Can Neural Networks Automatically Score Essay Traits?

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

Abstract

Essay traits are attributes of an essay that can help explain how well written (or badly written) the essay is. Examples of traits include Content, Organization, Language, Sentence Fluency, Word Choice, *etc*. A lot of research in the last decade has dealt with automatic holistic essay scoring - where a machine rates an essay and gives a score for the essay. However, writers need feedback, especially if they want to improve their writing - which is why trait-scoring is important. In this paper, we show how a deep-learning based system can outperform feature-based machine learning systems, as well as a string kernel system in scoring essay traits.

1 Introduction

An **essay** is a piece of text that is written in response to a topic, called a prompt. Writing a good essay is a very useful skill. However, evaluating the essay consumes a lot of time and resources. Hence, in 1966, Ellis Page proposed a method of evaluation of essays by computers (Page, 1966). The aim of automatic essay grading (AEG) is to have machines, rather than humans, score the text.

An AEG system is a software that takes an essay as input and returns a score as output. That score could either be an overall score for the essay, or a trait-specific score, based on essay traits like content, organization, style, etc. To the best of our knowledge, most of the systems today use feature engineering and ordinal classification / regression to score essay traits.

From the late 1990s / early 2000s onwards, there were many commercial systems that used automatic essay grading. Shermis and Burstein (2013) cover a number of systems that are used commercially, such as E-Rater (Attali and Burstein, 2006), Intelligent Essay Assessor (Landauer, 2003), Lightside (Mayfield and Rosé, 2013), etc.

In 2012, Kaggle conducted a competition called the Automatic Student Assessment Prize (ASAP), which had 2 parts - the first was essay scoring, and the second was short-answer scoring. The release of the ASAP AEG dataset[1] led to a large number of papers on automatic essay grading using a number of different techniques, from machine learning to deep learning. Section 3 lists the different work in automatic essay grading.

In addition to the Kaggle dataset, another dataset - the International Corpus of Learner's English (ICLE) - is also used in some trait-specific essay grading papers (Granger et al., 2009). Our work, though, makes use of only the ASAP dataset, and the trait specific scores provided by Mathias and Bhattacharyya (2018a) for that dataset.

The rest of the paper is organized as follows. In Section 2, we give the motivation for our work. In Section 3, we describe related work done for trait-specific automatic essay grading. In Section 4, we describe the Dataset. In Section 5, we describe the experiments, such as the baseline machine learning systems, the string kernel and super word embeddings, the Neural Network system, *etc*. We report the results and analyze them in Section 6, and conclude our paper and describe future work in Section 7.

2 Motivation

Most of the work dealing with automatic essay grading either deals with providing an overall score to the essay, but often doesn't provide any more feedback to the **essay's writer** (Carlile et al., 2018).

One way to resolve this is by using trait-specific scoring, where we either do feature engineering or construct a neural network, for individual traits.

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

Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pages 85–91
July 10, 2020.

Prompt ID	Trait Scores Range	Word Count	No. of Traits	No. of Essays	Essay Type
Prompt 1	1-6	350	5	1783	Argumentative / Persuasive
Prompt 2	1-6	350	5	1800	Argumentative / Persuasive
Prompt 3	0-3	100	4	1726	Source-Dependent Response
Prompt 4	0-3	100	4	1772	Source-Dependent Response
Prompt 5	0-4	125	4	1805	Source-Dependent Response
Prompt 6	0-4	150	4	1800	Source-Dependent Response
Prompt 7	0-6	300	4	1569	Narrative / Descriptive
Prompt 8	0-12	600	6	723	Narrative / Descriptive

Table 1: Properties of the dataset we used in our experiments.

However, coming up with different systems for measuring different traits is often going to be a challenge, especially if someone decides to come up with a new trait to score. Our work involves showing how we can take existing general-purpose systems, and use them to score traits in essays.

In our paper, we demonstrate that a neural network, built for scoring essays holistically, performs reasonably well for scoring essay traits too. We compare it with a task-independent machine learning system using task independent features (Mathias and Bhattacharyya, 2018a), as well as a state-of-the-art string kernel system (Cozma et al., 2018) and report statistically significant results when we use the attention based neural network (Dong et al., 2017).

3 Related Work

In this section, we describe related work in the area of automatic essay grading.

3.1 Holistic Essay Grading

Holistic essay grading is assigning an overall score for an essay. Ever since the release of Kaggle's Automatic Student Assessment Prize's (ASAP) Automatic Essay Grading (AEG) dataset in 2012, there has been a lot of work on holistic essay grading. Initial approaches, such as those of Phandi et al. (2015) and Zesch et al. (2015) made use of machine learning techniques in scoring the essays. A number of other works used various deep learning approaches, such as Long Short Term Memory (LSTM) Networks (Taghipour and Ng, 2016; Tay et al., 2018) and Convolutional Neural Networks (CNN) (Dong and Zhang, 2016; Dong et al., 2017). The current State-of-the-Art in holistic essay grading makes use of word embedding clusters, called super word embeddings, and string kernels (Cozma et al., 2018).

3.2 Trait-specific Essay Grading

Over the years, there has been a fair amount of work done in trait-specific essay grading, in essay traits such as organization (Persing et al., 2010; Taghipour, 2017), coherence (Somasundaran et al., 2014), thesis clarity (Persing and Ng, 2013; Ke et al., 2019), prompt adherence (Persing and Ng, 2014), argument strength (Persing and Ng, 2015; Taghipour, 2017; Carlile et al., 2018), stance (Persing and Ng, 2016), style (Mathias and Bhattacharyya, 2018b), and narrative quality (Somasundaran et al., 2018). Most of these works use feature engineering with classifiers to score the essay traits.

All the above mentioned works describe systems for scoring different traits ***individually***. In our paper, we compare three approaches to score essay traits, which are trait agnostic. The first uses a set of task-independent features as described by Zesch et al. (2015) and Mathias and Bhattacharyya (2018a). The second uses a string kernel-base approach as well as super word embeddings as described by Cozma et al. (2018). The third is a deep learning attention based neural network described by Dong et al. (2017). Our work is also, to the best of our knowledge, the first work that uses ***the same*** neural network architecture to automatically score essay traits.

4 Dataset

The dataset we use is the ASAP AEG dataset. The original ASAP AEG dataset only has trait scores for prompts 7 & 8. Mathias and Bhattacharyya (2018a) provide the trait scores for the remaining prompts [2]. Tables 1 and 2 describe the different essay sets and the traits for each essay set respectively.

[2]The dataset and scores can be downloaded from http://www.cfilt.iitb.ac.in/~egdata/.

Essay Set	List of Essay Traits
Prompt 1	Content, Organization, Word Choice, Sentence Fluency, & Conventions
Prompt 2	Content, Organization, Word Choice, Sentence Fluency, & Conventions
Prompt 3	Content, Prompt Adherence, Language, & Narrativity
Prompt 4	Content, Prompt Adherence, Language, & Narrativity
Prompt 5	Content, Prompt Adherence, Language, & Narrativity
Prompt 6	Content, Prompt Adherence, Language, & Narrativity
Prompt 7	Content, Organization, Style, & Conventions
Prompt 8	Content, Organization, Voice, Word Choice, Sentence Fluency & Conventions

Table 2: Traits that are present in each prompt in our dataset. The trait scores are taken from the original ASAP dataset, as well as from ASAP++ (Mathias and Bhattacharyya, 2018a).

5 Experiments

We use the following systems for our experiments:

1. **Feature-Engineering System**. This is a machine-learning system described by Mathias and Bhattacharyya (2018a).

2. **String Kernels and Superword Embeddings**. This is a state-of-the-art system on holistic essay grading developed by Cozma et al. (2018) using string kernels and superword embeddings.

3. **Attention-based Neural Network**. This is a system for holistic automatic essay grading described by Dong et al. (2017), that we adapt for trait-specific essay grading.

5.1 Baseline Feature-Engineering System

The baseline system we use is the one described by Mathias and Bhattacharyya (2018a). Their system used a Random Forest classifier to score the essay traits. The features that they used are length based features (word count, sentence count, sentence length, word length), punctuation features (counts of commas, apostrophes, quotations, etc.), syntax features (parse tree depth, number of clauses (denoted by $SBAR$ in the parse tree), *etc.*), sytlistic features (formality, type-token ratio, etc.), cohesion features (discourse connectives, entity grid (Barzilay and Lapata, 2008)), etc.

5.2 String Kernels and Superword Embeddings

Cozma et al. (2018) showed that using string kernels and a bag of super word embeddings drastically improved on the state-of-the-art for essay grading.

5.2.1 String Kernels

A string kernel is a similarity function that operates on a pair of strings a and b. The string kernel used is the histogram intersection string kernel ($HISK(a, b)$) that is given by the formula:

$$HISK(a, b) = \sum min(\#_x(a), \#_x(b)),$$

where $HISK(a, b)$ is the histogram intersection string kernel between two strings a and b, and $\#_x(a)$ and $\#_x(b)$ is the number of occurrences of the substring x in the strings a and b.

The string kernel is then normalized as follows:

$$\hat{k}(i, j) = \frac{k(i,j)}{\sqrt{k(i,i) \times k(j,j)}},$$

where $\hat{k}(i, j)$ is the normalized value of the string kernel $k(i, j)$ between the strings i and j.

5.2.2 Super Word Embeddings

A super word embedding is a word embedding created by making a cluster of word embeddings (Cozma et al., 2018). The clusters are created using the k means algorithm, with $k = 500$. For each essay, we use the count of words in each cluster as features.

5.3 Attention-based Neural Network

Figure 1 describes the architecture of Dong et al. (2017)'s neural network system. An essay is taken as input and the network outputs the grade for a particular trait. The essay is first split into sentences. For each sentence, we get the embeddings from the **word embedding** layer. The **4000 most frequent words** are used as the vocabulary, with all the other words mapped to a special unknown token.

This sequence of words is given as input to a 1-d CNN layer. The output from the CNN layer is pooled using an attention layer, which gets a word-level representation for every sentence in the essay. This is then sent through a sentence-level LSTM

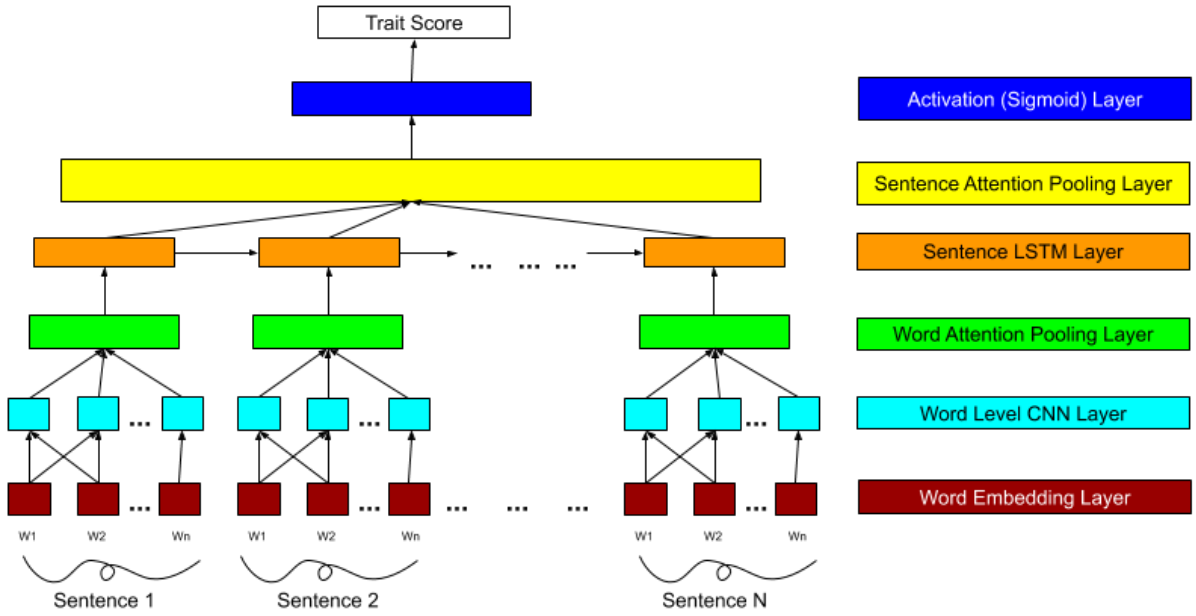

Figure 1: Architecture of Dong et al. (2017) neural network system

layer for getting a sentence-level representation of the essay.

We send the sentence-level representation of the essay through a sentence-level attention pooling layer, to get the representation for the essay. The essay representation is then sent through a Dense layer to score the essay trait. As the scores were converted to the range of $[0, 1]$, we use the **sigmoid activation function** in the activation layer, minimizing the **mean squared error loss** To evaluate the system, we convert the trait scores back to the original score range.

We use the **50 dimension** GloVe pre-trained word embeddings (Pennington et al., 2014). We run the experiments over a **batch size of 100**, for **100 epochs**, and set the learning rate as **0.001**, and a dropout rate of **0.5**. The Word-level CNN layer has a **kernel size of 5**, with **100 filters**. The Sentence-level LSTM layer and modeling layer both have **100 hidden units**. We use the **RMSProp Optimizer** (Dauphin et al., 2015) with a **momentum of 0.9**.

5.4 Experimental Setup

In this section, we describe different experiments.

5.4.1 Evaluation Metric

We choose to use Cohen's Kappa with quadratic weights (Cohen, 1968) - i.e. Quadratic Weighted Kappa (QWK) - as the evaluation metric. We use this as the evaluation metric because of the following reasons. Unlike accuracy and F-Score, Kappa takes into account if the classification happened by chance. Secondly, the accuracy and F-score metrics do not consider the fact that classes here are ordered. Thirdly, using weights allows Kappa to consider ordering among the classes. Lastly, by using quadratic weights, we reward matches and punish mismatches more than linear weights. Hence, we use QWK as the evaluation metric, rather than accuracy and F-score.

5.4.2 Evaluation Method

We evaluate the systems using **five-fold cross-validation**, with **60% training data**, **20% development data** and **20% testing data** for each fold. The folds that we use are the same as those used by Taghipour and Ng (2016).

6 Results and Analysis

Table 3 gives the results of the experiments using the different classification systems. In each cell, we compare the results of each of the 3 systems for a given trait and prompt. The bold value in each cell corresponds to the system giving the best value out of all the 3 systems. Traits which are not applicable to different prompts are marked with a "—".

We see that the attention-based neural network system is able to outperform both, the baseline system of Mathias and Bhattacharyya (2018a) and the histogram intersection string kernel system of

Prompt ID	System	Cont.	Org.	WC	SF	Conv.	PA	Lang.	Narr.	Style	Voice
Prompt 1	LREC 2018	0.628	0.606	0.618	0.594	0.588	—	—	—	—	—
	ACL 2018	0.686	0.637	0.659	0.639	0.620	—	—	—	—	—
	CoNLL 2017	**0.703**	**0.664**	**0.675**	**0.648**	**0.638**	—	—	—	—	—
Prompt 2	LREC 2018	0.563	0.551	0.531	0.495	0.486	—	—	—	—	—
	ACL 2018	0.600	0.570	0.583	0.544	0.530	—	—	—	—	—
	CoNLL 2017	**0.617**	**0.623**	**0.630**	**0.603**	**0.601**	—	—	—	—	—
Prompt 3	LREC 2018	0.586	—	—	—	—	0.575	0.534	0.594	—	—
	ACL 2018	0.659	—	—	—	—	0.658	0.590	0.645	—	—
	CoNLL 2017	**0.673**	—	—	—	—	**0.683**	**0.612**	**0.684**	—	—
Prompt 4	LREC 2018	0.646	—	—	—	—	0.636	0.577	0.641	—	—
	ACL 2018	0.702	—	—	—	—	0.702	0.571	0.687	—	—
	CoNLL 2017	**0.751**	—	—	—	—	**0.738**	**0.645**	**0.722**	—	—
Prompt 5	LREC 2018	0.667	—	—	—	—	0.639	0.618	0.647	—	—
	ACL 2018	0.713	—	—	—	—	0.700	0.620	0.635	—	—
	CoNLL 2017	**0.738**	—	—	—	—	**0.719**	**0.638**	**0.700**	—	—
Prompt 6	LREC 2018	0.579	—	—	—	—	0.581	0.555	0.592	—	—
	ACL 2018	0.759	—	—	—	—	0.711	0.624	0.635	—	—
	CoNLL 2017	**0.820**	—	—	—	—	**0.783**	**0.664**	**0.690**	—	—
Prompt 7	LREC 2018	0.495	0.528	—	—	0.533	—	—	—	0.577	—
	ACL 2018	0.737	0.659	—	—	0.504	—	—	—	0.609	—
	CoNLL 2017	**0.771**	**0.676**	—	—	**0.621**	—	—	—	**0.659**	—
Prompt 8	LREC 2018	0.510	0.571	0.518	0.507	0.431	—	—	—	—	0.507
	ACL 2018	0.573	0.572	0.494	0.477	0.455	—	—	—	—	0.489
	CoNLL 2017	**0.586**	**0.632**	**0.559**	**0.586**	**0.558**	—	—	—	—	**0.544**
Mean QWK	LREC 2018	0.584	0.564	0.556	0.532	0.510	0.608	0.571	0.619	0.577	0.507
	ACL 2018	0.679	0.610	0.579	0.553	0.527	0.693	0.601	0.651	0.609	0.489
	CoNLL 2017	**0.707**	**0.649**	**0.621**	**0.612**	**0.605**	**0.731**	**0.640**	**0.699**	**0.659**	**0.544**

Table 3: Results of each of the systems for scoring essay traits, namely Mathias and Bhattacharyya (2018a) (LREC 2018), Cozma et al. (2018) (ACL 2018) and Dong et al. (2017) (CoNLL 2017). — denote that the particular trait is not there for that particular prompt. The different traits are Content (Cont.), Organization (Org.), Word Choice (WC), Sentence Fluency (SF), Conventions (Conv.), Prompt Adherence (PA), Language (Lang.), Narrativity (Narr.), Style and Voice. **Mean QWK** is the mean QWK predicted for the trait across all essay sets.

Cozma et al. (2018) for all the traits, and across all 8 prompts. We also check if the improvements are statistically significant. We find that the improvements of the neural network system over the baseline system Mathias and Bhattacharyya (2018a) and histogram intersection string kernel system Cozma et al. (2018) to be statistically significant for $p < 0.05$ using the Paired T-Test.

Between the other 2 systems, the String Kernels system performed better than the baseline system in most of the cases. The only prompt in which it did not do so was in Prompt 8 - mainly because of the number of essays being very low and the size of the essay being very high compared to the other prompts.

Among the traits, the easiest to score are the traits of content and prompt adherence (where ever they are applicable) as they yielded the best agreement with the human raters. The hardest of the traits to score was Voice, which yielded the lowest QWK in the only prompt in which it was scored.

7 Conclusion and Future Work

In this paper, we describe a comparison between a feature-engineering system, a string kernel-based system, and an attention-based neural network to score different traits of an essay. We found that the neural network system provided the best results. To the best of our knowledge, this is the first work that describes how neural networks are used, in particular, to score essay traits.

As part of future work, we plan to investigate how to incorporate trait scoring as a means of helping to score essays holistically.

Acknowledgements

We would like to acknowledge the work done by the anonymous reviewers in evaluating our submission.

References

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

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more Feedback: Annotating Argument Persuasiveness and Related Attributes in Student Essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631. Association for Computational Linguistics.

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

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509, Melbourne, Australia. Association for Computational Linguistics.

Yann Dauphin, Harm De Vries, and Yoshua Bengio. 2015. Equilibrated adaptive learning rates for non-convex optimization. In *Advances in neural information processing systems*, pages 1504–1512.

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.

Fei Dong, Yue Zhang, and Jie Yang. 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. Association for Computational Linguistics.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. International Corpus of Learner English.

Zixuan Ke, Hrishikesh Inamdar, Hui Lin, and Vincent Ng. 2019. Give me more feedback II: Annotating thesis strength and related attributes in student essays. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3994–4004, Florence, Italy. Association for Computational Linguistics.

Thomas K Landauer. 2003. Automated scoring and annotation of essays with the intelligent essay assessor. *Automated Essay Scoring: A Cross-disciplinary Perspective*.

Sandeep Mathias and Pushpak Bhattacharyya. 2018a. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Sandeep Mathias and Pushpak Bhattacharyya. 2018b. Thank "Goodness"! A Way to Measure Style in Student Essays. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 35–41. Association for Computational Linguistics.

Elijah Mayfield and Carolyn Penstein Rosé. 2013. LightSIDE: Open source machine learning for text. In *Handbook of Automated Essay Evaluation*, pages 146–157. Routledge.

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. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2013. Modeling Thesis Clarity in Student Essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 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. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2015. Modeling Argument Strength in Student Essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2016. Modeling Stance in Student Essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2174–2184, Berlin, Germany. Association for Computational Linguistics.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 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. Association for Computational Linguistics.

Mark D Shermis and Jill Burstein. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge.

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 City University and Association for Computational Linguistics.

Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6:91–106.

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

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.

Yi Tay, Minh Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. SkipFlow: Incorporating Neural Coherence Features for End-to-End Automatic Text Scoring. In *Proceedings of the 32nd Annual AAAI Conference on Artificial Intelligence*.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 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. Association for Computational Linguistics.