Cognitively Aided Zero-Shot Automatic Essay Grading

Sandeep Mathias[1], **Rudra Murthy**[2], **Diptesh Kanojia**[3], and **Pushpak Bhattahcaryya**[1]
[1] Department of Computer Science & Engineering, IIT Bombay
[2] IBM Research India Limited
[3] IITB-Monash Research Academy
{sam,diptesh,pb}@cse.iitb.ac.in, rmurthyv@in.ibm.com

Abstract

Automatic essay grading (AEG) is a process in which machines assign a grade to an essay written in response to a topic, called the prompt. Zero-shot AEG is when we train a system to grade essays written to a new prompt which was not present in our training data. In this paper, we describe a solution to the problem of zero-shot automatic essay grading, using cognitive information, in the form of gaze behaviour. Our experiments show that using gaze behaviour helps in improving the performance of AEG systems, especially when we provide a new essay written in response to a new prompt for scoring, by an average of almost **5 percentage points** of QWK.

1 Introduction

One of the major challenges in machine learning is the requirement of a large amount of training data. AEG systems perform at their best when they are trained in a prompt-specific manner - i.e. the essays that they are tested on are written in response to the **same** prompt as the essays they are trained on (Zesch et al., 2015). These systems perform badly when they are tested against essays written in response to a different prompt.

Zero-shot AEG is when our AEG system is used to grade essays written in response to a completely different prompt. In order to solve this challenge of lack of training data, we use cognitive information learnt by gaze behaviour of readers to augment our training data and improve our model.

Automatic essay grading has been around for over half a century ever since Page (1966)'s work (Beigman Klebanov and Madnani, 2020). While there have been a number of commercial systems like E-Rater (Attali and Burstein, 2006) from the Educational Testing Service (ETS), most modern-day systems use deep learning and neural networks, like convolutional neural networks (Dong and Zhang, 2016), recurrent neural networks (Taghipour and Ng, 2016), or both (Dong and Zhang, 2016). However, all these systems rely on the fact that their training and testing data is from the same prompt.

Quite often, at run time, we may not have essays written in response to our target prompt (i.e. the prompt which our essay is written in response to). Because of the lack of training data, especially when training a model for essays written for a new prompt, many systems may fail at run time. To solve this problem, we propose a multi-task approach, similar to Mathias et al. (2020), where we learn a reader's gaze behaviour for helping our system grade new essays.

In this paper, we look at a similar approach proposed by Mathias et al. (2020) to grade essays using cognitive information, which is learnt as an auxiliary task in a multi-task learning approach. Multi-task learning is a machine-learning approach, where the model tries to solve one or more auxiliary tasks to solve a primary task (Caruana, 1998). Similar to Mathias et al. (2020), the scoring of the essay is the primary task, while learning the gaze behaviour is the auxiliary task.

Contribution. In this paper, we describe a relatively new problem - zero-shot automatic essay grading - and propose a solution for it using gaze behaviour data. We show a **5 percentage points** increase in performance when learning gaze behaviour, as opposed to without using it.

1.1 Gaze Behaviour Terminology

We use the following gaze behaviour terms as defined by Mathias et al. (2020). An ***Interest Area*** (IA) is a part of the screen that is of interest to us. These areas are where some text is displayed, and not the background on the left / right, as well as above / below the text. **Each word** is a separate

and unique IA. A ***Fixation*** is an event when the reader's eye fixates on a part of the screen. We are only concerned with fixations that occur inside interest areas. The fixations that occur in the background are ignored. ***Saccades*** are eye movements as the eye moves from one fixation point to the next. ***Regressions*** are a type of saccade where the reader moves from the current interest area to an *earlier* one.

1.2 Organization of the Paper

The rest of the paper is organized as follows. Section 2 describes the motivation for our work. Section 3 describes some of the related work in the area of automatic essay grading. Section 4 describes the essay dataset, as well as the gaze behaviour dataset. Section 5 describes our experiment setup. We report our results and analyze them in Section 6 and conclude our paper in Section 7.

2 Motivation

As stated earlier, in Section 1, one of the challenges for machine-learning systems is the requirement of training data. Quite often, we may not have training data for an essay, especially if the essay is written in response to a new prompt. Without any labeled data, in the form of scored essays, we cannot train a system properly to grade the essays.

Zero-shot automatic essay grading is a way in which we overcome this problem. In zero-shot automatic essay grading, we train our system on essays written to different prompts, and test it on essays written in response to the target prompt. One drawback of this approach is that it would not be able to use the properties of the target essay set in training the model. Therefore, as a way to alleviate this problem, we learn cognitive information, in the form of gaze behaviour, for the essays to help our automatic essay grading system grade the essays better.

3 Related Work

While there has been work done on developing systems for automatic essay grading, all of them describe systems which use some of the essays the system is tested on as part of the training data (as well as validation data, where applicable) (Chen and He, 2013; Phandi et al., 2015; Taghipour and Ng, 2016; Dong and Zhang, 2016; Dong et al., 2017; Zhang and Litman, 2018; Cozma et al., 2018; Tay et al., 2018; Mathias et al., 2020).

One of the solutions to solve the problem was using cross-domain AEG, where systems were trained using essays in a set of source prompt / prompts and tested on essays written in response to the target prompt. Some of the work done to study cross-domain AEG were Zesch et al. (2015) (who used task-independent features), Phandi et al. (2015) (who used domain adaptation), Dong and Zhang (2016) (who used a hierarchical CNN layers) and Cozma et al. (2018) (who used string kernels and super word embeddings). In all of their works, they defined a ***source prompt*** which is used for training and a ***target prompt*** which is used for validation and testing.

To the best of our knowledge, we are the first to explore the task of ***Zero-shot*** automatic essay grading, as a way to alleviate the challenge of a lack of graded essays (written in response to the target prompt) for an automatic essay grading system. In our approach, **we do not use the target prompt essays even for validation**, thereby making it truly zero-shot.

4 Datasets

In this section, we discuss our essay grading dataset and the gaze behaviour dataset which we used.

4.1 Essay Dataset Details

For our experiments, we use the Automatic Student Assessment Prize (ASAP)'s AEG dataset[1]. This dataset is one of the most widely-used essay grading datasets, consisting of 12,978 graded essays, written in response to 8 essay prompts. The prompts are either argumentative, narrative, and source dependent responses. Details of the dataset are summarized in Table 1.

4.2 Gaze Behaviour Dataset

For our experiments, we use the same essay grading dataset as Mathias et al. (2020). We use 5 attributes of gaze behaviour, namely dwell time (the total time that the eye has fixated on a word), first fixation duration (the duration of the first fixation of the reader on a particular word), IsRegression (whether or not there was a regression from a particular interest area or not), Run Count (the number of times an interest area was fixated on), and Skip (whether or not the interest area was skipped).

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

Prompt ID	Number of Essays	Score Range	Mean Word Count	Essay Type
Prompt 1	1783	2-12	350	Persuasive
Prompt 2	1800	1-6	350	Persuasive
Prompt 3	1726	0-3	150	Source-Dependent
Prompt 4	1770	0-3	150	Source-Dependent
Prompt 5	1805	0-4	150	Source-Dependent
Prompt 6	1800	0-4	150	Source-Dependent
Prompt 7	1569	0-30	250	Narrative
Prompt 8	723	0-60	650	Narrative
Total	12976	0-60	250	–

Table 1: Statistics of the 8 prompts from the ASAP AEG dataset.

Essay Set	0	1	2	3	4	Total
Prompt 3	2	4	5	1	N/A	12
Prompt 4	2	3	4	3	N/A	12
Prompt 5	2	1	3	5	1	12
Prompt 6	2	2	3	4	1	12
Total	8	10	15	13	2	48

Table 2: Number of essays for each essay set which we collected gaze behaviour, scored between 0 to 3 (or 4).

The gaze behaviour was collected from 8 different annotators, who read only 48 essays (out of the almost 13,000 essays in the ASAP AEG dataset) from the source dependent response essay sets. Table 2 summarizes the distribution of essays across the different essay sets that we collect gaze behaviour data for.

Table 3 gives the details of the different annotators used by Mathias et al. (2020). We evaluated the annotator's performance on 3 different metrics - QWK, Close and Correct. **QWK** is the Quadratic Weighted Kappa agreement (Cohen, 1968) between the score given by the annotator and the ground truth score from the dataset. **Correct** is the number of times (out of 48) that the annotator **exactly** agreed with the ground truth score, and **Close** is the number of times (out of 48) where the annotator disagreed with the ground truth score by **at most 1 score point**.

More details about the dataset and its creation are found in Mathias et al. (2020).

5 Experiment Setup

In this section, we describe our experiment setup, such as the evaluation metric, network architecture and hyperparameters, etc.

5.1 Evaluation Metric

For evaluating our system, we use Cohen's Kappa with Quadratic Weights, i.e. Quadratic Weighted Kappa (QWK) (Cohen, 1968). This evaluation metric is most frequently used for automatic essay grading experiments because it is sensitive to differences in scores, and takes into account chance agreements (Mathias et al., 2018).

5.2 Network Architecture

Figure 1 shows the architecture of our system. The essay is split into different sentences and each sentence is tokenized and given as input at the Embedding Layer. In this layer, for each token, we output the corresponding word embedding, which is given as input to the next layer - the Word-level CNN layer.

The Word-level CNN layer learns local representations of nearby words, as well as the gaze behaviour. The outputs of the word-level CNN layer are then pooled at the word-level pooling layer to get a sentence representation for each sentence.

Each sentence representation is then sent through an LSTM (Hochreiter and Schmidhuber, 1997) layer, whose output is pooled through a sentence-level attention layer, to get the essay representation.

The essay representation from the sentence-level attention layer is then sent through a Dense layer, from which we learn the essay scores. For both the tasks (learning gaze behaviour, as well as scoring the essay), we minimize the mean squared error loss.

ID	Sex	Age	Occupation	TA?	L1 Language	English Score	QWK	Correct	Close
Annotator 1	Male	23	Masters student	Yes	Hindi	94%	0.611	19	41
Annotator 2	Male	18	Undergraduate	Yes	Marathi	95%	0.587	24	41
Annotator 3	Male	31	Research scholar	Yes	Marathi	85%	0.659	21	43
Annotator 4	Male	28	Software engineer	Yes	English	96%	0.659	26	44
Annotator 5	Male	30	Research scholar	Yes	Gujarati	92%	0.600	19	42
Annotator 6	Female	22	Masters student	Yes	Marathi	95%	0.548	19	40
Annotator 7	Male	19	Undergraduate	Yes	Marathi	93%	0.732	21	46
Annotator 8	Male	28	Masters student	Yes	Gujarati	94%	0.768	29	45

Table 3: Profile of the annotators

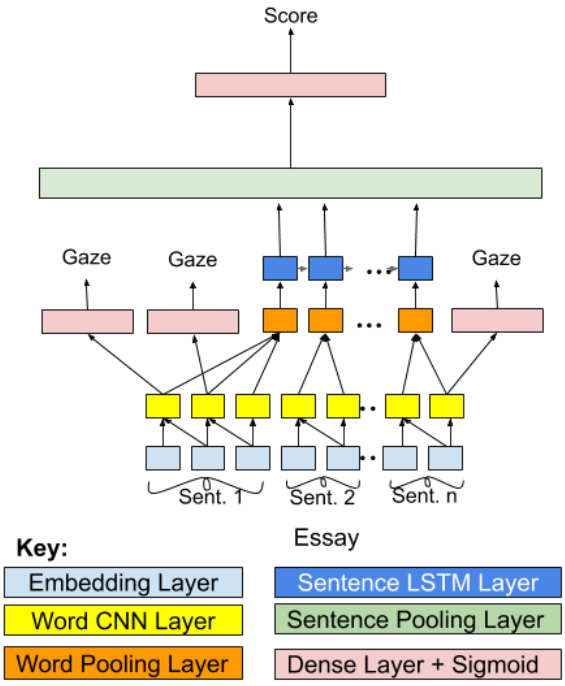

Figure 1: Architecture of our gaze behaviour system, showing an input essay of n sentences, with the outputs being the gaze behaviour (whenever applicable), and the overall essay score.

5.3 Network Hyperparameters

We use the **50 dimension** GloVe pre-trained word embeddings (Pennington et al., 2014). We run our experiments over a **batch size of 200**, for **50 epochs**. We set the **learning rate as 0.001**, and the **dropout rate as 0.5**. The word-level CNN layer has a **kernel size of 5**, with **100 filters**. The sentence-level LSTM layer has **100 hidden units**. We use the RMSProp Optimizer (Dauphin et al., 2015) with an **initial learning rate of 0.001** and **momentum of 0.9**. Along with the network hyperparameters, we also weigh the loss functions of the different gaze behaviour attributes differently, using the same weights as Mathias et al. (2020), namely **0.05 for DT and FFD**, **0.01 for IR and RC**, and **0.1 for Skip**.

5.4 Normalization and Binning

While training our model, we scale the essay scores for all the data (training, testing and validation) to a range of $[0, 1]$. For calculating the final scores, as well as the QWK, we rescale the predictions of the essay score back to the score range of the essays.

We also bin the gaze behaviour attributes as described in Mathias et al. (2020). Binning is done to take into account the idiosyncracies of the gaze behaviour of individual readers (i.e. some people may read faster, others slower, etc.). Whenever we use gaze behaviour, we scale the value of the gaze behaviour bins to the range of $[0, 1]$ as well.

5.5 Experiment Configurations

We run our experiments in the following configurations. **No Gaze** is a single-task learning experiment, where we only learn to score the essay. **Gaze** is the multi-task learning approach, where we learn gaze behaviour as an auxiliary task, and score the essay as the primary task.

5.6 Evaluation Method

We use **five-fold cross-validation** to evaluate our system. For each fold, the testing data consists of essays from the target prompt and the training data and validation data comprise of essays from the other 7 prompts.

6 Results and Analysis

Table 4 gives the results of our experiments. The results reported are on the target essay set for the mean of the 5 folds. For each fold, we record the performance of the model on the target essay set, corresponding to the epoch which had the best QWK for the development set. Table 4 reports the mean performance for all 5 folds.

From the table, we see that in most of the essay sets, we are able to see an improvement in perfor-

Target Essay Set	No Gaze	Gaze
Prompt 1	0.319	**0.423***
Prompt 2	0.391	**0.439***
Prompt 3	0.508	**0.545***
Prompt 4	0.548	**0.626***
Prompt 5	0.548	**0.628***
Prompt 6	0.599	**0.600**
Prompt 7	0.362	**0.420***
Prompt 8	**0.316**	0.286
Mean QWK	0.449	**0.498***

Table 4: Results of our experiments with and without using gaze behaviour. Improvements which are statistically significant (with $p < 0.05$), when gaze behaviour is used, are marked with a *

mance. In order to verify if the improvements were statistically significant, we use the 2-tailed Paired T-Test with a significance level of $p < 0.05$. Statistically significant improvements where we use gaze behaviour data are marked with a * next to the result.

Out of the 8 essay sets, the only essay set where the performance using gaze behaviour falls short compared to when we do not use gaze behaviour is in Prompt 8. One of the main reasons for this is that the essays in Prompt 8 are very long compared to the other essay sets. When they are absent from the training data, the system is unable to learn about the existence of long essays, which could also be the reason that those essays are scored badly.

7 Conclusion and Future Work

In this paper, we discussed an important problem for automatic essay grading, namely **zero-shot** automatic essay grading, where we have no labeled essays written in response to our target prompt, present at the time of training.

We showed that, by using gaze behaviour, we are able to learn cognitive information which can help improve our AEG system.

In the future, we plan to extend our work to other tasks, like grading of essay traits, using gaze behaviour.

References

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

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Rich Caruana. 1998. *Multitask Learning*, pages 95–133. Springer US, Boston, MA.

Hongbo Chen and Ben He. 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. 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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Sandeep Mathias, Diptesh Kanojia, Kevin Patel, Samarth Agrawal, Abhijit Mishra, and Pushpak Bhattacharyya. 2018. Eyes are the windows to the soul: Predicting the rating of text quality using gaze behaviour. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2352–2362, Melbourne, Australia. Association for Computational Linguistics.

Sandeep Mathias, Rudra Murthy, Diptesh Kanojia, Abhijit Mishra, and Pushpak Bhattacharyya. 2020. Happy are those who grade without seeing: A multi-task learning approach to grade essays using gaze behaviour. In *Proceedings of the 1st Conference of the*

Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 858–872, Suzhou, China. Association for Computational Linguistics.

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.

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.

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.

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.

Haoran Zhang and Diane Litman. 2018. Co-attention based neural network for source-dependent essay scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 399–409, New Orleans, Louisiana. Association for Computational Linguistics.