees. The resulting geometric interpretation is that the feature space is split into linear boxes (since this is a binary regression tree). Therefore the end result similar to an unsupervised clustering algorithm but the training phase is drastically different.

5. Experimental Results

We had used cross validation to ascertain which parameters would be the most useful. For example, we realized that too many proportional variables to emotions were actually bringing down the Kappa score, and few but important features were more efficient in rendering a better score.

5.1. Error Rates

The measure of error rate utilized is the quadratic weighted Kappa. The quadratic weighted Kappa measures the agreement between the automated essay grader and the human scores. Scores typically range between 0 (random agreement) and 1 (total agreement), although scores less than 0 can occur when there's less agreement than random.

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$$

The weight is the squared difference of the i and j scores over the square of the size of the set minus one. E is just a calculation based on the user's given vector based on score frequency. And $O_{i,j}$ is the number of occurrences when score i is assigned by grader one and score j is assigned by grader 2.(Kaggle.com, 2012)

6. Feature Selection and Evaluation

Above is the Kappa score comparison for the feature categories of style and their scores for each essay set. Please note that essay set 2 and 8 are excluded because essay set 2 asked for two separate ratings, which were furthermore dependent on specific trait rubrics. Essay set 8 was also not included due to the Matlab machine learning algorithms becoming rank deficient during the process and we felt that it was not a wholesome measurement.

6.1. Preliminary Evaluation

6.1.1. Individual features

Of the various individual feature sets in the graphs above, the imagery and age of acquisition features have the highest kappa score (even reaching 0.7 at times) on average while pronouns, emotions, and subjectivity all have pretty low scores (around 0.2 and below). This came as a surprise as we has believed that emotions in subjectivity would be the most relevant criteria for the reader to sympathize with.

The use of pronouns was also a disappointing use case with the only exception in pronouns for set 7; which was a narrative personal story, and references to other people were probably more crucial to the development of the essay.

We did not have too much hope for beauty, as the phonemes and character frequency collected by us were not thought to be as great statistics as those given by the other feature categories; however, they seem to be going just fine in set 1 of linear discriminant analysis and 7 of linear regression.

Many of the features are dependent on both the prompt and the age of the writers. This distinction is especially poignantly separated by the essay sets. Essays 3 through 6 are responses and evaluations of a prompt essay, thereby relating more to the context

of the given prompt and For instance, essay set 1 is a persuasive essay for grade 8 students on the effect of computers. This provides unrestricted access to anything that the child can come up with, instead of the context dependent essays from sets 3-6. The vocabulary shrinks in sets 3-6 due to the responsive nature of the essays since they are given an essay to respond to. Set 7 is once again narrative in nature, thereby the instead of subjective features are overall increased.

Interestingly, the lower the grade levels the better the Age of Acquisition scores are. Essay sets 1 and 5 are for grade 8 students while essay set 7 is for 7th grade children. This could be the product of children expanding their vocabulary at their ages, and the higher a vocabulary of a child would often result in a more varied and mature essay.

6.1.2. Algorithms' effects

The differences in the Machine Learning algorithms can come from the tokens in each essay. We were quite surprised at how well the Linear Regression algorithm did, considering that it was only fitting a plane for the dimensions. But since the features here are mostly proportional and not matrices or occurrence maps, then it is understandable that such data would be easy to fit. Regression Trees are difficult to consider, due to their binary nature. Because only a few important tokens can trigger a high regression tree score, the overall style is not necessarily reflection of the score. SVM was not expected to work in general for these concatenated and aggregated data sets as it was originally intended for predictions in high-dimensional data.

6.1.3. Combined features

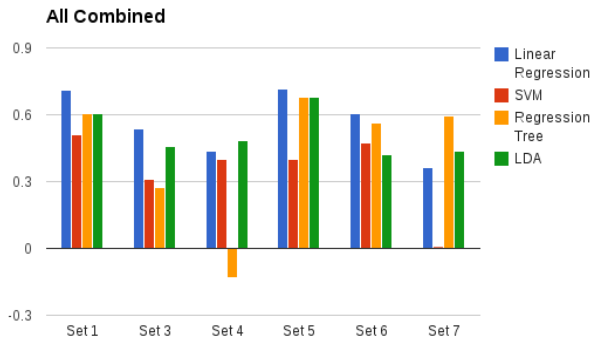

Figure 5. The Kappa set per Machine Learning algorithm for per essay set.

The overall first iteration Kappa score which includes all the categories using and comparing the various Machine Learning algorithms is included above. Linear Regression in on average the best resulting algorithm, which is the reason why we kept with it for our final evaluation.

6.2. Feature Selection

Judging by the data, we decided to keep only the Visual and Age of Acquisition categories with Linear Regression. Then for every feature in those categories, we added a feature one by two to see if the feature is helpful to increasing the Kappa score. If it doesn't help, then remove that feature (we deemed it helpful if it increased the essay set 1 score, since it was very difficult to determine the overall helpfulness given the different change in kappa scores and the number of essays in each set).

6.3. Final Product

This results in the following final prediction Kappa Scores for each problem set for Linear Regression.

Essay Set	Kappa Score
1	0.73
3	0.46
4	0.49
5	0.72
6	0.52
7	0.17
Overall	0.52

This is a pretty low disappointing score for the overall prediction Kappa scores, particularly essay set 7, which has a lower than expected Kappa Score since the original individual feature set scores were reasonable. This must be due to conflicting answers for the visual versus the age-of-acquisition which then intercepted both their scores incorrectly. Judging feature selection only by looking at essay set 1 also hurt the other essay sets; a feature applicable to one set does not necessarily help others.

7. Conclusion and Future Work

Essay grading is a hot topic that is known to be solvable by easy means such as word count, which according to a previous year's Machine Learning final project does better than the feature sets that we have used in this project (Mahana, Johns, and Apte, 2012).

The application of style to high school essay grading is not one of great success in this short venture into the territory. While the exploration of the age of acquisition as a reflection of the students' maturity has

possibilities for extensions, it can also be erroneous. This project has also followed closely on the features of the Louis and Nenkova paper, but perhaps instead of utilizing proportions of features, maybe just using those visual words, or other significant tokens as an version of visual count matrix could be a potential area of exploration. Furthermore, in the future, it would be worthwhile to explore not only the style by token content (as this project has mainly done so so far), but also on the structural and syntactical style. Perhaps, it is only with children that such mastery of style is not quite relevant, but perhaps judging Pulitzer prize winning essays would be more suited.

References

Cmu pronounciation dictionary, 1998. URL http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

English letter frequency, 2003. URL http://www.math.cornell.edu/~mec/2003-2004/cryptography/subs/frequencies.html.

Automated essay grading using machine learning. 2012. URL http://cs229.stanford.edu/proj2012/MahanaJohnsApte-AutomatedEssayGradingUsingMachineLearning.pdf.

What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association of Computational Linguistics*, 2013. URL paperhttp://www.transacl.org/wp-content/uploads/2013/07/paper341.pdf.

Kaggle.com. The hewlett foundation: Atumated essay scoring - evaluation, February 2012. URL http://www.kaggle.com/c/asap-aes/details/evaluation.

Kilgraff, Adam. Bnc database and word frequency lists, 1996. URL http://www.kilgarriff.co.uk/bnc-readme.html.

MPQA. Subjectivity lexicon. URL http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/.

Schölkopf, Bernhard, Smola, Alex J., Williamson, Robert C., and Bartlett, Peter L. New support vector algorithms. *Neural Comput.*, 12(5):1207–1245, May 2000. ISSN 0899-7667. doi: 10.1162/089976600300015565. URL http://dx.doi.org/10.1162/089976600300015565.

The MathWorks, Inc. Supervised learning, 2013. URL http://www.mathworks.com/help/stats/supervised-learning.html.

V, Kuperman. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 2012. URL http://link.springer.com/article/10.3758%2Fs13428-012-0210-4/fulltext.html.

Vapnik, Vladimir N. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. ISBN 0-387-94559-8.