Experimental Comparison of Automated Feedback Generation for Academic Writing: From Symbolic to Generative AI

Toru Sasaki, Rianne Conijn, Martijn C. Willemsen
Eindhoven University of Technology (TU/e)

Keywords: automated essay scoring (AES), automated feedback generation (AFG), generative AI, large language models (LLMs)

Introduction

Machine learning-based models for automated essay scoring (AES) and feedback generation (AFG) have been developed since 1960's, some of which have been commercially deployed. Such systems are expected to lighten the labor-intensive task of essay scoring by teachers and help them provide timely feedback to students [4,7]. However, despite the long history and active discussion, these tools are not used effectively enough in practice [3]. Latest attempts of AFG use generative large language models (LLMs) as well as conventional rule-based symbolic AI for writing evaluation [9,10]. We try to identify how accurately currently available models can spot micro-level errors such as in spelling or verb form and whether they can locate improvable segments of text at macro-level.

Past studies have examined the performance of existing tools either focusing on one single tool or comparing two or more. However, to our best knowledge, none of them includes generative LLM-based model in the comparison nor distinguishes improvable segment detection from error correction. To fill these gaps, we conduct an experimental comparison of three models developed during different stages of AI advancement, from symbolic to generative eras, with a clear distinction of micro- and macro-level feedback. Findings indicate that satisfactory feedback is available only for micro-level features and no model can effectively locate improvable segments of text for the next action by the users.

Method

The first-generation AES models are chiefly based on rule-based symbolic AI that are trained for prescribed topics [2]. After the introduction of ChatGPT in 2022, researchers began to use LLMs with a non-deterministic algorithm for AFG in natural language [9,10]. We pick two most widely used writing assistants, namely *Criterion*® by Educational Testing Service (ETS) and *Grammarly* by Grammarly, Inc. among commercially deployed services [1,5]. Since no LLM-based AFG system is currently available commercially, we use GPT-4o by OpenAI with prompting strategies employed in previous studies [8].

For the comparison, 20 essays are randomly chosen from an academic argumentative essay writing dataset, written by university bachelor students on a general topic of "multitasking". The chosen essays were then annotated by three experienced annotators in terms of (i) "minimum necessary correction" to make the text grammatically acceptable and (ii) "improvable segments" for better argumentation. The 20 essays are fed to *Criterion*® and *Grammarly* through their commercial user interface. As suggested in a past study, GPT-4o is given the same instruction in the prompt as provided to the annotators [6].

Results

Experiment results indicate that *Criterion*® provides both micro- and macro-level correction and suggestions, whereas *Grammarly* gives mostly micro-level feedback only. GPT-4o is capable of providing the both, yet is sensitive to prompts and its output quality is not stable across essays. Detailed analyses are still ongoing to calculate their feedback accuracy against the annotated gold standard, but some initial findings show that *Criterion*® fails to spot some grammatical errors that *Grammarly* does spot in terms of accuracy, and that its macro-level suggestions are mostly too generic to help users know what to do next for text revision. In contrast, *Grammarly* spots spelling/grammar errors at a higher accuracy, but does not provide macro-level suggestions.

Conclusion and discussion

Even though LLM-based technologies are increasingly available in some fields, AES/AFG tools are not used effectively enough in practice. To spot the reasons, we compare the two most widely used commercial services, namely *Criterion*® and *Grammarly*, and GPT-4o-based model with the same set of 20 student-written essays. The results indicate that no model provides satisfactory feedback both in spotting "minimum necessary correction" and locating "improvable segments" with acceptable stability in output accuracy and quality. Better non-deterministic finetuning of LLMs and application of model grounding techniques with external knowledge from earlier deterministic systems would be useful for the next step.

Additionally, one common observation is the models' inability to adjust their feedback to the skill level of the writer, which teachers are considered to be good at. This observation goes parallel with some past studies claiming that the lack of skill-adapted feedback is one of the reasons why AES/AFG tools are not used effectively enough in practice [11]. Guiding the model to focus on skill-dependent features would be a fruitful area of future studies.

Acknowledgement

This research project is funded by Netherlands Organization for Scientific Research (NWO).

References

[1] Attali, Y. (2004). Exploring the feedback and revision features of Criterion. *Journal of Second Language Writing*, *14*(3), 191-205.

[2] Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3).

[3] Choi, S., Jang, Y., & Kim, H. (2023). Influence of pedagogical beliefs and perceived trust on teachers' acceptance of educational artificial intelligence tools. *International Journal of Human–Computer Interaction*, *39*(4), 910-922.

[4] Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback?. *Assessing writing*, *22*, 1-17.

[5] Fitria, T. N. (2021). Grammarly as AI-powered English writing assistant: Students' alternative for writing English. *Metathesis: Journal of English Language, Literature, and Teaching*, *5*(1), 65-78.

[6] Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies*, *29*(7), 8435-8463.

[7] Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, *5*, e208.

[8] Jacobsen, L. J., & Weber, K. E. (2023). The promises and pitfalls of ChatGPT as a feedback provider in higher education: An exploratory study of prompt engineering and the quality of AI-driven feedback.

[9] Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 100050.

[10] Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)* (pp. 394-403).

[11] Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., ... & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*, 101894.