Towards Prompt Generalization: Grammar-aware Cross-Prompt Automated Essay Scoring

Heejin Do[1], **Taehee Park**[1], **Sangwon Ryu**[1], **Gary Geunbae Lee**[1,2]
[1]Graduate School of Artificial Intelligence, POSTECH, Republic of Korea
[2]Department of Computer Science and Engineering, POSTECH, Republic of Korea
{heejindo, taehpark, ryusangwon, gblee}@postech.ac.kr

Abstract

In automated essay scoring (AES), recent efforts have shifted toward cross-prompt settings that score essays on unseen prompts for practical applicability. However, prior methods trained with essay-score pairs of specific prompts pose challenges in obtaining prompt-generalized essay representation. In this work, we propose a grammar-aware cross-prompt trait scoring (GAPS), which internally captures prompt-independent syntactic aspects to learn generic essay representation. We acquire grammatical error-corrected information in essays via the grammar error correction technique and design the AES model to seamlessly integrate such information. By internally referring to both the corrected and the original essays, the model can focus on generic features during training. Empirical experiments validate our method's generalizability, showing remarkable improvements in prompt-independent and grammar-related traits. Furthermore, GAPS achieves notable QWK gains in the most challenging cross-prompt scenario, highlighting its strength in evaluating unseen prompts.

1 Introduction

Automated essay scoring (AES) emerged as a viable alternative to human graders to assist language learners in acquiring writing skills, alleviating the burden and costs of grading. To practically supply AES in educational situations, the model's capability to generalize well to new prompts (i.e., unseen in training) is essential yet challenging (Li and Ng, 2024). Accordingly, unlike the earlier prompt-specific AES systems, which aim to assess essays written on the seen prompts (Taghipour and Ng, 2016; Dong and Zhang, 2016; Wang et al., 2022; Do et al., 2024a,b), recent attention increasingly moves on cross-prompt AES to grade new prompts' essays (Ridley et al., 2021; Do et al., 2023; Chen and Li, 2024; Li and Ng, 2024).

To achieve cross-prompt scoring with multiple trait setting, previous studies primarily focus on learning essay representation with score labels in concatenation with prompt-independent features (Jin et al., 2018; Ridley et al., 2020; Li et al., 2020; Ridley et al., 2021; Chen and Li, 2023; Do et al., 2023; Li and Ng, 2024). To obtain consistent essay representations, some studies additionally developed contrastive learning (Chen and Li, 2023, 2024) or prompt-aware networks (Do et al., 2023; Jiang et al., 2022). However, as these models are typically trained on essays responding to specific prompts differ from the target, generalizability to unseen prompts still remains a challenge. Notably, they exhibit the lowest performance in the most prompt-agnostic *Conventions* trait (20% gap to the best trait), further highlighting the shortcomings.

In this work, we propose a grammar-aware cross-prompt essay trait scoring, which integrates grammar error correction (GEC) before the scoring process. By informing the model of the syntactic errors contained in the essay, our method facilitates capturing generic syntax information during scoring. Internally, we design a shared structure to trade knowledge between original and corrected essays, facilitating accurate score derivation. As non-semantic aspects are less dependent on prompts, our grammar-aware learning via directly providing error-corrected essays leads to the intrinsic acquisition of generic essay representation.

Empirical experiments demonstrate that our grammar-aware method assists in capturing generic aspects, enhancing related trait-scoring performances in cross-prompt settings. Notably, significant enhancements observed in prompt-agnostic traits, such as *Conventions* and *Sentence Fluency*, support the advancement towards prompt generalized representation. Surprisingly, informing error-corrected essays also improves semantic traits, such as *Content* and *Narrativity*, suggesting that referring to a revised essay has the potential for facili-

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2818–2824
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

tating a deeper contextual understanding.

2 Related Works

To improve the performance of AES, several studies have auxiliary trained the model with various other tasks such as morpho-syntactic labeling, type and quality prediction, and sentiment analysis (Craighead et al., 2020; Ding et al., 2023; Muangkammuen and Fukumoto, 2020). Instead of jointly training auxiliary tasks, our direct use of corrected text output significantly reduces the training burden.

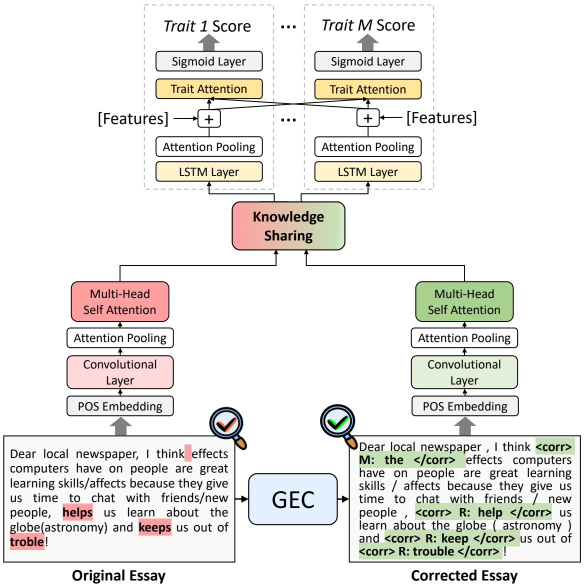
There have been attempts to apply grammatical error information. Suggesting that detecting grammatical errors is a beneficial indicator for the quality of the essay, Cummins and Rei (2018) jointly trained grammatical error detection task with the scoring model. Also, Doi et al. (2024) utilize grammatical features proposed by Hawkins and Filipović (2012) and Liu et al. (2019) use GEC to measure the number of grammar corrections. Unlike the existing studies, we directly utilize the text output generated by the GEC without additional training as input for the scoring model.

3 Method: GAPS

Our method comprises two main steps: (1) Essay correction and (2) Grammar-aware Essay Scoring. Initially, we automatically identify the grammar errors included in the essay and then pass them to the scoring model along with the original essay.

3.1 Essay Correction

We employ the T5-based pre-trained GEC model (Rothe et al., 2021) to obtain the grammar-corrected essay text without additional training. The student's original essay, which contained diverse types of errors, is input into the model, and the cleaned essay is output. Grounded on one of the representative error types presented in the error annotation toolkit (ERRANT) (Bryant et al., 2017)[1], we classify errors into three major categories: Missing (M), Replacement (R), and Unnecessary (U). Missing refers to a required token that is not present but must be inserted, replacement indicates the substituted token that is revised, and unnecessary means the deleted token that does not fit in the syntax. For the input essay, we add the correction tag, `<corr> Category: Token </corr>`, for the identified error corrections. For instance, in Figure 1, for the missed token "`the`" in the essay, `<corr> M: the </corr>` is applied. Explicitly notifying the revisions could place greater emphasis during training.



Dear local newspaper, I think effects computers have on people are great learning skills/affects because they give us time to chat with friends/new people, **helps** us learn about the globe(astronomy) and **keeps** us out of **troble**!

Dear local newspaper , I think **<corr> M: the </corr>** effects computers have on people are great learning skills / affects because they give us time to chat with friends / new people , **<corr> R: help </corr>** us learn about the globe (astronomy) and **<corr> R: keep </corr>** us out of **<corr> R: trouble </corr>** !

Figure 1: The overview of the proposed GAPS method.

3.2 Grammar-aware Essay Scoring

Essay encoders We construct individual essay encoders for the original and corrected essays, respectively, but with the same structure. Our design intends to first understand each document and then share the informed knowledge. We believe enabling the model to internally distinguish each element separately, rather than combining or concatenating them, facilitates more sophisticated information exchange in subsequent layers.

We adopt a hierarchical structure for essay encoding, which obtains trait-specific document-level representations based on sentence-level representations (Dong et al., 2017; Ridley et al., 2021). To obtain a generalized representation, we employ part-of-speech (POS) embedding[2], as in previous studies (Ridley et al., 2021; Do et al., 2023; Chen and Li, 2024). After passing through POS embedding, the output c_i from the 1D convolution layer (Kim, 2014) is subjected to attention pooling layer (Dong et al., 2017): $\mathbf{s} = \text{Pooling}_{\text{att}}([c_1 : c_w])$, where w denotes the number of words in the sentence. To effectively capture all parts of the essays, we adopt multi-head self-attention (Vaswani et al.,

[1] `https://github.com/chrisjbryant/errant`

[2] NLTK toolkit is used: `https://www.nltk.org/`

2017), motivated by (Do et al., 2023):

$$\mathrm{H_i} = \mathrm{att}(SW_i^1, SW_i^2, SW_i^3) \quad (1)$$

$$\mathrm{M} = \mathrm{concat}(\mathrm{H_1}, ..., \mathrm{H_h})W^O \quad (2)$$

where $\mathrm{H_i}$ and $\mathtt{att}$ indicate the i-th head and the scaled-dot product attention, respectively. $W_i^{1\cdot 3}$ is the parameter matrices. Then, the LSTM layer (Hochreiter and Schmidhuber, 1997) is applied, followed by the attention pooling, obtaining the original essay vector, E_o, and the grammar-corrected essay vector, E_g.

Knowledge-sharing layers Given two representations of the original and the grammar-corrected essay, we introduce the knowledge-sharing layers via the cross-attention leveraging multi-head attention mechanism. Specifically, with the original essay vector E_o as the key and value and the grammar-corrected vector E_g as the query, the knowledge-sharing layer is defined as follows:

$$\mathrm{H_i} = \mathrm{att}(E_g W_i^1, E_o W_i^2, E_o W_i^3) \quad (3)$$

$$\mathrm{M} = \mathrm{concat}(\mathrm{H_1}, ..., \mathrm{H_h})W^O \quad (4)$$

Subsequently, m trait-specific layers are obtained for m distinct traits. Following previous studies, we concatenate the prompt-independent features of Ridley et al. (2021) to each trait-wise essay representation vector. To refer to other traits' representations during training, we employ the trait-attention mechanism (Ridley et al., 2021).

Training For the loss function, we use mean squared error: $\mathrm{L}(y, \hat{y}) = \frac{1}{\mathrm{n \cdot m}} \sum_{i=1}^{\mathrm{n}} \sum_{j=1}^{\mathrm{m}} (\hat{y}_{ij} - y_{ij})^2$, with n number of essays and m trait scores. As different prompts are evaluated by different traits (Apendix 1), the masking mechanism is applied to mark empty traits as 0 (Ridley et al., 2021).

4 Experiments

For experiments, we use the Automated Student Assessment Prize (ASAP[3]) and ASAP++[4] (Mathias and Bhattacharyya, 2018) dataset, which are publicly available and representative for AES. The dataset includes eight prompts and corresponding essays written in English, and multiple trait scores are assigned by human raters (Table 1).

Pr	Evaluated Traits	# Essays	Essay Type
P1	Cont, Org, WC, SF, Conv	1,783	Argumentative
P2	Cont, Org, WC, SF, Conv	1,800	Argumentative
P3	Cont, PA, Lan, Nar	1,726	Source-Dependent
P4	Cont, PA, Lan, Nar	1,772	Source-Dependent
P5	Cont, PA, Lan, Nar	1,805	Source-Dependent
P6	Cont, PA, Lan, Nar	1,800	Source-Dependent
P7	Cont, Org, Conv, Style	1,569	Narrative
P8	Cont, Org, WC, SF, Conv, Voice	723	Narrative

Table 1: ASAP/ASAP++ combined dataset statistics. *Pr* denotes the prompt number. WC: *Word Choice*; PA: *Prompt Adherence*; Nar: *Narrativity*; Org: *Organization*; SF: *Sentence Fluency*; Conv: *Conventions*; Lang: *Language*.

In the cross-prompt setting, each target prompt is used for testing, while the other seven prompts are used for training. For instance, when the target prompt is 8, only 1–7 prompts are used in training and testing with P8. We used the 2080ti GPU, batch 10, epoch 50, selecting the model with the best validation. We use the efficient GEC model proposed by Rothe et al. (2021), which is pre-trained on sentence-level corrupted mC4 corpus and fine-tuned on cLang-8[5] dataset. As the official code is absent, we used an open-source implementation of the model[6], fine-tuned on English benchmark data on T5 (Raffel et al., 2020), achieving F0.5 scores of 65.01 on CoNLL-2014-test (Ng et al., 2014) and 70.32 on BEA-19 test set (Bryant et al., 2019).

5 Results and Discussions

Comparison with single encoder As our goal is to validate the impact of the proposed grammar-aware approach, we primarily compare GAPS against the *Single Encoder* model, which processes only the original essay without incorporating grammatical error-corrected versions yet within our designed structure. Trait-wise results in Table 2 highlight that referring to corrected essays with GAPS remarkably enhances the scoring performance for all traits except for *Overall*. Notably, the improvement is more pronounced in syntactic and lexical-related traits; nevertheless, the observed QWK enhancements in contextual assessment traits suggest that our method also facilitates the capture of semantic aspects. Prompt-wise results in Table 3 demonstrate that GAPS consistently outperforms the single-encoder model, confirming the efficacy of referring to error correction information.

Generalizability across prompts Grammar serves as a universal, prompt-agnostic criterion for evaluation, largely unaffected by the specific in-

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

[4] https://lwsam.github.io/ASAP++/lrec2018.html

[5] https://github.com/google-research-datasets/clang8

[6] https://github.com/gotutiyan/gec-t5

Model	**Traits**									**AVG**	**SD(↓)**
	Overall	Content	Org	WC	SF	Conv	PA	Lang	Nar		-
Hi att	0.453	0.348	0.243	0.416	0.428	0.244	0.309	0.293	0.379	0.346	-
AES aug	0.402	0.342	0.256	0.402	0.432	0.239	0.331	0.313	0.377	0.344	-
PAES (Ridley et al., 2020)	0.657	0.539	0.414	0.531	0.536	0.357	0.570	0.531	0.605	0.527	-
CTS (Ridley et al., 2021)	0.670	0.555	0.458	0.557	0.545	0.412	0.565	0.536	0.608	0.545	-
PMAES (Chen and Li, 2023)	0.671	0.567	0.481	0.584	0.582	0.421	0.584	0.545	0.614	0.561	-
PLAES (Chen and Li, 2024)	0.673	0.574	0.491	0.579	0.580	0.447	0.601	0.554	0.631	0.570	-
ProTACT **[TA+ PA]** (Do et al., 2023)	**0.674**	0.596	**0.518**	**0.599**	0.585	0.450	0.619	0.596	0.639	0.586	±0.009
Single Encoder	0.673	0.567	0.480	0.578	0.573	0.437	0.571	0.548	0.612	0.560	±0.012
GAPS [GA]	0.672	0.573	0.485	0.580	0.586	0.451	0.582	0.567	0.630	0.570	±0.014
GAPS [TA+ GA]	0.669	0.595	0.514	0.585	0.579	0.465	0.615	0.603	0.648	0.586	±0.017
GAPS [TA+ PA + GA]	0.670	**0.597**	0.515	0.595	**0.590**	**0.472**	**0.621**	**0.608**	**0.650**	**0.591**	±0.011

Table 2: Five runs averaged QWK scores over all prompts for each **trait**. TA and PA denote the prompt-aware and trait-aware methods in ProTACT (Do et al., 2023), respectively, while GA represents our grammar-aware approach.

Model	**Prompts**								**AVG**	**SD(↓)**
	1	2	3	4	5	6	7	8		
Hi att	0.315	0.478	0.317	0.478	0.375	0.357	0.205	0.265	0.349	-
AES aug	0.330	0.518	0.299	0.477	0.341	0.399	0.162	0.200	0.341	-
PAES (Ridley et al., 2020)	0.605	0.522	0.575	0.606	0.634	0.545	0.356	0.447	0.536	-
CTS (Ridley et al., 2021)	0.623	0.540	0.592	0.623	0.613	0.548	0.384	0.504	0.553	-
PMAES (Chen and Li, 2023)	0.656	0.553	0.598	0.606	0.626	0.572	0.386	0.530	0.566	-
PLAES (Chen and Li, 2024)	0.648	0.563	0.604	0.623	0.634	**0.593**	0.403	0.533	0.575	-
ProTACT **[TA+ PA]**	0.647	0.587	0.623	0.632	**0.674**	0.584	0.446	**0.541**	0.592	±0.016
Single Encoder	0.633	0.562	0.595	0.620	0.616	0.562	0.406	0.534	0.566	±0.016
GAPS [GA]	0.631	0.587	0.610	0.637	0.614	0.580	0.421	0.520	0.575	±0.016
GAPS [TA+ GA]	0.627	**0.626**	0.633	0.640	0.660	0.591	**0.469**	0.494	0.593	±0.022
GAPS [TA+ PA + GA]	**0.654**	0.614	**0.636**	**0.646**	0.665	0.590	**0.469**	0.498	**0.597**	±0.019

Table 3: Five runs averaged QWK scores over all traits for each **prompt**; *SD* is the averaged standard deviation for five seeds, and **bold** text indicates the highest value.

structions within a prompt; thus, it can be a great indicator for prompt generalization. This is particularly evident in the *Convention* trait, which evaluates writing conventions such as spelling and punctuation, independent of prompt-relevant information (Mathias and Bhattacharyya, 2018). While even robust previous models, such as PMAES (Chen and Li, 2023), PLAES (Chen and Li, 2024), and ProTACT (Do et al., 2023), have shown significantly lower performance in this trait, GAPS demonstrates substantial improvements in the *Convention*. These results emphasize the robustness of our method's prompt generalization capabilities. Furthermore, in a prompt-wise examination, we observe substantial performance gains for the challenging Prompt 7, which presents a difficult cross-prompt setting due to its differences in type and evaluated trait compositions (Table 1). Although Prompt 8 shares the same type, it is constrained by a smaller dataset of only 723 samples. Thus, the notable improvements in Prompt 7 indicate GAPS' ability to effectively evaluate essays of new, unseen prompts, even in more challenging settings.

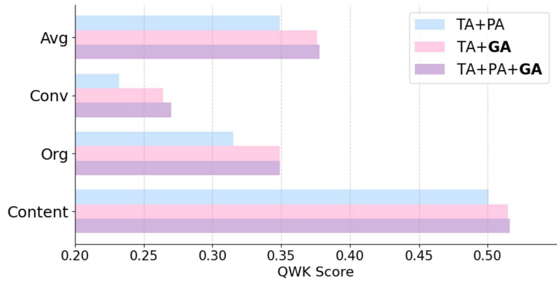


Figure 2: QWK scores for traits evaluated in Prompt 7.

Impact of grammar-aware vs. prompt-aware approaches We directly compare GAPS, our grammar-aware (GA) approach, with ProTACT's prompt-aware (PA) method, which leverages prompt information directly. Since ProTACT also introduces trait-relation-aware (TA) methods, such as trait-similarity loss, we incorporate TA into our model for a fair comparison (i.e., TA+GA vs. TA+PA). Results in Table 2 show that PA excels in *Organization*, *Word Choice*, and *Sentence Fluency*, indicating its strength in capturing logical flow and prompt adherence. In contrast, GA out-

Model	**Traits**									**AVG**	**SD(↓)**
	Overall	Content	Org	WC	SF	Conv	PA	Lang	Nar		
GAPS	**0.672**	**0.573**	0.485	**0.580**	**0.586**	**0.451**	**0.582**	0.567	0.630	**0.570**	±0.014
w/o KS	0.672	0.570	**0.488**	0.561	0.564	0.446	0.570	**0.571**	**0.632**	0.564	±0.015
w/o GCT	0.667	0.559	0.467	**0.580**	0.569	0.420	0.571	0.546	0.611	0.554	±0.011

Table 4: Five runs averaged ablation QWK results over all prompts for each **trait**. KS and GCT denote the *Knowledge Sharing* and *Grammar Correction Tagging*, respectively.

performs PA in *Conventions*, *Language*, and *Narrativity*, demonstrating its superiority in enhancing grammatical correctness and structural coherence. Notably, GA's impact on *Conventions* emphasizes the direct benefits of referring to grammatically corrected contexts. For most traits, using GA with PA yields better performance.

We also investigate the effects of GA on the traits evaluated in Prompt 7 (Figure 2). Interestingly, in this low-resource cross-prompt scenario, where similar types are scarce, GA outperforms PA in all traits. This result suggests that incorporating grammar-revised essays is much more beneficial than relying on prompt information alone, especially in challenging cross-prompt settings.

Effects of knowledge sharing We further examine the impact of the designed knowledge-sharing layer by comparing GAPS with a version that omits the knowledge-sharing component (Table 4; w/o KS). Instead of using Equations 3 and 4, we simply concatenate the obtained E_o and E_g vectors and subsequently input them to the LSTM layer. Removing the KS module results in a marked decline in the *Word Choice*, *Sentence Fluency*, and *Convention* traits, underscoring the pivotal role of knowledge sharing in effectively capturing both structural and syntactic features. Without the KS module, the model struggles to integrate the original and grammar-corrected essay representations, which hinders its ability to make accurate judgments of these traits.

Effects of grammar correction tagging To investigate whether the inclusion of grammar correction tags is effective, we conducted an ablation study to eliminate the tags, utilizing only the pure corrected essay (Table 4). Notably, specifying the correction tags in the essay significantly improves scoring performance across most traits, revealing the importance of key entity identification for balanced generalization. These findings are consistent with existing studies, which show that underscoring the key entities improves performance on downstream tasks (Ryu et al., 2024).

6 Conclusion

We propose a grammar-aware cross-prompt trait scoring to enhance prompt generalizability. By directly utilizing grammar error-corrected essays as the input, the model can learn more syntactic-aware representations of essays. In addition, we introduce tagging the corrected tokens, which leads the model to better focus on critical parts for grading. Our experiments demonstrate that grammar-aware essay representation obtained with our straightforward model structure remarkably assists the scoring of lexical or grammatical traits. Further, the notable performance increase in the most challenging prompt implies our model's internal acquisition of prompt-independent features.

7 Limitations

We have explored the use of grammar error correction to assist in obtaining invariant essay representation for cross-prompt trait scoring. Our limitation relates to the possible dependency on the GEC performance, which is not handled in this work. Although we used the robust and effective GEC method, further experiments with different models will provide another room for scoring quality improvement on cross-prompt settings.

8 Ethical Statement

We used publicly available datasets in this work.

Acknowledgments

This research was partially supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-2020-0-01789) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) (50%). It was also supported by an IITP grant funded by the Korea government (MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH)) (50%).

References

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Yuan Chen and Xia Li. 2023. 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)*, pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.

Yuan Chen and Xia Li. 2024. PLAES: Prompt-generalized and level-aware learning framework for cross-prompt automated essay scoring. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12775–12786, Torino, Italia. ELRA and ICCL.

Hannah Craighead, Andrew Caines, Paula Buttery, and Helen Yannakoudakis. 2020. Investigating the effect of auxiliary objectives for the automated grading of learner English speech transcriptions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2258–2269, Online. Association for Computational Linguistics.

Ronan Cummins and Marek Rei. 2018. Neural multi-task learning in automated assessment. *ArXiv*, abs/1801.06830.

Yuning Ding, Marie Bexte, and Andrea Horbach. 2023. Score it all together: A multi-task learning study on automatic scoring of argumentative essays. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13052–13063, Toronto, Canada. Association for Computational Linguistics.

Heejin Do, Yunsu Kim, and Gary Lee. 2024a. Autoregressive score generation for multi-trait essay scoring. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1659–1666, St. Julian's, Malta. Association for Computational Linguistics.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Heejin Do, Sangwon Ryu, and Gary Lee. 2024b. Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16427–16438, Miami, Florida, USA. Association for Computational Linguistics.

Kosuke Doi, Katsuhito Sudoh, and Satoshi Nakamura. 2024. Automated essay scoring using grammatical variety and errors with multi-task learning and item response theory. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 316–329, Mexico City, Mexico. Association for Computational Linguistics.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring-an empirical study. In *EMNLP*, volume 435, pages 1072–1077.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *CoNLL*, pages 153–162.

John A. Hawkins and Luna Filipović. 2012. Criterial features in l2 english: Specifying the reference levels of the common european framework.

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

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. Prompt-BERT: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. 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)*, pages 1088–1097.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7661–7681, Bangkok, Thailand. Association for Computational Linguistics.

Xia Li, Minping Chen, and Jian-Yun Nie. 2020. Sednn: shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowledge-Based Systems*, 210:106491.

Jiawei Liu, Yang Xu, and Lingzhe Zhao. 2019. Automated essay scoring based on two-stage learning.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Panitan Muangkammuen and Fumiyo Fukumoto. 2020. Multi-task learning for automated essay scoring with sentiment analysis. 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: Student Research Workshop*, pages 116–123, Suzhou, China. Association for Computational Linguistics.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. 2014. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. *arXiv preprint arXiv:2008.01441*.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online. Association for Computational Linguistics.

Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Geunbae Lee, and Jungseul Ok. 2024. Key-element-informed sllm tuning for document summarization. In *Interspeech 2024*, page 1940–1944.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yongjie Wang, Chuan Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. *arXiv preprint arXiv:2205.03835*.