Integration of Automated Essay Scoring Models using Item Response Theory

Itsuki Aomi, Emiko Tsutsumi, Masaki Uto, and Maomi Ueno

The University of Electro-Communications, Tokyo, Japan
{aomi,tsutsumi,uto,ueno}@ai.lab.uec.ac.jp

Abstract. Automated essay scoring (AES) is the task of automatically grading essays without human raters. Many AES models offering different benefits have been proposed over the past few decades. This study proposes a new framework for integrating AES models that uses item response theory (IRT). Specifically, the proposed framework uses IRT to average prediction scores from various AES models while considering the characteristics of each model for evaluation of examinee ability. This study demonstrates that the proposed framework provides higher accuracy than individual AES models and simple averaging methods.

Keywords: automated essay scoring · item response theory · model averaging

1 Introduction

In recent years, various studies have examined automated essay scoring (AES) models to reduce the costs involved in scoring essays in mass testing. Most AES models can be roughly divided into two approaches: *feature-engineering approach* and *automatic feature extraction approach* [5, 7]. The features-engineering approach manually extracts features (e.g., essay length and number of spelling errors) from given essays and uses these features to predict scores. An important benefit of this approach is its explicability. The approach, however, generally requires careful feature creation and selection to achieve high accuracy. To obviate the need for feature engineering, the automatic feature extraction approach using neural networks has been recently proposed [1, 2, 4, 13, 14, 21]. Such conventional AES models are known to provide different advantages. Therefore, averaging the scores of various AES models is expected to improve scoring accuracy. However, scores that are simply averaged might be inaccurate because each AES model has different accuracy for evaluating examinee ability.

To resolve this problem, we propose a framework that aggregates various AES models using item response theory (IRT) [10], which is a test theory based on mathematical models. In recent years, IRT models that are able to estimate scores while considering the characteristics of human raters, such as rater severity and consistency, have been proposed [3, 8, 11, 15, 17–19]. The present study focuses on the use of such IRT models with AES models instead of human raters. The proposed framework is expected to provide scores that are more accurate

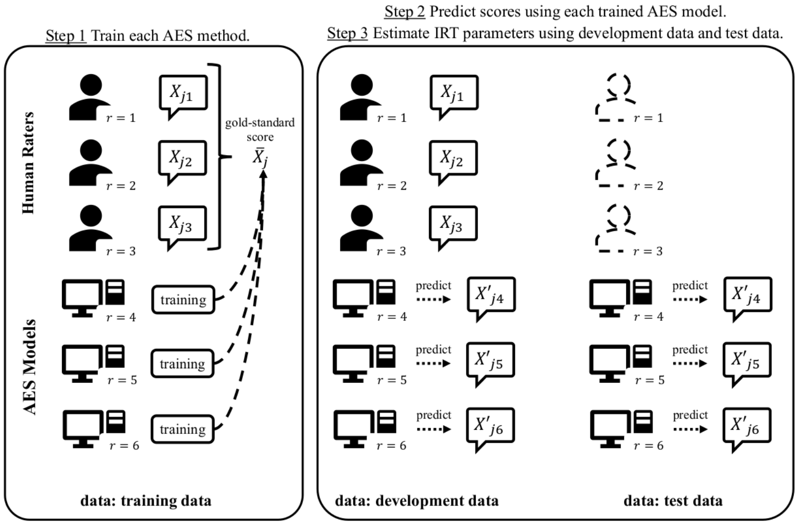

Fig. 1: Proposed framework for three human raters ($r = 1, 2, 3$) and three AES models ($r = 4, 5, 6$). X_{jr} indicates the score given by human-rater r for the essay of examinee j. $\bar{X}_j$ is the average of all the scores given by all the human-raters. X'_{jr} is the prediction score given by the r-th AES model for the essay of examinee j.

than those obtained by simple averaging or a single AES model because the framework can integrate prediction scores from various AES models while considering the characteristics of each model for each examinee's ability level. Our experiments demonstrate that the proposed framework provides higher accuracy than individual AES models and than simply averaged scores.

Of note, Uto and Okano have recently proposed another AES framework that uses IRT [16]. They, however, use IRT to remove rater bias effects within training data to improve the robustness of the model training process. The research objective and the developed framework are completely different from those of the present study.

2 Proposed Framework

In this section, we propose a framework for averaging scores of various AES models in consideration of the characteristics of each model. Fig. 1 shows the outline of the proposed framework. As shown in the figure, the proposed framework executes model training and score prediction through the following four steps: 1) Train each AES model individually using gold-standard scores in training data. 2) Predict scores for essays using development data and test data in each trained AES model. 3) Estimate IRT model parameters from the prediction scores obtained in Step 2. In this estimation, human scores for development data are also used, whereas human scores for test data are not used because they are not given in advance. The IRT models used in this study are the many facet Rasch model (MFRM) [8] and the generalized MFRM (g-MFRM) [18, 19]. The g-MFRM defines the probability that human-rater or AES model $r \in \mathcal{R} = \{1, \ldots, R\}$ gives score k for the essay of examinee $j \in \mathcal{J} = \{1, \ldots, J\}$ as follows.

$$P_{jrk} = \frac{\exp \sum_{m=1}^{k} [\alpha_r(\theta_j - \beta_r - d_{rm})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_r(\theta_j - \beta_r - d_{rm})]}. \tag{1}$$

where θ_j represents the latent ability of examinee j, α_r denotes the consistency of rater r, β_r denotes the strictness of rater r, and d_{rk} represents the severity of rater r within category $k \in \mathrm{K} = \{1, \ldots, K\}$. The MFRM is a special case of g-MFRM when $\alpha_r = 1$ and $d_{rm} = d_m$ for all rater. 4) Calculate the following expectation score $\hat{X}_j$ for essays in test data.

$$\hat{X}_j = \frac{1}{|\mathcal{R}_{\text{human}}|} \sum_{r \in \mathcal{R}_{\text{human}}} \sum_{k=1}^{K} k \cdot P_{jrk}, \tag{2}$$

where $\mathcal{R}_{\text{human}}$ is the set of human raters. This calculation is performed given IRT parameter estimates including the latent examinee ability $\hat{\theta}_j$, which are estimated from multiple AES model predictions in Step 3.

3 Experiments

We evaluate the effectiveness of the proposed framework using the Automated Student Assessment Prize (ASAP) dataset, which has been used in various AES studies [6, 13, 14, 21] and Kaggle competitions[1]. We use five-fold cross validation to evaluate scoring accuracy in terms of quadratic weighted kappa (QWK) which is the common evaluation metric in the ASAP competition.

The following AES models are used in our experiment: Feature-engineering approach models, including **EASE (SVR)**, **EASE (BLRR)** [12], and **XG-Boost** [6, 9]. Automatic feature extraction approach models, including **LSTM-**based model [13] and **SkipFlow** model [14]. We also used a hybrid model **BERT+F** [20] that integrates the feature-engineering approach and automatic feature extraction approach. Model settings, including hyperparameter settings, are the same as those used in the original studies.

The present experiment compares the proposed framework incorporating the model described above with the individual AES models (hereinafter, BASE models), and with two simple model averaging methods; **MEAN** (arithmetic averaging of AES scores) and **VOTING** (hard voting of AES scores). Hereinafter, we call the simple averaging methods as AVG methods.

In the proposed framework, we examine two IRT models: MFRM and g-MFRM. We refer to the proposed frameworks using these IRT models respectively as **Proposal (MFRM)** and **Proposal (g-MFRM)**. The IRT parameter estimation was conducted by Markov chain Monte Carlo following [19].

[1] https://www.kaggle.com/c/asap-aes

Table 1: QWK score of the BASE models and the AVG methods.

		Prompts								
	AES models	1	2	3	4	5	6	7	8	Avg.
BASE	EASE (SVR)	0.558	0.533	0.564	0.571	0.659	0.749	0.545	0.350	0.566
	EASE (BLRR)	0.804	0.603	0.656	0.717	0.784	0.761	0.730	0.675	0.716
	XGBoost	0.814	0.640	0.593	0.660	0.763	0.657	0.692	0.676	0.687
	LSTM	0.777	0.619	0.651	0.730	0.770	0.760	0.750	0.460	0.690
	SkipFlow	0.798	0.652	0.657	0.729	0.783	0.778	0.751	0.614	0.720
	BERT+F	0.827	0.637	0.672	0.620	0.780	0.673	0.720	0.681	0.701
AVG	MEAN	0.820	0.667	0.673	0.730	**0.805**	0.774	0.768	0.678	0.739*
	VOTING	0.833	0.660	**0.675**	0.731	0.794	0.770	0.745	0.666	0.734*
	Proposal (MFRM)	0.821	0.626	0.663	0.685	0.777	0.728	0.768	0.674	0.718*
	Proposal (g-MFRM)	**0.838**	**0.686**	0.668	**0.743**	0.796	**0.785**	**0.793**	**0.717**	**0.753**

Table 1 presents the experimentally obtained results. * indicates that the performance of Proposal (g-MFRM) is higher than that of the other AVG methods at the 5 % significance level by one-tailed paired *t*-test. The results show that Proposal (g-MFRM) provides a higher QWK score than that of all the BASE models except for only one case (BERT+F in prompt 3). Furthermore, Proposal (g-MFRM) achieves the highest QWK score on average.

In Table 1, simple averaging methods are shown to also outperform the BASE models for almost all prompts. Compared with the simple averaging methods, Proposal (g-MFRM) provides a higher QWK score for prompts 1, 2, 4, 6, 7, and 8, but it provides a slightly lower QWK score for prompts 3, and 5. The reason for this improvement is that the proposed framework can estimate scores while considering the characteristics of the respective BASE models. In prompts where Proposal (g-MFRM) provides higher QWK score, the difference in QWK score among the BASE models tends to be large. For example, EASE (SVR) in prompts 1 and 7, XGBoost and BERT+F in prompt 6, and EASE (SVR) and LSTM in prompt 8 show much lower QWK score. Thus, Proposal (g-MFRM) can maintain high scoring accuracy even when models with various characteristics exist, although simple averaging methods can not.

4 Conclusion

In this work, we proposed a new framework for integrating AES models that uses IRT. We described how simply averaged scores can lower evaluating accuracy because each AES model has a different assessment accuracies for scoring examinee ability. To resolve this issue, we presented the idea of estimating scores using IRT models while considering the characteristics of the AES models. Based on experiment results, we demonstrated that the proposed framework with a latent IRT model provides higher accuracy than individual AES models and higher accuracy than simply averaged scores.

References

1. Alikaniotis, D., Yannakoudakis, H., Rei, M.: Automatic text scoring using neural networks. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 715–725 (2016)
2. Dasgupta, T., Naskar, A., Dey, L., Saha, R.: Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In: Proceedings of the Fifth Workshop on Natural Language Processing Techniques for Educational Applications. pp. 93–102 (2018)
3. Eckes, T.: Introduction to Many-Facet Rasch Measurement. Peter Lang, Bern (2015)
4. Farag, Y., Yannakoudakis, H., Briscoe, T.: 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). pp. 263–271 (2018)
5. Hussein, M.A., Hassan, H.A., Nassef, M.: Automated language essay scoring systems: a literature review. PeerJ Computer Science **5**, e208 (2019)
6. Jin, C., He, B., Hui, K., Sun, L.: 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). pp. 1088–1097 (2018)
7. Ke, Z., Ng, V.: Automated essay scoring: A survey of the state of the art. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. pp. 6300–6308 (2019)
8. Linacre, J.M.: Many-facet Rasch measurement. MESA Press, Chicago (1989)
9. Liu, J., Xu, Y., Zhu, Y.: Automated Essay Scoring based on Two-Stage Learning. arXiv e-prints **arXiv:1901.07744**, arXiv:1901.07744 (Jan 2019)
10. Lord, F.M.: Applications of item response theory to practical testing problems. Routledge, Abingdon-on-Thames (1980)
11. Myford, C.M., Wolfe, E.W.: Detecting and measuring rater effects using many-facet rasch measurement: part I. Journal of Applied Measurement **4**(4), 386–422 (2003)
12. Phandi, P., Chai, K.M.A., Ng, H.T.: Flexible domain adaptation for automated essay scoring using correlated linear regression. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 431–439 (2015)
13. Taghipour, K., Ng, H.T.: A neural approach to automated essay scoring. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1882–1891 (2016)
14. Tay, Y., Phan, M., Luu, A.T., Hui, S.C.: SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring. In: Thirty-Second AAAI Conference on Artificial Intelligence. pp. 5948–5955 (2018)
15. Ueno, M., Okamoto, T.: Item response theory for peer assessment. In: 2008 Eighth IEEE International Conference on Advanced Learning Technologies. pp. 554–558 (2008). https://doi.org/10.1109/ICALT.2008.118
16. Uto, M., Okano, M.: Robust neural automated essay scoring using item response theory. In: Artificial Intelligence in Education. pp. 549–561 (2020)
17. Uto, M., Ueno, M.: Item response theory for peer assessment. IEEE Transactions on Learning Technologies **9**(2), 157–170 (2016)

18. Uto, M., Ueno, M.: Item response theory without restriction of equal interval scale for rater's score. In: Artificial Intelligence in Education. pp. 363–368 (2018)
19. Uto, M., Ueno, M.: A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. Behaviormetrika **47**, 469–496 (2020)
20. Uto, M., Xie, Y., Ueno, M.: Neural automated essay scoring incorporating handcrafted features. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 6077–6088 (2020)
21. Wang, Y., Wei, Z., Zhou, Y., Huang, X.: Automatic essay scoring incorporating rating schema via reinforcement learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 791–797 (2018)