ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores

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

Abstract

In this paper, we describe the creation of a resource - ASAP++ - which is basically annotations of the Automatic Student Assessment Prize's Automatic Essay Grading dataset. These annotations are scores for different attributes of the essays, such as content, word choice, organization, sentence fluency, *etc*. Each of these essays is scored by an annotator. We also report the results of each of the attributes using a Random Forest Classifier using a baseline set of task independent features as described by Zesch et al. (2015). We release and share this resource to facilitate further research into these attributes of essay grading.

Keywords: Automatic Essay Grading, Attribute-specific Essay Grading

1. Introduction

Automatic essay grading (AEG) is one of the most challenging activities in natural language processing (NLP). AEG makes use of many NLP and machine learning (ML) techniques in predicting the score of an essay - a piece of text that is written by a human on a given topic (called a prompt). It has been around since the 1960s, with the first AEG system - Project Essay Grade - proposed by Ellis Page (Page, 1966). Since then, there have been multiple systems that look at providing either a holistic score to the essay, or to score individual attributes of the essay. Examples of a few online systems include Grammarly[1] and Paper-Rater[2].

Essay grading systems rely on training and test data in order to grade essays. Most of the research today makes use of the Automated Student Assessment Prize's (ASAP) Automated Essay Grading (AEG) dataset to train and test systems[3]. The ASAP AEG dataset comprises of approximately 13,000 essays, written across 8 prompts. The essays were written by students of class 7 to 10. Each essay was evaluated by 2 evaluators. 6 out of the 8 prompts only have overall scores. Only 2 of them have scores for individual essay attributes, like content, organization, style, *etc*.

Our **contribution** is the scoring of individual attributes of the essays, like content, organization, style, etc. in the ASAP dataset for the remainder of the essays.

2. Motivation

A lot of the work in essay grading today makes use of the ASAP AEG dataset. However, most of the essays only have an overall score, not attribute-specific scores. This limitation limits the utility of this dataset for predicting the scores of particular attributes of essays.

Shermis and Burstein (2013) in chapter 19 (Contrasting State-of-the-Art Automatic Essay Grading Systems) of their book describe the ASAP dataset, as well as the results of current commercial AEG systems in scoring those essays. Since then, a large amount of work has been done using that dataset for evaluating the overall score of essays, from using machine learning techniques (Chen and He, 2013; Phandi et al., 2015) to deep learning systems (Dong et al., 2017; Dong and Zhang, 2016; Alikaniotis et al., 2016; Taghipour and Ng, 2016).

One common feature that all the above work has in common is the fact that the essay grading dataset that they used was the ASAP AEG dataset. However, most of them (in particular the deep learning systems) are constrained by the fact that there are very few prompts to handle scoring of individual attributes.

3. Related Work

While there has been a lot of work done in overall essay scoring, not much has been done with respect to scoring particular attributes of essays. Some of the attributes that have been scored include organization (Persing et al., 2010), prompt adherence (Persing and Ng, 2014), coherence (Somasundaran et al., 2014).

4. Dataset

The entire ASAP dataset has nearly 13,000 essays across 8 prompts. 6 of those 8 prompts, comprising nearly 10,400 essays, only have an overall score.

4.1. Essay Topics

The following is the list of topics of the 8 prompts in the dataset:

1. **Prompt 1** - The writers had to write a letter to their local newspaper in which they stated their opinion on the effects computers have on people.

2. **Prompt 2** - The writers had to write a persuasive essay reflecting their views on censorship in libraries.

3. **Prompt 3** - The writers had to read an extract from *Rough Road Ahead: Do Not Exceed Posted Speed Limit* by Joe Kurmaskie. They then had to explain how the features of the setting affected the cyclist.

[1] www.grammarly.com

[2] www.paperrater.com

[3] The dataset can be downloaded from here: https://www.kaggle.com/c/asap-aes/data.

Prompt ID	Essay Type	No. of Essays	Avg. Length	Score Range	Attribute Scores Contributed
Prompt 1	Argumentative	1785	350	1 - 6	**Yes**
Prompt 2	Argumentative	1800	350	1 - 6	**Yes**
Prompt 3	Source-Dependent	1726	150	0 - 3	**Yes**
Prompt 4	Source-Dependent	1772	150	0 - 3	**Yes**
Prompt 5	Source-Dependent	1805	150	0 - 4	**Yes**
Prompt 6	Source-Dependent	1800	150	0 - 4	**Yes**
Prompt 7	Narrative	1569	300	0 - 3	**No**
Prompt 8	Narrative	723	650	1 - 6	**No**

Table 1: Description of the ASAP AEG dataset. The Avg. Length column gives the average length of the essay, in terms of number of words. The score range column lists the scoring range of the various attributes that we score. We use the same score range as the overall score range of the essays. The last column tells us the prompts whose attribute scores we contribute. All the essays were written by native English speaking children from classes 7 to 10.

Feature Type	Feature List
Length	Word Count, Sentence Count, Sentence Length, Word Length
Punctuation	Counts of Commas, Quotations, Apostrophes, etc.
Syntax	Parse Tree Depth, Subordinate Clauses, etc.
Stylistic Features	Formality, Word Frequency, Type-Token Ratio
Cohesion Features	Discourse Connectives, Entity Grid, etc.
Coherence Features	Average Similarity between adjacent sentences of PoS tags, Lemmas, etc.
n-Gram Features	Word n-Grams and PoS n-Grams
Entity Grid Features	Bigrams, trigrams, and 4-grams of entity-sentence sequences

Table 2: Different features used in our experiment

4. **Prompt 4** - The writers had to read an extract from *Winter Hibiscus* by Minfong Ho. They then had to explain why the author concludes the story in the way that she did.

5. **Prompt 5** - The writers had to read an extract from *Narciso Rodriguez* by Narciso Rodriguez. They then had to describe the mood created by the author with supporting evidence from the extract.

6. **Prompt 6** - The writers had to read an extract from *The Mooring Mast* by Marcia Amidon Lusted. They then had to answer a question about the difficulties faced by the builders of the Empire State Building in allowing dirigibles to dock there.

7. **Prompt 7** - Write a story about a time when you, or someone you know, was patient.

8. **Prompt 8** - Write a story in which laughter plays a part.

Table 1 gives a description of the different essay prompts. Since scores are already present for prompts 7 & 8, we mainly provide scores for prompts 1 to 6.

4.2. Types of Essays

There are 3 types of essays in the dataset.

1. **Argumentative / Persuasive essays** - These are essays where the prompt is one in which the writer has to convince the reader about their stance for or against a topic (for example, free speech in public colleges).

2. **Source-dependent responses** - These essays are responses to a source text, where the writer responds to a question about the text (for instance, describing the writer's opinion about an incident that happened to him in the text).

3. **Narrative / Descriptive essays** - These are essays where the prompt requires us to describe / narrate a story.

Based on the type of the essay, we have a different set of attributes for evaluation. The ASAP dataset already contains attribute scores for the narrative essays, namely content, organization, word choice, sentence fluency, conventions, *etc*. Since we already have scores present for the narrative essays, we describe the scores for the other types of essays.

4.3. Attributes of Essays

Based on the types of essays, there are 2 sets of attributes[4].

4.3.1. Attributes of Argumentative / Persuasive Essays

There are 5 attributes for narrative essays, namely

1. **Content**: The quantity of relevant text present in the essay.

2. **Organization**: The way the essay is structured.

3. **Word Choice**: The choice and aptness of the vocabulary used in the essay.

[4]Details of the attributes and their scoring are also shared as a part of the resource.

Prompt ID	Content	Organization	Word Choice	Sentence Fluency	Conventions	Overall
Prompt 1	0.67	0.60	0.64	0.62	0.61	0.74
Prompt 2	0.61	0.58	0.60	0.59	0.62	0.62

Table 3: Results of the 5-fold cross-validation using the Random Forest classifier for argumentative / persuasive essays.

Prompt ID	Content	Prompt Adherence	Language	Narrativity	Overall
Prompt 3	0.59	0.59	0.57	0.63	0.54
Prompt 4	0.66	0.66	0.56	0.67	0.68
Prompt 5	0.67	0.64	0.60	0.63	0.76
Prompt 6	0.60	0.56	0.59	0.62	0.63

Table 4: Results of the 5-fold cross-validation results using the Random Forest classifier for source-dependent essays.

4. **Sentence Fluency**: The quality of the sentences in the essay.

5. **Conventions**: Overall writing conventions to be followed, like spelling, punctuations, etc.

4.3.2. Attributes of Source-dependent Responses

There are 4 attributes for source-dependent responses, namely

1. **Content**: The amount of relevant text present in the essay.

2. **Prompt Adherence**: A measure of how the writer sticks to the question asked in the prompt.

3. **Language**: The quality of the grammar and spelling in the response.

4. **Narrativity**: A measure of the coherence and cohesion of the response to the prompt.

We consider organization as an important attribute for the argumentative essays mainly because the average length of those essays is far more that of the source-dependent responses. The argumentative essays also have the scope for a wider vocabulary compared to the source dependent essays. Hence, we use word choice and organization as useful attributes for the argumentative / persuasive essays.

On the other hand, the source-dependent responses are constrained to respond to the source text. Hence, we have attributes like prompt adherence here, rather than word choice. The sentence fluency and conventions attributes are present in the language attribute of the source-dependent responses. The narrativity attribute attempts to ensure that the response is well-connected and makes sense. Hence, it is similar to organization, except that the organization attribute of the argumentative essays also requires that the essay have a good structure, like *introduction* $\rightarrow$ *body* $\rightarrow$ *conclusion*, while the source-dependent response would just be the *body*[5].

5. Creation of the Dataset

Each of the essays in a particular prompt were scored by an annotator. Each prompt was split into sets of 100 essays each, with the assumption that a set would correspond to a week's worth of time for the annotator. Thus, each prompt had a total of 18 sets[6].

Unlike the ASAP AEG dataset in which every essay was annotated by 2 annotators, we use only 1 annotator here for each essay. For the ground truth, we make use of the overall score of the essays given by the original annotators of the ASAP AEG dataset. In case the scoring of a particular attribute for a particular prompt differs from either of the original scorers by 2 or more points, the essay is then annotated by another annotator. The final score that is chosen is the one from the annotator that is closest to the overall scores. One of the reasons that we do this is because, in the 2 prompts that were rated by the original raters, there is a very high Pearson correlation (nearly 0.9) between the overall scores and the individual attribute scores.

5.1. Annotator Details

We made use of a total of three annotators to annotate the essays. Each of the annotators had competence in English, either by scoring quite high marks in their high school exams (over 90% in English), or scoring over 110 in ToEFL. Each of them also had some experience in evaluating texts, such as interning at *The Hindu*[7] (a top English newspaper in India), being the chief editor of the college magazine, *etc.* All the annotators have either studied or are studying English at a Master of Arts (MA) level.

6. Experiments

6.1. Features Used

After creating the resource, we ran experiments to get some baseline results. We used the task-independent feature set provided by Zesch et al. (2015). In addition to those features, we also made use of entity grid features described in Barzilay and Lapata (2005). Table 2 summarizes the list of features that we used in our experiments. All the features were extracted using Stanford Core NLP (Manning et al., 2014).

[5]Prompt #6, for instance, requires the writer to enumerate some of the difficulties faced by the builders of the Empire State Building in docking dirigibles

[6]Prompt 5, with a total of 1805 essays had 105 essays in its last set.

[7]http://www.thehindu.com/

Prompt ID	Content	Organization	Word Choice	Sentence Fluency	Conventions
Prompt 1	Coherence	Length	Coherence	Syntax	Coherence
Prompt 2	Coherence	Coherence	Coherence	Syntax	Coherence
Average	Coherence	Coherence	Coherence	Syntax	Coherence

Table 5: Results of the ablation tests using the Random Forest classifier for argumentative / persuasive essays to determine the most important feature set for each task in each prompt.

Prompt ID	Content	Prompt Adherence	Language	Narrativity
Prompt 3	Length	Coherence	Coherence	Style
Prompt 4	Punctuation	Language Model	Coherence	Complexity
Prompt 5	Length	Coherence	Punctuation	Language Model
Prompt 6	Coherence	Language Model	Coherence	Coherence
Average	Length	Coherence	Coherence	Coherence

Table 6: Results of the ablation tests using the Random Forest classifier for source-dependent responses to determine the most important feature set for each task in each prompt

6.2. Evaluation Metric

We evaluate each of the annotators using Cohen's Kappa, with quadratic weights - i.e. Quadratic Weighted Kappa (QWK) (Cohen, 1968).

We chose this as the evaluation metric (as compared to accuracy and weighted F-Score) because of the following reasons:

- Unlike accuracy and F-Score, Kappa takes into account random agreement. For example, a majority class classification will result in a Kappa of 0, while accuracy and F-Score will be the percentage of the majority class in the test set.
- Weighted Kappa, takes into account the distance between the actual score and the reported score. Quadratic weights reward matches and penalize mismatches more than linear weights.

Due to these reasons, this is one of the most used evaluation metrics to evaluate the performance of essay grading systems. To the best of our knowledge, all of the papers using the ASAP dataset make use of this as the evaluation metric.

6.3. Classifier Used

We made use of the Ordinal Class Classifier (Frank and Hall, 2001) in Weka (Frank et al., 2016). This classifier is a meta-classifier, that first converts ordinal data into categorical data, before running an internal classifier on the data. We used the Random Forest classifier (Breiman, 2001) as the internal classifier. We used 5-fold cross validation to get the results for each attribute for each prompt. We report the results in Table 3.

7. Results and Analysis

Tables 3 and 4 shows the Quadratic Weighted Kappa (QWK) with respect to predicting the score using the Random Forest Classifier. Since the feature set was designed specifically for the overall score of the essays, it is expected that the overall score usually has the best result (this is true with the exception of Prompt #3).

Most of the essays required only a single annotator. Only about a sixth of the essays required a second annotator.

One of the major problems that the annotators faced was the fact that all the essays were anonymized. Named entities, like *The New York Times* would be referred to as *@ORGANIZATION1*, *Donald Trump* would be referred to as *@PERSON1*, *etc*. The annotators were instructed not to penalize the essays because of the anonymizations, but were told to replace them with placeholders (like *@PERSON1* being replaced by either *Joe*, or *Jane*, *etc.* wherever applicable).

7.1. Ablation Tests

In order to see which feature sets are important for each of the attributes, we conducted ablation tests on each of the feature sets for each of the attributes. Tables 5 and 6 show which features are important for which attribute and which type of essay.

For source-dependent essays, we found out that the most important feature for content was length, while for argumentative / persuasive essays, it was coherence and cohesion features, followed by length. This is mainly because source-dependent essays are highly dependent on the source text, while argumentative / persuasive essays can utilize arguments from beyond the scope of any text, and so, those arguments have to be coherent and cohesive.

For source-dependent essays, the coherence and cohesion feature set is the most important feature set for each of the other 3 attributes. While narrativity is a measure of the coherence and cohesion of the text (and hence, it is self-evident that these features would be the most important for scoring this attribute), the language and prompt adherence scores also happen to be affected by this. This is mainly because source-dependent responses should adhere to the prompt that they are written as a response to.

For persuasive / argumentative essays, coherence and cohesion features are the most important features for 4 of the 5 attributes. This is mainly because coherence and cohesion are important for organization of the argument. The syntactic features though, were found to be the most important features for sentence fluency, since they measure how well

written the individual sentences of the essay are.
In fact, overall, the most important feature set is the coherence and cohesion features.

8. Conclusion

In this paper, we present a manually annotated dataset for automated essay grading. The annotation was done for different attributes of the essays. Most of the essays were annotated by a single annotator. However, about a sixth of them were annotated by a second annotator. These annotations can be used as a gold standard for future experiments in predicting different attribute scores.
The resource is available online at `https://cfilt.iitb.ac.in/~egdata/`. The resource is available for non-commercial research use under the Creative Commons Attribution-NonCommercial-ShareAlike License[8].

9. Acknowledgements

We thank the annotators of our task - Advaith Jayakumar, Janice Pereira and Elaine Mathias - for their help in creating this resource. We also thank members of CFILT at IIT Bombay for their valuable comments and suggestions. We also acknowledge the support of various funding agencies, like the Department of Information Technology (DIT), Ministry of Human Resource Development (MHRD), etc. of the Government of India.

10. Bibliographical References

Alikaniotis, D., Yannakoudakis, H., and Rei, M. (2016). Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany, August. Association for Computational Linguistics.

Barzilay, R. and Lapata, M. (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, June. Association for Computational Linguistics.

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

Chen, H. and He, B. (2013). Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752, Seattle, Washington, USA, October. Association for Computational Linguistics.

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

Dong, F. and Zhang, Y. (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, November. Association for Computational Linguistics.

Dong, F., Zhang, Y., and Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada, August. Association for Computational Linguistics.

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

Frank, E., Hall, M., and Witten, I. (2016). The weka workbench. *Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", 4th edition. Morgan Kaufman, Burlington*.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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

Persing, I. and Ng, V. (2014). Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543, Baltimore, Maryland, June. Association for Computational Linguistics.

Persing, I., Davis, A., and Ng, V. (2010). Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239, Cambridge, MA, October. Association for Computational Linguistics.

Phandi, P., Chai, K. M. A., and Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Lisbon, Portugal, September. Association for Computational Linguistics.

Shermis, M. D. and Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Somasundaran, S., Burstein, J., and Chodorow, M. (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, August. Dublin City University and Association for Computational Linguistics.

Taghipour, K. and Ng, H. T. (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, November. Association for Computational Linguistics.

Zesch, T., Wojatzki, M., and Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 224–232, Denver, Colorado, June. Association for Computational Linguistics.

[8] https://creativecommons.org/licenses/by-nc-sa/4.0/