

Enhancing Cross-prompt Automated Essay Scoring by Selecting Training Data Based on Reinforcement Learning

Takumi Shibata, Masaki Uto

The University of Electro-Communications, 1-5-1 Chofugaoka, Chofu, Tokyo, Japan

Abstract

Automated essay scoring (AES) aims to automatically grade essays, thereby reducing the time and cost associated with manual scoring. The most common AES methods are classified under the prompt-specific approach, which involves developing a scoring model exclusively for a target prompt by using a dataset of scored essays corresponding to that prompt. Meanwhile, recent studies have emphasized the cross-prompt approach, which leverages scored essay data from other prompts, referred to as source prompts, to build an AES model for the target prompt. However, these cross-prompt methods have limitations in that they do not consider the presence of source prompt essays that can potentially have a negative impact on the construction of the AES model for the target prompt. To address this limitation, we propose a novel cross-prompt AES method that utilizes data valuation with reinforcement learning (DVRL). The proposed method enables the selective use of source prompt essays, which positively contributes to improving the scoring accuracy of the AES for the target prompt. Experiments on a benchmark dataset demonstrate that the proposed method enhances the performance of various AES models in cross-prompt scoring settings.

Keywords

Cross-prompt automated essay scoring, reinforcement learning, data valuation, transfer learning, educational measurement

1. Introduction


In recent years, dynamic changes in social structures have led to a growing emphasis on practical skills such as critical thinking and expressive abilities in educational settings. The essay exam has gained attention as a popular method for assessing these practical abilities [1, 2]. However, grading essays incurs substantial costs in terms of personnel, time, and money, and it is also challenging to ensure consistency and fairness in scoring [3]. To address these issues, automated essay scoring (AES) methods, which employ artificial intelligence technologies to automatically score essays, have been extensively explored in recent years (e.g., [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]).

AES methods can be broadly classified into two categories [22]: prompt-specific and cross-prompt methods. Prompt-specific AES methods construct a specialized scoring model for a single target prompt by using a training dataset consisting of scored essays corresponding to

EvalLAC'24: Workshop on Automatic Evaluation of Learning and Assessment Content, July 08, 2024, Recife, Brazil

✉ shibata@ai.lab.uec.ac.jp (T. Shibata); uto@ai.lab.uec.ac.jp (M. Uto)

ORCID 0000-0002-9330-5158 (M. Uto)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

that prompt¹. Traditional prompt-specific AES methods have relied on feature-based methods, which involve extracting specific features such as essay length and grammatical error rate from essays and training machine learning models using these features [4, 5]. However, these methods require substantial effort in feature engineering and their performance depends heavily on manually designed features. To address these limitations, deep learning-based approaches have gained popularity in recent years. These methods directly input the word sequences of essays into deep neural networks, eliminating the need for manual feature design [3, 14, 15]. In particular, pre-trained transformer–encoder-based models, such as those using BERT [24] or its variants, have been widely adopted over the past few years, and have demonstrated high performance [25]. Furthermore, recent research has begun to explore the potential of large language models (LLMs) for AES, investigating their enhanced knowledge retention and language-understanding capabilities [26, 27], although they are not necessarily superior to the AES models using BERT or its variants.

Although these prompt-specific AES models demonstrate high performance on the target prompt for which they were trained, there is no guarantee that directly applying the trained model to other prompts will yield high performance. To enhance the scoring performance for other prompts, it is generally necessary to collect an additional scored essay dataset tailored to each prompt and subsequently retrain the AES model using those data. To avoid such retraining processes, cross-prompt AES methods have recently been proposed [11, 17, 22, 23, 28, 29]. Cross-prompt AES methods build an AES model for a target prompt by leveraging scored essay data collected from other prompts, referred to as source prompts. The effective use of source prompt data can enhance the performance of an AES model for a target prompt, even when there are no or only a limited number of scored essays corresponding to that prompt.

Various cross-prompt AES methods have been explored recently. For example, Li et al. [23] proposed a feature-based AES model using prompt-independent features, constructed by domain adversarial neural networks (DANN) [30]. Furthermore, Ridley et al. [11] proposed a deep neural network model that integrates prompt-independent features and is designed to receive sequences of part-of-speech (POS) tags instead of word sequences as input in order to mitigate the influence of prompt-specific information. More recently, Chen et al. [22] introduced a technique that employs a contrastive learning approach to obtain more consistent prompt-independent features, thereby achieving the current state-of-the-art.

However, these existing cross-prompt AES methods are assumed to utilize all source prompt essays, ignoring the presence of essays that can potentially have a negative impact on the construction of the AES model for the target prompt [30, 31, 32]. Because some essays from source prompts that exhibit significantly different characteristics compared with the target prompt essays can act as noise, proper data selection to omit such essays is expected to improve scoring accuracy.

For this reason, we propose a cross-prompt AES method that follows the approach of data valuation by using reinforcement learning (DVRL) [32] to select source prompt essays that are valuable in constructing AES models for the target prompt. DVRL is a reinforcement learning framework that estimates the value of each data sample based on its contribution to

¹Note that the term *prompt* refers to the writing task or instructions given to a student, distinct from prompts used as inputs for large language models.

performance improvement in a specific target task. In our method, we adapt DVRL to construct a data value estimator, which assigns higher values to source prompt essays that positively contribute to AES performance on the target prompt and assigns lower values to those that might negatively impact the AES performance. The data selected using our DVRL framework can be used to construct any type of AES model, enhancing their AES performance on the target prompt compared with scenarios that use all source prompt data. In this study, we evaluate the effectiveness of our proposed method, using a benchmark dataset and several popular AES models, including BERT, Llama-2 [33], and the models proposed by Ridley et al. [11] and Chen et al. [22]. The experimental results show that the proposed method succeeded in improving performance across all AES models.

The remainder of this paper is structured as follows: Section 2 provides further details on conventional cross-prompt AES models. Section 3 explains the data valuation methods. Section 4 describes the proposed method, and Section 5 evaluates its effectiveness, using a benchmark dataset. Finally, Section 6 summarizes the study.

2. Conventional Cross-Prompt AES Methods

This section provides an overview of conventional cross-prompt AES methods and discusses the limitations and drawbacks of these approaches.

Jin et al. [17] proposed a cross-prompt AES method based on a two-stage approach. In the first stage, a RankSVM [34] is trained using essays from source prompts. This RankSVM is then used to generate prediction scores for essays of the target prompt, which serve as pseudo-scores for the next stage. In the second stage, a prompt-specific AES model is trained for the target prompt, using these pseudo-scores.

Li et al. [23] also proposed a two-stage AES method that utilizes DANN in the first stage. DANN is a deep learning approach that learns domain-independent features through an adversarial training process. This adversarial training uses two models: a main model that solves a target task and a domain classifier that identifies the domain each datum belongs to. These models are trained to maximize the performance of the main model while minimizing that of the domain classifier. The first stage of the method of Li et al. [23] uses the DANN to construct a feature extractor that produces prompt-independent features. Then, an AES model is constructed using source prompt data of essays that are vectorized by the feature extractor to generate pseudo-scores for the target prompt essays. The second stage trains a prompt-dependent AES model for the target prompt, using the target prompt essays with the pseudo-scores.

Meanwhile, Ridley et al. [11] introduced a model called *the prompt-agnostic essay scorer (PAES)*, which learns an AES model in an end-to-end fashion. PAES is a deep neural network model that integrates manually-designed prompt-independent features. This neural model is designed to receive sequences of POS tags instead of word sequences as input in order to mitigate the influence of prompt-specific information.

Chen et al. [22] proposed a model called *the prompt-mapping contrastive learning for cross-prompt automated essay scoring (PMAES)*, which uses contrastive learning to learn more consistent prompt-independent features. PMAES utilizes PAES as an encoder to generate feature vectors for essays. It then employs contrastive learning to bring the vectors from the essays

of source prompts closer to those from the target prompt. This process contributes to the construction of more consistent prompt-independent features, which are effective for cross-prompt scoring. PMAES has achieved state-of-the-art performance in cross-prompt AES methods.

As discussed above, conventional cross-prompt AES methods have focused primarily on learning prompt-independent features in order to extract transferable knowledge in essay scoring from source prompt data to target prompt data. However, these existing cross-prompt AES methods are assumed to utilize all source prompt essays, ignoring the presence of essays that can negatively impact the construction of the AES model for the target prompt [30, 31, 32]. Although these methods assume the source prompts to be a mixture of multiple prompts [11, 17, 22, 23, 28, 29], not all of the source prompts will necessarily share similar characteristics with the target prompt. Thus, the inclusion of source prompt essays that are greatly dissimilar to the target prompt essays can act as noise in the construction of an AES model for the target prompt. This issue becomes particularly relevant in conditions where there is a large variety of source prompts in terms of topics and writing styles. These insights suggest that a careful selection of source prompt essays would be effective for obtaining accurate cross-prompt AES models. The idea of our study is thus to apply data valuation methods to construct a selector of valuable source prompt essays.

3. Data Valuation Methods

Data valuation is a method for quantifying the importance of each sample in a dataset. Quantifying the value of data is regarded as an important task in various machine learning problems, including domain adaptation, discovering noisy samples, learning robust models, and improving the quality of datasets.

Representative data valuation methods include leave-one-out and data Shapley [35]. Leave-one-out is a method that estimates the importance of each sample by calculating the change in performance of a target task when removing each sample one by one. Data Shapley evaluates the value of data, using the Shapley value from cooperative game theory. Specifically, data Shapley calculates the marginal contribution of each sample by evaluating the prediction performance of a target task when using each possible combination of samples. Moreover, another method using the Banzhaf value, which originates from cooperative game theory as well, has also been proposed [36].

Several data valuation methods based on meta-learning have also been proposed. One example is ChoiceNet [37], a valuation method that identifies noisy data within training datasets by separately estimating the distributions of meaningful data and noise data. Learning to reweight [38] is another method that calculates the weights of each sample in the source dataset based on the performance of a target task on a validation dataset. Furthermore, as a recent meta-learning-based data valuation method, Yoon et al. [32] proposed a method called *data valuation using reinforcement learning (DVRL)*. DVRL employs a reinforcement learning strategy that simultaneously optimizes a data value estimator and a predictor model for a target task. In this study, we apply the framework of DVRL to cross-prompt AES.

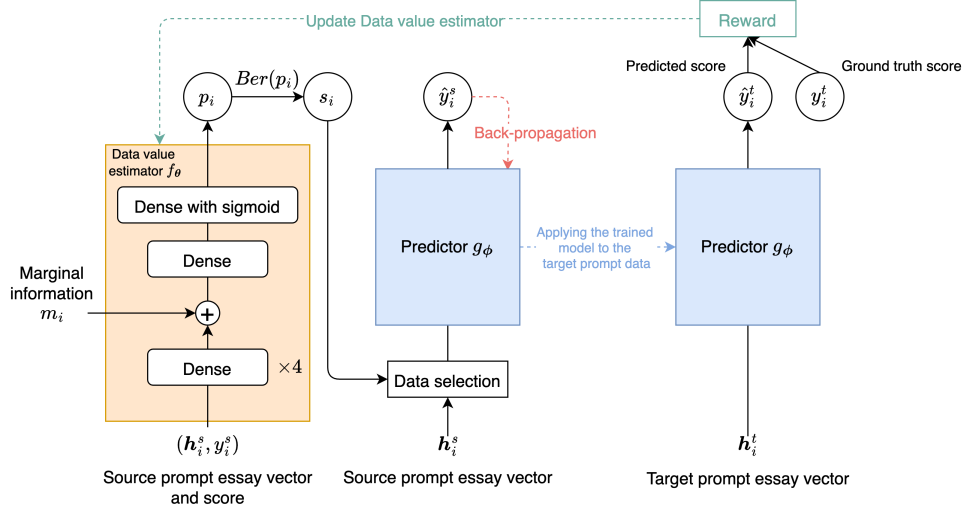


Figure 1: Model architecture of DVRL

4. Proposed Method

4.1. Task Definition

This study assumes that a large number of scored essays from a mixture of source prompts $\mathcal{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and a small number of scored essays for the target prompt $\mathcal{D}^t = \{(x_i^t, y_i^t)\}_{i=1}^{N_t}$ are given. Here, x_i^s and x_i^t represent the i -th essay in the source and target prompt essays, respectively, while y_i^s and y_i^t denote their corresponding scores. N_s and N_t represent the total numbers of essays for the source prompts and target prompt, respectively.

Our study aims to develop an AES model that can accurately predict scores for unscored essays corresponding to the target prompt by executing the following two steps.

1. Construct a data value estimator, using DVRL to assign value scores to each essay in the source prompt essays.
2. Train an AES model for the target prompt, using a subset of source prompt essays assigned high-value scores by the data value estimator.

Note that this study exclusively uses \mathcal{D}^s in the AES training process, while both \mathcal{D}^s and \mathcal{D}^t are used in the DVRL process². The following sections describe the details of each step.

4.2. Data Valuation Using DVRL

Figure 1 illustrates the outline of our DVRL framework. It consists of two models: a *data value estimator* f_θ that estimates the value of each scored essay data, and a *predictor* g_ϕ that outputs

²It should be noted that \mathcal{D}^t is also available to train the AES model constructed in step 2. However, we do not use \mathcal{D}^t because this study focuses on how data selection by the proposed method affects AES performance compared with scenarios in which all source prompt data are used. A detailed evaluation of the effect of integrating \mathcal{D}^t as AES training data remains a subject for future research.

the predicted score of the essay. Here, θ and ϕ are the model parameters of the data value estimator and predictor, respectively.

In the figure, \mathbf{h}_i^s and \mathbf{h}_i^t represent feature vectors corresponding to x_i^s and x_i^t , respectively. The method for creating these feature vectors depends on the type of AES model that will ultimately be constructed. Specifically, when we intend to use AES models that accept word sequences as input, we use distributed essay representation vectors obtained from DeBERTa-v3-large [39, 40] as the feature vectors. Meanwhile, when we intend to use cross-prompt AES models such as PAES and PMAES, we utilize manually designed prompt-independent features.

The learning process of DVRL is formulated as the following optimization problem:

$$\max_{f_\theta} \mathbb{E}_{(\mathbf{h}^t, y^t) \sim \mathcal{P}^t} [R(\phi)] \text{ s.t. } g_\phi^* = \arg \min_{g_\phi} \mathbb{E}_{(\mathbf{h}^s, y^s) \sim \mathcal{P}^s} [f_\theta(\mathbf{h}^s, y^s) \mathcal{L}(g_\phi(\mathbf{h}^s), y^s)]. \quad (1)$$

Here, $R(\phi)$ represents the reward, which is the performance of the predictor g_ϕ trained using the source prompt data \mathcal{D}^s and evaluated using \mathcal{D}^t as test data. The reward is measured using the quadratic weighted kappa (QWK) metric, which assesses the agreement between the predicted scores and the ground truth scores and is widely used in AES studies [3, 10]. \mathcal{L} denotes the mean squared error (MSE) loss function used to train the predictor, as explained in Section 4.2.2. \mathcal{P}^s and \mathcal{P}^t represent the distributions of the source prompt data and the target prompt data, respectively. Solving this formulation offers a data value estimator that estimates the value score of each essay. The following subsections explain the specific calculation procedures.

4.2.1. Data Value Estimator

For each essay vector \mathbf{h}_i^s and its score y_i^s for the source prompt essays in \mathcal{D}^s , the data value estimator f_θ outputs its data value $p_i \in [0, 1]$ as $p_i = f_\theta(\mathbf{h}_i^s, y_i^s)$. The data value estimator f_θ is implemented using a deep neural network with six stacked dense layers, where the output layer is designed as a linear layer with sigmoid activation; it also incorporates marginal information m_i into its intermediate layer. The marginal information m_i is a quantity expected to correlate with the data value of each essay i and can be written as $m_i = |y_i^s - \hat{g}_\phi(\mathbf{h}_i^s)|$, where \hat{g}_ϕ is a predictor trained on \mathcal{D}^t .

Using the calculated data value p_i , the selection indicator $s_i \in \{0, 1\}$ for each essay is determined by sampling from a Bernoulli distribution with probability p_i ; that is, $s_i \sim \text{Ber}(p_i)$, where $s_i = 1$ means that the i -th data is selected, and $s_i = 0$ means that it is not selected.

4.2.2. Predictor

The source prompt data selected through the above procedure are used to train the predictor g_ϕ . The predictor is designed as a multi-layer perceptron with a linear output layer with sigmoid activation³. The weighted loss function \mathcal{L}_{pred} used for learning is calculated as follows:

$$\mathcal{L}_{pred}(\phi) = \frac{1}{N_s} \sum_{(x_i^s, y_i^s) \in \mathcal{D}^s} s_i \cdot \mathcal{L}(\hat{y}_i^s, y_i^s), \quad (2)$$

³In our study, we used different multi-layer perceptrons depending on the input data types. Specifically, a two-layer perceptron is used for cases inputting distributed essay representation vectors obtained from DeBERTa-v3-large, while a single-layer perceptron is used for cases inputting manually designed prompt-independent features.

where \hat{y}_i^s is the predicted score of the predictor g_ϕ for the i -th essay of the source prompt data. As the loss function \mathcal{L} , we use the MSE between the predicted score \hat{y}_i^s and the ground truth score y_i^s . Note that the ground truth scores y_i^s are assumed to be normalized to the range $[0, 1]$ because the predicted scores are within this range too, as a result of the sigmoid activation in the output layer.

4.2.3. Reinforcement Learning

Using the trained predictor, our method computes the reward $R(\phi)$ for reinforcement learning as the QWK between the predicted scores and the ground truth scores evaluated using the dataset \mathcal{D}^t . The reward $R(\phi)$ is used to update the parameters θ of the data value estimator f_θ . Specifically, the parameters θ are updated using the REINFORCE algorithm [41], a reinforcement learning algorithms, with the following loss function [32]:

$$\mathcal{L}_{RL}(\theta) = R(\phi) * \log P((s_1, s_2, \dots, s_{N_s}) | \theta), \quad (3)$$

where $P((s_1, s_2, \dots, s_{N_s}) | \theta)$ represents the joint probability of the selection indicators given the parameters θ . Note that each essay is selected independently, meaning that the joint probability can be written as $\prod_{i=1}^{N_s} p_i^{s_i} (1 - p_i)^{1-s_i}$. Using this loss function, the parameters θ are updated by gradient ascent as follows:

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} \mathcal{L}_{RL}(\theta), \quad (4)$$

where α represents the learning rate, which is set to 0.001 in this study. Adam [42] is used as the optimization method for parameter updates.

Finally, by repeating the above steps until the model converges, the data value estimator f_θ is trained.

4.3. Train an Arbitrary AES Model Based on Estimated Data Values

Through the above process, we can obtain the data value estimator f_θ and the resulting data value scores for essays in the source prompt data \mathcal{D}^s . Thus, our last step is to construct an AES model for the target prompt, using source prompt essays with high-value scores. However, it is not clear how much data should be selected based on their value scores. Thus, we employ the following approach, which is inspired by that described in [32], to select essays based on their value scores.

1. Sort the source prompt essays in descending order based on their estimated value scores.
2. Train an AES model using essays with top 10% value scores and repeat this process with different data usage percentages, ranging from 10% to 100%, in increments of 10%.
3. For the ten constructed models, evaluate their MSE loss, using \mathcal{D}^t as test data. The model with the lowest MSE loss is selected as the optimal one and is used for scoring the unscored target prompt essays.

Table 1
Details of the ASAP.

Prompt	No. of essays	Avg. len.	Genre	Score range
1	1783	350	Argumentative	2–12
2	1800	350	Argumentative	1–6
3	1726	150	Source-dependent	0–3
4	1772	150	Source-dependent	0–3
5	1805	150	Source-dependent	0–4
6	1800	150	Source-dependent	0–4
7	1569	250	Narrative	0–30
8	723	650	Narrative	0–60

5. Experiment

We conducted an evaluation experiment using real-world data to demonstrate the score prediction performance of the proposed method compared with the conventional method, which uses all source data.

5.1. Dataset

In this experiment, we used the ASAP (Automated Student Assessment Prize)⁴ dataset as real-world data. The ASAP dataset is used in Kaggle’s automated essay-scoring competition and is widely used as a benchmark dataset in many AES studies. The ASAP contains a total of 8 essay prompts for 3 genres: argumentative, source-dependent responses, and narrative. Each prompt also includes student’s essays and their scores. The details of the dataset characteristics are shown in Table 1.

5.2. Performance Evaluation of our Proposed Method

In line with previous cross-prompt AES studies, the present experiment was conducted using prompt-wise cross-validation [11, 17, 22]. In prompt-wise cross-validation, one prompt is used as the target prompt, while all remaining prompts are used as source prompts for training. This operation is performed sequentially for all prompts, and the average is calculated to evaluate performance.

Our proposed method needs \mathcal{D}^t , a small number of scored essay data sampled from the target prompt. In this experiment, the size of \mathcal{D}^t was set to 30, and the set of samples was selected so that the sum of the Euclidean distances between each distributed essay representation vector obtained from DeBERTa-v3-large was maximized.

Our proposed method can be used for any AES model. The present experiment used four representative AES models: BERT, Llama-2-7B [33], PAES, and PMAES. Note that the PMAES with the same hyper-parameters as in [22] could not be implemented using our GPU (RTX4090). Thus, we changed some hyper-parameters. Specifically, the number of mini-batches was changed from 2 to 20.

⁴<https://www.kaggle.com/c/asap-aes>

Table 2
Experimental results.

Model	Setting	Prompts								Avg.
		1	2	3	4	5	6	7	8	
BERT	<i>All source</i>	.513	.541	.578	.582	.637	.600	.529	.431	.551
	<i>Proposed</i>	.640	.581	.684	.631	.683	.636	.597	.628	.635
Llama-2-7B	<i>All source</i>	.481	.556	.545	.610	.690	.582	.583	.424	.559
	<i>Proposed</i>	.530	.522	.661	.589	.704	.574	.686	.558	.603
PAES	<i>All source</i>	.654	.583	.612	.605	.730	.565	.706	.542	.625
	<i>Proposed</i>	.787	.600	.588	.588	.747	.573	.737	.560	.648
PMAES	<i>All source</i>	.799	.634	.591	.589	.716	.567	.658	.366	.615
	<i>Proposed</i>	.800	.627	.559	.606	.749	.613	.664	.523	.643

The experiments were conducted in two settings: *All source*, and *Proposed*, and the score prediction accuracy was compared. *All source* is a setting in which each AES model is trained using all source prompt data, which is equivalent to the case where all essays are selected in the proposed method. *Proposed* is a setting in which each AES model is trained using a subset of source prompt data selected using our method. The prediction performance of each trained model is evaluated by QWK using the target prompt essays, excluding 30 data in \mathcal{D}^t .

Table 2 shows the experimental results. The results show that the proposed method outperforms the *All source* settings for all models. The improvement is particularly significant for BERT and Llama-2-7B. These models use the word sequence as the input data, increasing the difference in feature vector characteristics between the source and target prompts. This would enhance the negative impact of using source prompt essays irrelevant to the target prompt, thereby deteriorating the AES model trained using all source prompt data.

For PAES and PMAES, the improvement margin is smaller because they mitigate the difference in the feature space between prompts by using prompt-independent features and POS sequences as input. However, even for these models, the proposed method succeeds in improving their performances by selecting relevant essays that align better with the target prompt’s characteristics.

Moreover, BERT achieves higher performance with the proposed method than does PAES and PMAES without the proposed method. This suggests that the proposed method applied to BERT can achieve performance comparable to these cross-prompt AES models. This is a significant result because it indicates that by simply selecting essays that are effective for the target prompt, it is possible to achieve performance comparable to conventional cross-prompt AES models without relying on complex techniques to align features across prompts.

These results demonstrate the effectiveness of the proposed method in selecting the most relevant essays from source prompts, leading to improved performance of conventional AES models.

5.3. Validity Evaluation of Estimated Data Values

In this section, we investigate whether the value estimates of the proposed method appropriately relate to the score prediction performance. To confirm this point, we examined the prediction

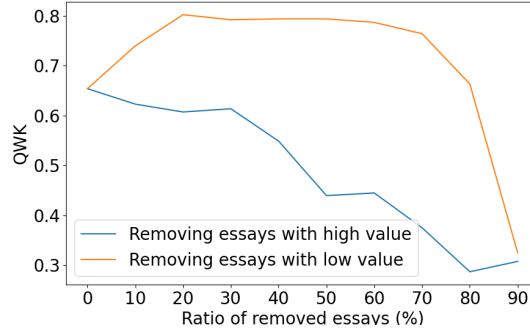


Figure 2: Relationship between the ratio of essays and QWK for Prompt 1.

accuracy, QWK, of an AES model trained using source prompt essays, excluding those with top or bottom $n\%$ value scores. The removing ratio n was changed from 0% to 90% in increments of 10%. This analysis uses PAES as the AES model because, as reported above, it demonstrated the highest performance among the models to which the proposed method was applied.

The experimental results for Prompt 1 are presented in Figure 2, which shows the ratio of excluded essays on the horizontal axis and the QWK on the vertical axis. The blue line represents the QWK when essays are excluded in order of the highest value scores, while the orange line represents the QWK when essays are excluded in order of the lowest value scores.

The figure demonstrates that, for the range where the ratios of removed essays are small to medium, QWK tends to increase as essays with low value scores are sequentially excluded, whereas it tends to decrease when essays with high value scores are sequentially excluded. For the range where the ratios of removed essays are extremely large, both cases revealed low QWK values due to the removal of too many training data, which is a reasonable trend.

These results suggest that the value scores estimated by the proposed method appropriately relate to the effectiveness of the scoring performance of the constructed AES model for the target prompt.

6. Conclusion

This study introduced a novel cross-prompt AES approach that leverages the data valuation method to select source prompt essays valuable to improving the accuracy of the AES model for the target prompt. The experimental results demonstrate the effectiveness of our method in improving the performance of AES models.

In future work, we will perform further analyses of the proposed model aimed at gaining a deeper understanding of its characteristics and behavior. Additional experiments are needed to evaluate the effects of utilizing a small set of scored essays for the target prompt, denoted as \mathcal{D}^t , to train the AES model, in addition to its usage in our DVRL process. We also aim to explore methods that do not rely on \mathcal{D}^t because this requirement may not always be feasible in real-world scenarios. Furthermore, we intend to develop an end-to-end model that integrates the data value estimation and AES components into a single, unified framework. This will enable a more streamlined and efficient approach to cross-prompt AES.

References

- [1] I. D. Erguvan, B. Aksu Dunya, Analyzing rater severity in a freshman composition course using many facet Rasch measurement, *Language Testing in Asia* 10 (2020) 1–20. doi:<https://doi.org/10.1186/s40468-020-0098-3>.
- [2] M. Uto, M. Ueno, A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo, *Behaviormetrika* 47 (2020) 469–496. doi:<https://doi.org/10.1007/s41237-020-00115-7>.
- [3] K. Taghipour, H. T. Ng, A neural approach to automated essay scoring, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1882–1891. doi:10.18653/v1/D16-1193.
- [4] Y. Attali, J. Burstein, Automated essay scoring with e-rater® v.2, *The Journal of Technology, Learning and Assessment* 4 (2006) 1–30. doi:<https://doi.org/10.1002/j.2333-8504.2004.tb01972.x>.
- [5] H. Chen, B. He, Automated essay scoring by maximizing human-machine agreement, in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2013, pp. 1741–1752.
- [6] P. Phandi, K. M. A. Chai, H. T. Ng, Flexible domain adaptation for automated essay scoring using correlated linear regression, in: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 431–439. doi:10.18653/v1/D15-1049.
- [7] M. Dascalu, W. Westera, S. Ruseti, S. Trausan-Matu, H. Kurvers, ReaderBench learns Dutch: Building a comprehensive automated essay scoring system for Dutch language, in: *International Conference on Artificial Intelligence in Education*, Springer, 2017, pp. 52–63. doi:10.1007/978-3-319-61425-0_5.
- [8] P. Hastings, S. Hughes, M. A. Britt, Active learning for improving machine learning of student explanatory essays, in: *International Conference on Artificial Intelligence in Education*, Springer, 2018, pp. 140–153. doi:10.1007/978-3-319-93843-1_11.
- [9] L. Yao, S. J. Haberman, M. Zhang, Prediction of writing true scores in automated scoring of essays by best linear predictors and penalized best linear predictors, *ETS Research Report Series 2019* (2019) 1–27. doi:<https://doi.org/10.1002/ets2.12248>.
- [10] M. Uto, A review of deep-neural automated essay scoring models, *Behaviormetrika* 48 (2021) 1–26. doi:10.1007/s41237-021-00142-y.
- [11] R. Ridley, L. He, X. Dai, S. Huang, J. Chen, Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring, 2020. [arXiv:2008.01441](https://arxiv.org/abs/2008.01441).
- [12] D. Alikaniotis, H. Yannakoudakis, M. Rei, Automatic text scoring using neural networks, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2016, pp. 715–725. doi:10.18653/v1/P16-1068.
- [13] F. Dong, Y. Zhang, Automatic features for essay scoring—an empirical study, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1072–1077. doi:10.18653/v1/D16-1115.
- [14] F. Dong, Y. Zhang, J. Yang, Attention-based recurrent convolutional neural network for automatic essay scoring, in: *Proceedings of the 21st Conference on Computational Natural*

- Language Learning, 2017, pp. 153–162. doi:10.18653/v1/K17-1017.
- [15] Y. Tay, M. Phan, L. A. Tuan, S. C. Hui, SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring, Proceedings of the AAAI Conference on Artificial Intelligence 32 (2018). doi:10.1609/aaai.v32i1.12045.
- [16] Y. Farag, H. Yannakoudakis, T. Briscoe, 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, Volume 1 (Long Papers), Association for Computational Linguistics, 2018, pp. 263–271. doi:10.18653/v1/N18-1024.
- [17] C. Jin, B. He, K. Hui, L. Sun, TDNN: A two-stage deep neural network for prompt-independent automated essay scoring, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 1088–1097. doi:10.18653/v1/P18-1100.
- [18] P. U. Rodriguez, A. Jafari, C. M. Ormerod, Language models and automated essay scoring, 2019. arXiv:1909.09482.
- [19] M. Uto, Y. Xie, M. Ueno, Neural automated essay scoring incorporating handcrafted features, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6077–6088. doi:10.18653/v1/2020.coling-main.535.
- [20] M. Uto, M. Okano, Learning automated essay scoring models using item-response-theory-based scores to decrease effects of rater biases, IEEE Transactions on Learning Technologies 14 (2021) 763–776. doi:10.1109/TLT.2022.3145352.
- [21] T. Shibata, M. Uto, Analytic automated essay scoring based on deep neural networks integrating multidimensional item response theory, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, 2022, pp. 2917–2926.
- [22] Y. Chen, X. Li, PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2023, pp. 1489–1503. doi:10.18653/v1/2023.acl-long.83.
- [23] X. Li, M. Chen, J.-Y. Nie, SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring, Knowledge-Based Systems 210 (2020) 106491. doi:https://doi.org/10.1016/j.knsys.2020.106491.
- [24] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
- [25] R. Yang, J. Cao, Z. Wen, Y. Wu, X. He, Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 1560–1569. doi:10.18653/v1/2020.findings-emnlp.141.
- [26] G.-G. Lee, E. Latif, X. Wu, N. Liu, X. Zhai, Applying large language models and chain-of-thought for automatic scoring, Computers and Education: Artificial Intelligence 6 (2024)

100213. doi:<https://doi.org/10.1016/j.caeai.2024.100213>.

- [27] M. Stahl, L. Biermann, A. Nehring, H. Wachsmuth, Exploring llm prompting strategies for joint essay scoring and feedback generation, 2024. [arXiv:2404.15845](https://arxiv.org/abs/2404.15845).
- [28] Y. Cao, H. Jin, X. Wan, Z. Yu, Domain-adaptive neural automated essay scoring, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, 2020, pp. 1011–1020. doi:10.1145/3397271.3401037.
- [29] Z. Jiang, T. Gao, Y. Yin, M. Liu, H. Yu, Z. Cheng, Q. Gu, Improving domain generalization for prompt-aware essay scoring via disentangled representation learning, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2023, pp. 12456–12470. doi:10.18653/v1/2023.acl-long.696.
- [30] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, Springer International Publishing, 2017. doi:10.1007/978-3-319-58347-1_10.
- [31] X. Glorot, A. Bordes, Y. Bengio, Domain adaptation for large-scale sentiment classification: A deep learning approach, in: Proceedings of the 28th International Conference on International Conference on Machine Learning, Omnipress, 2011, pp. 513–520.
- [32] J. Yoon, S. O. Arik, T. Pfister, Data valuation using reinforcement learning, in: Proceedings of the 37th International Conference on Machine Learning, JMLR.org, 2020.
- [33] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, 2023. [arXiv:2307.09288](https://arxiv.org/abs/2307.09288).
- [34] T. Joachims, Optimizing search engines using clickthrough data, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, 2002, pp. 133–142. doi:10.1145/775047.775067.
- [35] A. Ghorbani, J. Zou, Data shapley: Equitable valuation of data for machine learning, in: International conference on machine learning, PMLR, 2019, pp. 2242–2251.
- [36] J. T. Wang, R. Jia, Data Banzhaf: A robust data valuation framework for machine learning, in: International Conference on Artificial Intelligence and Statistics, 2022.
- [37] S. Choi, S. Hong, K. Lee, S. Lim, ChoiceNet: Robust learning by revealing output correlations, 2020. [arXiv:1805.06431](https://arxiv.org/abs/1805.06431).
- [38] M. Ren, W. Zeng, B. Yang, R. Urtasun, Learning to reweight examples for robust deep learning, in: International conference on machine learning, PMLR, 2018, pp. 4334–4343.
- [39] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, 2021. [arXiv:2006.03654](https://arxiv.org/abs/2006.03654).
- [40] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-

- training with gradient-disentangled embedding sharing, 2021. [arXiv:2111.09543](https://arxiv.org/abs/2111.09543).
- [41] R. J. Williams, Simple statistical gradient-following algorithms for connectionist reinforcement learning, *Machine learning* 8 (1992) 229–256.
- [42] D. Kingma, J. Ba, Adam: A method for stochastic optimization, in: *International Conference on Learning Representations*, 2015.