A Multi-Task Automated Assessment System for Essay Scoring

Shigeng Chen[1][0009−0003−4332−5817], Yunshi Lan[2][0000−0002−0192−8498], and Zheng Yuan[1][0000−0003−2406−1708]

[1] Department of Informatics, King's College London, UK
{shigeng.chen, zheng.yuan}@kcl.ac.uk
[2] School of Data Science and Engineering, East China Normal University, China
yslan@dase.ecnu.edu.cn

Abstract. Most existing automated assessment (AA) systems focus on holistic scoring, falling short in providing learners with comprehensive feedback. In this paper, we propose a Multi-Task Automated Assessment (MTAA) system that can output detailed scores along multiple dimensions of essay quality to provide instructional feedback. This system is built on multi-task learning and incorporates Orthogonality Constraints (OC) to learn distinct information from different tasks. To achieve better training convergence, we develop a training strategy, Dynamic Learning Rate Decay (DLRD), to adapt the learning rates for tasks based on their loss descending rates. The results show that our proposed system achieves state-of-the-art performance on two benchmark datasets: ELLIPSE and ASAP++. Furthermore, we utilize ChatGPT to assess essays in both zero-shot and few-shot contexts using an ELLIPSE subset. The findings suggest that ChatGPT has not yet achieved a level of scoring consistency equivalent to our developed MTAA system and that of human raters.

Keywords: Automated Essay Scoring · Multi-Task Learning · ChatGPT Automated Assessment · Zero-Shot Learning · Few-Shot Learning

1 Introduction

Automated assessment (AA), mimicking the judgment of examiners evaluating the quality of student writing, is one of the most important educational Natural Language Processing (NLP) applications. Originally used for summative purposes in standardised testing such as the TOEFL[3] and GRE[4], these systems are now frequently found in classrooms [4, 5].

Traditional ML-based AA systems typically rely on hand-crafted features and models like SVM and linear regression have been proposed [6–9], while neural-based AA systems often employ word embeddings like GloVe [10] and incorporate deep neural networks such as CNN [11] and LSTM [12] to achieve better system

[3] https://www.ets.org/toefl.html
[4] https://www.ets.org/gre.html

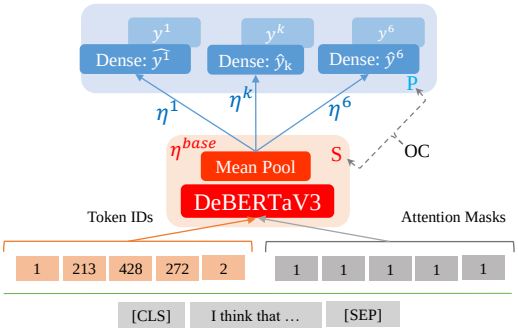

Fig. 1: The architecture of MTAA.

performance [13–16]. In contrast, Transformer-based AA systems which leverage large pre-trained language models like DistilBERT [23] and XLNet [22] outperform their counterparts by handling long-distance relationships and exhibiting strong generalization abilities [17–19]. Most prior AA systems focusing on holistic scoring are unable to provide in-depth feedback to learners. A few studies [29–31] have explored the development of an AA system to support multi-dimensional essay evaluation. However, they often neglect the inter-connectedness among various assessment measures.

In this paper, we propose a Multi-Task Automated Assessment (MTAA) system that eliminates the need for feature engineering and evaluates essays across various dimensions of essay quality. Specifically, a multi-task learning (MTL) framework has been designed where Dynamic Learning Rate Decay (DLRD) has been employed to promote balanced training across different tasks, and Orthogonality Constraints (OC) [24] have been employed to facilitate the encoding of various facets of the inputs from the shared and task-specific networks. We evaluate our system on two public benchmarks, ELLIPSE [1] and ASAP++ [2], and new state-of-the-art results have been achieved. In addition, we engage ChatGPT, which has recently been used for automatic scoring [27,28], in a comparative evaluation under both zero-shot and few-shot conditions. The results show that ChatGPT has not yet reached the scoring consistency of our developed MTAA system and that of human raters.

2 Multi-Task Automated Assessment

2.1 Architecture Design

MTAA, as shown in Figure 1, is a model with hard parameter sharing [25] that utilizes a backbone as a shared encoder to optimize multiple tasks and task-specific decoders to perform predictions. The shared encoder of the model consists of a pre-trained base version of DeBERTaV3 [26] and a mean pooling layer with the former capturing intricate information from the inputs and the

latter extracting compressed shared representations. The task-specific encoder is structured with various branches, each one consisting of two densely connected layers for a task. These layers are utilized to extract knowledge from shared representations and to evaluate an essay across various dimensions. OC and DLRD are integrated into the MTL model to extract specific task-related information and promote faster training convergence.

2.2 Orthogonality Constraints

To encourage the shared and task-specific encoders to encode diverse information facets of tasks within the MTL framework, we incorporate Orthogonality Constraints (OC), as introduced in [24], into our MTAA system:

$$\ell_{oc}^{k} = \sum_{k=1}^{m} \left\| \mathbf{S}^{\top} \mathbf{P}^{k} \right\|_{F}^{2} \tag{1}$$

where $\| \cdot \|_F^2$ refers to the squared Frobenius norm. $\mathbf{S}$ and $\mathbf{P}^k$ are two matrices, whose rows are the shared and task-specific representations, as shown in Figure 1.

2.3 Dynamic Learning Rate Decay

To ensure effective and balanced MTL learning, as well as preventing certain tasks from overpowering others during model optimization, we propose Dynamic Learning Rate Decay(DLRD) to adapt different learning rates for different tasks. Specifically, DLRD keeps a moderate learning rate for the shared encoder and assigns smaller learning rates to task-specific encoders exhibiting high learning speeds. The DLRD involves two steps:

1. Calculating task weights:

$$\omega^{k}(t) = \frac{r^{k}(t-1)^{\alpha}}{\bar{r}}, r^{k}(t-1) = \frac{\ell^{k}(t-2)}{\ell^{k}(t-1)} \tag{2}$$

   where t is the index of training iteration. Exponent α serves as a factor for adjusting the magnitude of differences in task weights and a greater value of α (>1) amplifies the disparities among task weights. The average value $\bar{r}$ scales the weights to prevent dominance by tasks with higher loss descent rates.
2. Dynamical learning rate decay:

$$\eta^{k} = \eta^{\text{base}}\, \omega^{k}(t), \eta^{\text{base}} \; = \eta_0 \gamma^{\left\lfloor \frac{t}{n} \right\rfloor} \tag{3}$$

   where γ represents the decay factor. Floor function $\left\lfloor \frac{t}{n} \right\rfloor$ represents the frequency of learning rate decay, occurring every certain number of epochs. η^{base} is the learning rate for the shared network, and learning rate η^k is calculated for the task-specific network based on its task weight. In our implementation, we set $\eta_0 = 1e-5$, $\alpha = 10$, and $\gamma = 0.3$. Additionally, we set $n = 1$ to perform learning rate decay in every epoch.

3 Experiments

3.1 Datasets

Our system is developed using two multi-dimensional AA benchmark datasets. We further divide both datasets into 80% for training and 20% for testing.[5]

ELLIPSE [1] consists of 3,911 argumentative essays written by English language learners in grades 8-12, with each sample comprising a text and a corresponding score set representing Cohesion, Grammar, Vocabulary, Phraseology, Syntax, and Conventions levels.[6] Each essay was independently rated by two expert annotators using a five-point scoring rubric. Score discrepancies of two or more points were resolved through annotating team discussion. The average word count of the essays is 430, with most falling within 250 to 500 words. The scores range from 1.0 to 5.0 in increments of 0.5, where each of the six dimensions demonstrates an approximate normal distribution, with the mean at about 3.0. We also observe that the Pearson correlation coefficients among all six evaluation dimensions exceed 0.6, suggesting a substantial positive linear relationship.

ASAP++ [2] has been developed on top of ASAP [3], offering multi-dimensional scores for first six prompts. Prompts 1-2 assess argumentative essays on Content, Organization, Word Choice, Sentence Fluency, and Conventions. Prompts 3-6 evaluate source-dependent essays on Content, Prompt Adherence, Language, and Narrativity. For details on attribute definitions and essay statistics, please refer to the original paper.

3.2 Metrics

The performance of the models is assessed by a broad range of AA metrics, including the root mean square error (RMSE), Pearson correlation coefficient (PCC) and Spearman's rank correlation coefficient (SCC), and the Quadratic Weighted Kappa (QWK), all of which are widely adopted in AA [6,8,9,13–17,29–32]. We evaluate the model performance using a column-wise mean technique. Specifically, for each task, we compute metric scores based on actual-predicted pairs and then average these scores across all tasks/dimensions to obtain an overall assessment metric value.

3.3 Results

We compare our proposed MTAA system with several robust baselines, including BERT and RoBERTa, as well as MTL_vanilla, which maintains the same network structure as the MTAA but does not incorporate the OC and DLRD mechanisms.

[5] The train/test split can be found at https://github.com/Aries-chen/MTAA/blob/main/README.md.

[6] The ELLIPSE rubric is available at: https://docs.google.com/document/d/1OSbRELoWKlq8chYmujAaHJqMwFZnwt2PnnbSXfOJkIY/edit.

Table 1: Performance on ELLIPSE. The abbreviations are: Coh. for Cohesion, Syn. for Syntax, Voc. for Vocabulary, Phr. for Phraseology, Gra. for Grammar, and Con. for Conventions. Avg. represents the average scores across all dimensions. ChatGPT_0 and ChatGPT_3 denote the use of ChatGPT in zero-shot and few-shot settings, respectively. The best scores for each metric are highlighted in bold.

Datasets	Models	RMSE ↓							PCC ↑						
		Coh.	Syn.	Voc.	Phr.	Gra.	Con.	Avg.	Coh.	Syn.	Voc.	Phr.	Gra.	Con.	Avg.
ELLIPSE	BERT	.54	.47	.48	.47	.52	.48	.49	.60	.69	.65	.68	.65	.68	.66
	RoBERTa	.51	.46	.43	.45	.49	.46	.47	.65	.70	.68	.71	.72	.73	.70
	$\text{MTL}_{\text{vanilla}}$	.51	.45	.43	.44	.48	.46	.46	.66	.72	.69	.73	.73	.73	.71
	MTAA	.51	.45	.42	.44	.46	.44	**.45**	.66	.72	.70	.73	.74	.75	**.72**
$\text{ELLIPSE}_{\text{Subset}}$	ChatGPT_0	.89	.67	.76	.89	.88	.74	.80	.26	.64	.63	.50	.52	.62	.53
	ChatGPT_3	.81	.59	.58	.69	.81	.62	.68	.26	.63	.62	.58	.46	.56	.52
	MTAA	.58	.38	.36	.44	.50	.46	**.45**	.54	.72	.74	.70	.72	.68	**.68**
Datasets	**Models**	**SCC ↑**							**QWK ↑**						
		Coh.	Syn.	Voc.	Phr.	Gra.	Con.	Avg.	Coh.	Syn.	Voc.	Phr.	Gra.	Con.	Avg.
ELLIPSE	BERT	.57	.66	.63	.65	.63	.66	.63	.56	.64	.61	.65	.61	.63	.62
	RoBERTa	.62	.67	.65	.69	.71	.71	.67	.61	.66	.65	.67	.69	.71	.66
	$\text{MTL}_{\text{vanilla}}$	.63	.69	.67	.72	.72	.71	**.69**	.62	.69	.64	.69	.70	.70	.67
	MTAA	.63	.69	.67	.71	.73	.72	**.69**	.63	.69	.67	.69	.72	.71	**.68**
$\text{ELLIPSE}_{\text{Subset}}$	ChatGPT_0	.21	.59	.64	.49	.48	.61	.50	.12	.33	.28	.22	.32	.42	.29
	ChatGPT_3	.19	.58	.63	.55	.41	.53	.48	.25	.56	.56	.51	.43	.55	.48
	MTAA	.40	.59	.66	.70	.74	.64	**.62**	.49	.67	.68	.69	.69	.65	**.64**

Furthermore, we compare our proposed system with ChatGPT in both zero-shot and few-shot (i.e. 3-shot) manners.[7] Due to budget constraints, we utilized a representative subset of ELLIPSE that maintains the percentage of samples at each score level from the test set.[8]

Results on ELLIPSE are presented in Table 1. We can see that the proposed MTAA demonstrates superior performance compared to the baselines (i.e., BERT, RoBERTa, and $\text{MTL}_{\text{vanilla}}$) on the evaluated ELLIPSE dataset. Specifically, it achieves scores of 0.45 for RMSE, 0.72 for PCC, 0.69 for SCC, and 0.68 for QWK, setting new state-of-the-art performance. When evaluated on the subset, our MTAA model significantly outperforms ChatGPT in both zero-shot and few-shot settings. While these methods show similar performance in terms of PCC and SCC, the few-shot approach significantly excels over the zero-shot one for RMSE and QWK. We also notice that all models yield the worst performance on Cohesion compared to other dimensions.

[7] We used the GPT-4-0613 API. The prompts used in our experiments are available at https://github.com/Aries-chen/MTAA/blob/main/Few-shot_prompt.txt.

[8] This comparison is excluded from ASAP++ due to its lack of a clear evaluation rubric, making it difficult to provide precise prompts for ChatGPT inputs.

Table 2: Performance on ASAP++. The abbreviations are: Cont. for Content, Org. for Organization, WoCh. for Word Choice, SeFl. for Sentence Fluency, Conv. for Conventions, PrAd. for Prompt Adherence, Lang. for Language, and Narr. for Narrativity. Avg. represents the average scores across all dimensions. The best scores for each metric are highlighted in bold.

Metrics	Models	Argumentative essays						Source-dependent essays				
		Cont.	Org.	WoCh.	SeFl.	Conv.	Avg.	Cont.	PrAd.	Lang.	Narr.	Avg.
RMSE ↓	MTAA	.76	.73	.73	.68	.71	**.72**	.56	.57	.62	.58	**.58**
PCC ↑	MTAA	.76	.76	.75	.75	.74	**.75**	.84	.83	.80	.81	**.82**
SCC ↑	MTAA	.75	.75	.73	.74	.73	**.74**	.84	.83	.79	.80	**.82**
QWK ↑	MTAA	.72	.70	.70	.72	.70	**.71**	.80	.80	.74	.76	**.77**
	Ridley et al. (2020) [30]	.54	.41	.53	.54	.36	.48	.54	.57	.53	.61	.56
	Ridley et al. (2021) [31]	.56	.46	.56	.55	.41	.51	.56	.57	.54	.61	.57
	Chen & Li. (2023) [32]	.57	.48	.58	.58	.42	.53	.57	.58	.55	.61	.58

Results on ASAP++ are presented in Table 2. Again, our proposed MTAA system outperforms all its competitors [30–32] by a large margin when evaluated on QWK and yields new state of the art.[9]

4 Discussion

MTL is particularly well-suited for multi-dimensional AA tasks due to their ability to first extract shared information before exploiting task-specific information. This advantage stems from the fact that the assessment measures in these tasks are often related, but not necessarily identical. The benefits of our approach are further enhanced by the integration of OC and DLRD, which promote task-specific representations while effectively balancing learning speeds across tasks. Furthermore, the proposed design eliminates the need for manual task weight adjustment, thereby making the model more robust and generalizable.

Regarding the low performance in Cohesion compared to all the other dimensions, we speculate that the abstract and complex scoring measure poses significant challenges for AA systems. Making use of detailed and concrete features, e.g. 'reference and transitional words and phrases' (as outlined in the ELLIPSE Cohesion Rubric), might be beneficial.

The performance of ChatGPT in our multi-dimensional AA task is much lower than those reported in other NLP tasks. Upon analyzing the zero-shot outputs, we discovered that the scores across all measures were consistently lower than those provided by human raters. This observation suggests that ChatGPT acts as a more "stringent" assessor. However, when we provided it with three essay-score pairs for few-shot evaluation, ChatGPT became more "lenient" and the scores aligned more closely with those given by human raters. This is evidenced by the improvements in RMSE and QWK as shown in Table 1. Despite

[9] Previous work has only reported QWK.

ChatGPT's underperformance in our task, it possesses unique strengths, such as providing more specific feedback [27], including grammar corrections and word suggestions.

5 Conclusion

We proposed a MTAA system that supports multi-dimensional essay scoring. Specifically, we introduced OC to obtain more task-specific representations and designed DLRD to dynamically adjust the learning rates for the tasks to achieve balanced training. Our system achieves state of the art on ELLIPSE and ASAP++ public benchmarks. Additionally, we explored the potential of ChatGPT in multi-dimensional AA and found that ChatGPT has not yet matched the consistency of our MTAA system or that of human raters. Our future research interests lie in investigating the performance of ChatGPT in providing multi-dimensional feedback, such as offering detailed and constructive suggestions in addition to scoring. We aim to develop a more comprehensive and effective AA system that not only assigns scores but also guides students towards improving their writing skills across multiple dimensions.

References

1. Franklin, A., Maggie, Benner, M., Rambis, N., Baffour, P., Holbrook, R., Crossley, S., Boser, U.: Feedback Prize - English Language Learning. Kaggle (2022). https://kaggle.com/competitions/feedback-prize-english-language-learning
2. Mathias, S., Bhattacharyya, P.: ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In: LREC (2018)
3. Hamner, B., Morgan, J., Vandev, L., Shermis, M., Vander Ark, T.: The Hewlett Foundation: Automated Essay Scoring. Kaggle (2012). https://kaggle.com/competitions/asap-aes
4. Ramineni, C., Trapani, C., Williamson, D., Davey, T., Bridgeman, B.: Evaluation of the e-rater® Scoring Engine for the TOEFL® Independent and Integrated Prompts. ETS Research Report Series, Wiley Online Library, vol. 2012, no. 1, pp. i–51 (2012)
5. Ramineni, C., Trapani, C., Williamson, D., Davey, T., Bridgeman, B.: Evaluation of e-rater for the GRE issue and argument prompts. Educational Testing Service Princeton, NJ (2012)
6. Yannakoudakis, H., Briscoe, T., Medlock, B.: A New Dataset and Method for Automatically Grading ESOL Texts. In: ACL-HLT, pp. 180–189 (2011)
7. Contreras, J. O., Hilles, S., Abubakar, Z. B.: Automated Essay Scoring with Ontology based on Text Mining and NLTK tools. In: ICSCEE, pp. 1–6. IEEE (2018)
8. Kumar, Y., Aggarwal, S., Mahata, D., Shah, R. R., Kumaraguru, P., Zimmermann, R.: Get It Scored Using AutoSAS - An Automated System for Scoring Short Answers. CoRR, abs/2012.11243 (2020)
9. Phandi, P., Chai, K. M. A., Ng, H. T.: Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression. In: EMNLP, pp. 431–439 (2015)
10. Pennington, J., Socher, R., Manning, C.: GloVe: Global Vectors for Word Representation. In: EMNLP, pp. 1532–1543 (2014)

11. Kim, Y.: Convolutional Neural Networks for Sentence Classification. CoRR, abs/1408.5882 (2014)
12. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation, 9(8), pp. 1735-1780 (1997)
13. Taghipour, K., Ng, H. T.: A neural approach to automated essay scoring. In: EMNLP, pp. 1882–1891 (2016)
14. Dong, F., Zhang, Y., Yang, J.: Attention-based Recurrent Convolutional Neural Network for Automatic Essay Scoring. In: CoNLL, pp. 153–162 (2017)
15. Wang, Y., Wei, Z., Zhou, Y., Huang, X.: Automatic Essay Scoring Incorporating Rating Schema via Reinforcement Learning. In: EMNLP, pp. 791–797 (2018)
16. Tay, Y., Phan, M., Tuan, L.A., Hui, S.C.: Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In: AAAI, 32(1) (2018)
17. Rodriguez, P.U., Jafari, A., Ormerod, C.M.: Language models and Automated Essay Scoring. CoRR, abs/1909.09482 (2019)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems, 30. Curran Associates, Inc. (2017)
19. Andersen, Ø. E., Yuan, Z., Watson, R., Cheung, K. Y. F.: Benefits of Alternative Evaluation Methods for Automated Essay Scoring. In: International Educational Data Mining Society (2021)
20. Zhang, A., Chan, A., Tay, Y., Fu, J., Wang, S., Zhang, S., Shao, H., Yao, S., Lee, R. K.-W.: On Orthogonality Constraints for Transformers. In: ACL-IJCNLP '21 (Vol. 2: Short Papers), pp. 375–382 (2021)
21. Hamner, B., Morgan, J., lynnvandev, Shermis, M., Vander Ark, T.: The Hewlett Foundation: Automated Essay Scoring. Kaggle (2012)
22. Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., Le, Q.V.: XL-Net: Generalized Autoregressive Pretraining for Language Understanding. CoRR, abs/1906.08237 (2019)
23. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: CoRR, abs/1910.01108 (2019)
24. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain Separation Networks. In: CoRR, abs/1608.06019 (2016)
25. Caruana, R.: Multitask learning. Machine learning, vol. 28, pp. 41–75. Springer (1997)
26. He, P., Gao, J., Chen, W.: DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In: CoRR, abs/2111.09543 (2021)
27. Yoon, S., Miszoglad, E., Pierce, L.R.: Evaluation of ChatGPT Feedback on ELL Writers' Coherence and Cohesion. arXiv preprint arXiv:2310.06505 (2023)
28. Wu, X., He, X., Liu, T., Liu, N., Zhai, X.: Matching exemplar as next sentence prediction (mensp): Zero-shot prompt learning for automatic scoring in science education. In: AIED, Springer, pp. 401–413 (2023)
29. Ke, Z., Inamdar, H., Lin, H., Ng, V.: Give me more feedback II: Annotating thesis strength and related attributes in student essays. In: ACL, pp. 3994–4004 (2019)
30. Ridley, R., He, L., Dai, X., Huang, S., Chen, J.: Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring. CoRR, abs/2008.01441 (2020)
31. Ridley, R., He, L., Dai, X., Huang, S., Chen, J.: Automated cross-prompt scoring of essay traits. In: AAAI, vol. 35, no. 15, pp. 13745–13753 (2021)
32. Chen, Y., Li, X.: PMAES: Prompt-mapping Contrastive Learning for Cross-prompt Automated Essay Scoring. In: ACL, pp. 1489–1503 (2023)