Thank "Goodness"! A Way to Measure Style in Student Essays

Sandeep Mathias, Pushpak Bhattacharyya
Centre for Indian Language Technology
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
{sam,pb}@cse.iitb.ac.in

Abstract

Essays have two major components for scoring - content and style. In this paper, we describe a property of the essay, called **goodness**, and use it to predict the score given for the style of student essays. We compare our approach to solve this problem with baseline approaches, such as language modeling and also a state-of-the-art deep learning system, proposed by Taghipour and Ng (2016). We show that, despite being quite intuitive, our approach is very powerful in predicting the style of the essays.

1 Introduction

The first Automatic Essay Grading (AEG) system was Project Essay Grade developed by Ellis Page in 1966 (Page, 1966). Page (1966) believed that there are two major components to an essay, namely content (what the essay is about) and style (how well the essay is written). Style consists of two major parts, namely sentence fluency and word choice.

In 2012, a competition was organized by Kaggle. This competition, called the Automated Student Assessment Prize (ASAP), had multiple essays written by high school students of classes 7 to 10. The dateset for this competition has led to a large amount of research in AEG and automatic short-answer scoring in the last few years.

In this paper, we discuss one of the aspects of essay-writing, namely style, and how we can predict it automatically. In addition, we also look at two of the major components of style, namely word choice and sentence fluency. Style is necessary for providing a rich and diverse structure to the writing of the essay. Proficient and crisp vocabulary, as well as good sentence fluency is a mark of a writer being able to express his / her thoughts in the language of their writing.

Style is necessary for providing a rich and diverse structure to the writing of the essay. Proficient and crisp vocabulary, as well as good sentence fluency is a mark of a writer being able to articulate his / her thoughts well in the language of their writing.

There has been a fair bit of recent work in predicting other aspects of the essay, such as coherence (Somasundaran et al., 2014), organization (Persing et al., 2010), etc. However, not much work has been done for grading either style, sentence fluency, or word choice in student essays.

The central contribution of our paper is the definition of **goodness** and its use in predicting the style, word choice and sentence fluency scores of student essays. We define the **goodness** of a word (or phrase) as the weighted average of the count of the word (or phrase), weighted by score of the essay (either style, or word choice, or sentence fluency). In this way, words or phrases that occur more often in essays with a better score, get scored higher. Using this property, we show a significant improvement over our baseline measures, as well as a state-of-the-art deep learning system, developed by Taghipour and Ng (2016).

The rest of the paper is organized as follows. Section 2 defines the problem statement of our paper - in particular the terms style, word choice and sentence fluency. Section 3 describes our approach to predict the **goodness** scores of essays. Section 4 describes other features that we use, as well as a state-of-the-art system. Section 5 describes the dataset used. Section 6 describes the experiments that we performed. Section 7 gives our results and provides an analysis on the goodness of words and other features, and how they impact the sentence fluency score of essays. We also use ablation tests to find out which is the most

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 35–41
Melbourne, Australia, July 19, 2018.

important feature set. Section 8 describes related work in solving this problem. We conclude the paper in Section 9.

2 Problem Definition

We define style as the quality that measures how well the essay is written with respect to its language, vocabulary, sentences, etc. Hence, we say that style consists of 2 parts, namely word choice and sentence fluency.

Word choice is a quality in the essay where precise vocabulary is used. For example an essay using the word "express" ("Sally Yates expressed her concern about Michael Flynn's ties with Russia.") has a better word choice than if it were to use the word "say" ("Sally Yates said that she was concerned about Michael Flynn's ties with Russia.").

Sentence fluency is the quality of an essay that measures the writer's command of the language that they are writing in. A writer who is proficient in writing, will be able to form good quality phrases, construct sentences quite easily, and show a flow between the sentences that they write.

We model each of these as an ordinal classification problem, where each score point corresponds to a class.

3 Goodness

We hypothesize that essays with a better score in style, word choice or sentence fluency make use of words and phrases that have a higher goodness score. Goodness of a word (or phrase) W, is defined as the weighted average of W, weighted by the score of the essay. Hence, goodness is calculated using the formula:

$$Goodness(W) = \frac{\sum_i i * C_i(W)}{\sum_i C_i(W)},$$

where $Goodness(W)$ is the goodness of the word (or phrase) W, $C_i(W)$ is the count of word (or phrase), in essays scored with a score of i with respect to the relevant task (either style, word choice or sentence fluency).

For training, we run two passes over our dataset. In the first pass, we assign each word the same score of the essay (i.e. all words are assigned a score of i in essays with a score of i). Once this is done, we then construct the vocabulary in the second pass. In the second pass, we assign a score for each word in the vocabulary as the mean of the scores of the word throughout its occurrence in the training data. In this way, we learn the **goodness** scores of words and phrases.

For an unknown essay, we first score each word with the same score it has in the training data, it occurs in the training data set. Unknown words (or phrase) are scored as follows:

1. In case it is an **unknown word**, we find the most similar word to the unknown word using GloVe word vectors (Pennington et al., 2014) that is also present in the training data.

2. In case it is a **spelling mistake**. In case an unknown word does not exist in our set of word embeddings, we tag such a word as a spelling mistake, and assign a goodness score of 0.

3. In case it is an **unknown phrase**. In case there is a phrase that is not present in the training data, then it is marked as an unknown phrase. The score given to it is the mean score of its corresponding words.

We calculate the overall goodness score of the essay as the mean of the goodness scores of all the relevant words and phrases in the essay.

4 Additional Features

In addition to calculating the goodness, we also include the following add-on features to help improve our predictions of style, word choice and sentence fluency:

4.1 Essay statistics

These are length-based statistics about the essays, namely the number of words and sentences. We use these statistics because we observed that essays which were scored low (i.e. getting a 1) have a very low length, as compared to the average length of the essay. Similarly, essays that are scored high have a large number of words and sentences as well.

4.2 Punctuation features

In addition to the length-based features, we also count the number of commas, explanation points, question marks, and quotations. We believe that usage of these punctuation marks will help in detecting different kinds of sentences, like questions, exclamations, etc.

Prompt ID	Score Range	Essays	Average Length	Quantities Predicted
7	1-4	1569	250	Style
8	1-6	723	600	Word Choice & Sentence Fluency

Table 1: Properties of the data that we used.

4.3 Complexity features

Complexity measures, like the Flesch Reading Ease Score (FRES) are also used as features in our system. In addition to those, we also looked at parse tree features, like the average parse tree depth and the number of subordinate clauses (SBAR) in the text.

4.4 Language modeling features

These are language modeling features of the essay using the English Wikipedia from the Leipzig corpus (Goldhahn et al., 2012). These features are the output from the SRILM toolkit (Stolcke et al., 2002). We use the following features:

1. Number of sentences per essay.
2. Number of words per sentence.
3. Number of OOVs in the sentence.
4. Language model score.
5. Perplexity of the text.
6. Average perplexity per words of the text.

4.5 Coherence-based Features

We define sentence flow as the content word similarity between two adjacent sentences. For every pair of adjacent sentences, we find out $MaxSim$ and $MeanSim$, which are the maximum and mean similarity values between the content words of the 2 sentences (Pitler et al., 2010). We use the GloVe pre-trained word embeddings (Pennington et al., 2014) for the vectors of the content words.

In addition to the above, we also construct PoS-tag and lemma vectors of each of the sentences, and calculate the average similarity between adjacent sentences (Pitler et al., 2010).

We also look at entity grid features (Barzilay and Lapata, 2005). An entity grid is a 1-0 grid of sentences $\times$ entities. A cell ($E[i][j]$) in the grid is a 1 if the entity i is present in the sentence j, and 0 otherwise. We count the number of sequences of length between 2 to 4, that have at least one 1 and use them as features. A sequence of multiple 1s denote that an entity is referred to in a lot of consecutive sentences. On the other hand, sequences with a solitary 1 mean that the entity is mentioned just once, and never again in the adjacent sentences. The length of the sequence determines how many adjacent sentences we are considering at a time.

4.6 LSTMs - The State-of-the-Art

Deep learning networks, like LSTMs are quite good in predicting the score of the essays. We perform the experiments done by Taghipour and Ng (2016)[1]. We ran multiple configurations of their system. We used the default hyperparameters as described in Section 5.1 of Taghipour and Ng (2016). For pre-trained word embeddings, we ran experiments using

1. No pre-trained word embeddings
2. The same word embeddings that Taghipour and Ng (2016) used; and
3. GloVe word embeddings (Pennington et al., 2014)

The word-embeddings dimension for the look-up table layer was 50 for the first 2 experiments, and 300 for the experiment using GloVe.

5 Dataset

The complete ASAP training data set consists of nearly 13,000 essays, across 8 different essay prompts. The essays were written by students from classes 7 to 10. Things like dates, times, percentages, numbers, etc. were also anonymized.

Despite the fact that there are nearly 13,000 essays that have been graded in the data set, there are only two prompts (prompts #7 and #8) of 1569 and 723 essays, in which individual scores are given for each attribute or the essay. Since the scoring range is between 0 - 3 for prompt #7, we transform it to a range of 1 - 4, so that we can assign a **goodness** score of 0 to spelling errors, rather than to words belonging to the lowest-scoring essays.

[1]The system can be downloaded from https://github.com/nusnlp/nea

Experiment	Style	Word Choice	Sentence Fluency
Baseline Experiments			
Taghipour and Ng (2016)	0.4902	0.2511	0.3463
All features other than Goodness	0.5485	0.3433	0.3886
Goodness			
Goodness using only content words	0.2259	0.3323	0.3586
Goodness using all words	0.2821	0.3557	0.3984
Goodness using all words and content phrases	0.0792	0.1785	0.2241
ALL features	**0.5617**	**0.4233**	**0.4443**
Other human rater	0.5444	0.4816	0.5091

Table 2: Results of our experiments. These are the mean QWK scores. Numbers in **bold** denote the best system (excluding the human inter-rater agreement).

Table 1 describes the properties of the different different from which we score style, word choice and sentence fluency. Each of these scores were assigned by 2 annotators. For our experiments, we make use of Cohen's Kappa with Quadratic Weights - the Quadratic Weighted Kappa (QWK) (Cohen, 1968). The human inter-annotator agreement for style was 0.5444, word choice was 0.4816, and sentence fluency was 0.5091 between the human raters.

6 Experiment Setup

We model this problem as an ordinal classification problem where we consider each score to correspond to a class. We then classify the essay into the appropriate class that corresponds to its score.

This is not a run-of-the-mill classification problem as the values of the scores are ordered ($1 < 2 < 3 < ...$), and not independent. This is also not a regression problem, because the scores are discrete variables, and not continuous values. In regression, for instance, we could end up with scores higher than the maximum score possible. For instance, if the highest score was 4, if we are to use regression, we could end up scoring that essay 4.5!

We make use of the Ordinal Class Classifier (Frank and Hall, 2001) on Weka (Frank et al., 2016). The Ordinal Class Classifier is a meta-classifier that pre-processes the input data and transforms the input classes from ordinal to categorical classes before running the classification on an internal classifier. We ran our experiments using three classifiers, namely a Naïve Bayes Classifier (John and Langley, 1995), a Random Forest Classifier (Breiman, 2001), and a Multinomial Logistic Regression Classifier (le Cessie and van Houwelingen, 1992) as the internal classifier. The best classifiers were the Naïve Bayes Classifier for measuring style, and the Random Forest Classifier for measuring word choice and sentence fluency. We use *stratified* five-fold cross-validation. The results of our classification are given in Table 2.

7 Results and Analysis

The results of the 5-fold cross-validation of the training set are as shown in Table 2. The first block is the baseline experiments. The reported result for the neural network corresponds to the **best** neural network architecture - namely an LSTM with a CNN layer using GloVe pre-trained word embeddings due to space constraints. Block 2 features only goodness, and block 3 shows the results with all the features and compares it to the agreement with the other human rater.

In 2 out of the 3 tasks, using **goodness** without any additional features, we are able to outperform the baseline and Taghipour and Ng (2016)'s system. In the third task, while goodness is not able to outperform the baseline as well as the deep learning system, with the aid of language modeling, we are able to outperform the baseline when predicting style. This is because language modeling is able to reward / penalize style by itself.

7.1 Analysis of Goodness Scores

Table 3 gives examples of different words and their corresponding goodness scores for a single training fold for sentence fluency. Words with the lowest goodness scores tend to be spelling mistakes or out-of-context words. For instance, the word *computers* has the lowest goodness score of 1. This is because, in that fold it only occurs in a single training essay with word choice and sentence fluency scores of 1.

Range	Example Words	Example Phrases
1 - 2	ower, rumers, computers	sameting funing, adefokil stoeshi, feel happy we
2 - 3	tho, trash, reward	love laughter, a good thing, laugh that much
3 - 4	ok, fair, forever	make me happy, a joke, love to laugh
4 - 5	cherish, role, obvious	cherish forever, the center of attention
5 - 6	dire, aggressively, anguish	one of utter sarcasm, went on similarly, something ridiculous

Table 3: Example words with goodness scores for a single training fold in sentence fluency.

An interesting feature with respect to phrases is that the constituents of a phrase may have a lower score as compared to the overall goodness score of the phrase. For example, the words *cherish* and *forever* have mean goodness scores of 4.4 and 3.9 respectively, while the phrase *cherish forever* has a mean goodness score of 4.5.

7.2 Predictions Using Goodness Scores

If an essay contains a significant number of spelling errors (like *rumers*), or out-of-context words (like *computers*), the goodness score of the essay will be lowered and it will be predicted to have a lower style, word choice and sentence fluency score.

Unknown word handling allows us to handle spelling errors, as well as score words that are not present in the training data. For example *aggressively* has a mean goodness score of 5.5 across all training folds for both reviewers in the task of sentence fluency (out of 6). However, there may be a training fold in which it is not present. In one such fold, the synonym was *vigorously*, which also had a very high score of 4.5. In the absence of unknown word handling, we would skip it entirely.

When it came to using phrases, one of the challenges that we faced was data sparsity. For example, a phrase with a goodness score of 4.5, like *cherish forever* was ignored because the only essays that it occurred in were in the same fold. Hence, when any of those essays were encountered in testing, the phrases were tagged as an unknown phrase and skipped. Because of this, the results degraded when we used phrases.

To find out which of the feature sets worked best, we also ran ablation tests. We found out, that for style and word choice, goodness was the most important feature, and was the second-most important feature after the entity grid feature set, for sentence fluency.

Overall, we were able to consistently outperform the State-of-the-Art system, by using all our features in all three tasks.

7.3 Adversarial Essays

An adversarial essay is one where a human rater would rate it low but our system would be fooled into rating it high. A key question to ask here is: *Can a cunning student easily con the entire system into giving a good grade by submitting rubbish?* The answer is probably no. At least not easily. While it is possible for the writer to write an essay using *only* good words, this may not necessarily translate to a higher score than what he would have scored had he written the essay sincerely.

There are many ways to generate adversarial essays. Taghipour (2017) suggests using context-free grammars, and language modeling to create spurious essays, before trying to detect whether an input essay is spurious or not. Farag et al. (2018) construct adversarial essays by permuting the sentences of good scoring essays.

We created our own version of adversarial essays, by constructing essays that were long, but contained only "good" words (i.e. words with a high goodness score).

In order to see if such a thing would be possible we generated a set of 100 essays (50 from each prompt). These essays were generated from a vocabulary of *good* words, having above average length sentences and a reasonably large word count. We then graded these essays, using the original ASAP data for training. Table 4 shows ***how much*** is the average score, over the median score of the original essays.

Output	Goodness	Goodness++
Style	1.20	0.42
Sentence Fluency	1.96	1.22
Word Choice	2.05	1.36

Table 4: Adversarial Essays Average score increase using ONLY goodness scores (Goodness) and ALL features (Goodness++).

From this table we see that using all our features tends to make an average gain in score of about 1 point (out of 6 in sentence fluency and word choice) and 0.42 points (out of 4 in style) when we make use of all our features. In short, the easiest way for a cunning student to *beat* our system is for him / her to **write well**.

8 Related Work

As mentioned in Section 1, one of the major components of an essay is its style. While there has been work done in evaluating different subproblems with respect to style, there hasn't been too much work done with respect to evaluating style.

With respect to sentence fluency, Chae and Nenkova (2009) came up with a set of syntactic features to predict sentence fluency. They focused mainly on machine translation and articles written by people. However, the source of their articles was *published* articles from the Wall Street Journal (WSJ). WSJ articles are written by adults, proofread, and edited before publication. We focus on essays written by children studying in class 10 ***as is*** without any proof-reading or editing. Hence, they are expected to have a large number of errors, as compared to WSJ articles, which can serve as a discriminating factor between well and badly written essays.

In sentiment analysis, properties of adjectives have been used to predict the intensity of sentiment of a review as well (i.e. does the review *just* like the item or does he *really* like the item). Sharma et al. (2015) showed how intensity of adjectives could be a good predictor of deciding how positive or negative something is. Our approach - measuring the **goodness** of words / phrases to predict the style score of essays - is analogous to the Weighted Normal Polarized Intensity (WNPI) that they used.

In recent years, there has been a reasonable amount of research work done using deep learning to solve the problem of overall essay grading. However, not much has been done in the area of style, word choice or sentence fluency. Dong and Zhang (2016) describe a system for calculating the overall essay score using CNNs while Taghipour and Ng (2016) use LSTMs for predicting the overall score of essays.

9 Conclusions

We have defined a property of the essay called the **goodness** score, and use it as a way to score the style, word choice and sentence fluency of essays. We show that, by using goodness, we are able to predict the scores of the essays significantly better than the state-of-the-art system in essay grading, namely Taghipour and Ng (2016)'s essay grading system. Our system was able to achieve results that were close to human inter-rater agreement with respect to sentence fluency and word choice, and outperformed the human raters with respect to style.

References

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 141–148, Ann Arbor, Michigan. Association for Computational Linguistics.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

S. le Cessie and J.C. van Houwelingen. 1992. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201.

Jieun Chae and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: Case studies of machine translation and human-written text. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 139–147, Athens, Greece. Association for Computational Linguistics.

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

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. 2018. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *12th European Conference on Machine Learning*, pages 145–156. Springer.

Eibe Frank, Mark A. Hall, and Ian H. Witten. 2016. The weka workbench. online appendix for "data mining: Practical machine learning tools and techniques " morgan kaufmann, fourth edition, 2016.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, pages 759–765.

George H. John and Pat Langley. 1995. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo. Morgan Kaufmann.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

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, Cambridge, MA. Association for Computational Linguistics.

Emily Pitler, Annie Louis, and Ani Nenkova. 2010. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554, Uppsala, Sweden. Association for Computational Linguistics.

Raksha Sharma, Mohit Gupta, Astha Agarwal, and Pushpak Bhattacharyya. 2015. Adjective intensity and sentiment analysis. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2520–2526, Lisbon, Portugal. Association for Computational Linguistics.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 950–961, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Andreas Stolcke et al. 2002. Srilm-an extensible language modeling toolkit. In *Interspeech*, volume 2002, page 2002.

Kaveh Taghipour. 2017. *Robust Trait-Specific Essay Scoring Using Neural Networks and Density Estimators*. Ph.D. thesis.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.