Enhancing Consistency in Academic English Writing Feedback Generation with a MacBERT-large Model Combining Adversarial Training and Contrastive Learning

Jun Chen
College of Humanities, Gongqing College of Nanchang University, Gongqingcheng 332020, Jiangxi, China
Corresponding author: Jun Chen (e-mail: dream_0216@163.com)

ABSTRACT Academic English writing feedback generation requires robust semantic understanding, stable feedback output, and accurate discrimination among similar error types. Existing feedback generation systems often produce inconsistent suggestions for semantically equivalent inputs and show limited generalization to complex academic expressions. To improve feedback consistency, this study proposes an enhanced MacBERT-large encoder–decoder model integrating Fast Gradient Method adversarial training and supervised contrastive learning. The MacBERT-large encoder extracts contextual semantic representations of academic text, while a Transformer decoder generates feedback sequences using a dedicated academic vocabulary. FGM adversarial training introduces controlled perturbations into the embedding layer, enabling the model to maintain stable predictions under paraphrased or slightly varied inputs. A supervised contrastive learning module maps text samples into a representation space where feedback cases with the same error type are pulled closer and different error types are separated through NT-Xent loss. A multi-task learning framework jointly optimizes cross-entropy loss, adversarial loss, and contrastive loss to balance generation quality, robustness, and category discrimination. Experiments on the AEW-Feedback dataset containing 15,000 academic papers show that the proposed model achieves 67.3% BLEU-4, 71.2% ROUGE-L, 74.8% METEOR, and a feedback consistency score of 0.891, outperforming MacBERT-large and single-enhancement variants. The method provides a semantic consistency modeling framework for intelligent text generation and academic writing assistance systems.

INDEX TERMS macbert-large, adversarial training, contrastive learning, semantic consistency, feedback generation

I. INTRODUCTION

WRITING academic English is a core competency among researchers and it is not uncommon that non-English speakers have difficulties in writing their papers because of grammatical mistakes, non-standard terms and use of the wrong terminology. Online feedback systems which are automated can also give feedback to the learners in the form of suggestions on what to revise and these are far cheaper than the manual correction. But the current systems have severe inconsistencies in producing feedback: given even slightly different texts which contain the same type of error, the model tends to suggest contradictory revision advice; the model is incapable of distinguishing between similar issues like grammatical errors and vocabulary misuse; and the generalization performance of the model decreases significantly when exposed to language phenomena that are not represented in the training set [1-2]. These issues result in a crisis of confidence in the system on the part of the learners, which significantly limits the value of automated feedback technology in teaching academic writing.

To overcome the challenges mentioned above, this paper suggests a better MacBERT-large model that incorporates adversarial training as well as contrastive learning. The model is able to finely balance the judgment through semantic equivalence transformation over the embedding layer by introducing FGM adversarial perturbation; the supervised contrastive learning module can optimize the representation space by using the NT-Xent loss function to create clean semantic boundaries between the various error types; and the multi-task learning framework is able to jointly optimize the three tasks of generation quality, robustness and discrim-

inability. The key contributions of the paper are: developing a consistency enhancement mechanism, which integrates adversarial training with contrastive learning; developing a multi-task joint optimization strategy to trade-off between multiple training goals; and demonstrating the substantial gains on the feedback quality, consistency performance, and ability to detect errors on real academic writing datasets [3-4].

II. RELATED WORK

Language models with pre-trained language models have been found to have strong semantic understanding abilities in academic writing feedback tasks. The BERT family of models is a series of models that encode contextual information with a bidirectional encoding mechanism that attempts to present effective features to text correction and generation of feedback. MacBERT is Chinese and cross-lingual-friendly, with full-word masking strategy and N-gram masking mechanism, and has enhanced recognition of word boundaries in academic text processing [5-6]. While originally optimized for Chinese, its masking strategy is particularly effective for English academic writing, as it forces the model to learn complete technical terms and formulaic expressions rather than individual subwords. Furthermore, its performance in this task surpasses traditional English-specific models like RoBERTa in capturing the complex collocations prevalent in scholarly discourse. The encoder-decoder model is popular in text generation problems, wherein the encoder identifies semantic representation of the input text, and the decoder produces target sequences word-by-word in an autoregressive manner. Nevertheless, the classic pre-trained models have revealed evident weaknesses in writing feedback generation: the model is overly sensitive to fine variations in the input text, producing varying feedback responses to semantically equivalent input; the model does not learn the representation of error types, so it is hard to distinguish between similar problems, like grammatical errors and poor vocabulary choice; and the feedback generation process does not incorporate an explicit consistency check mechanism [7-8]. Contrasting learning and adversarial training provide new options to enhance the performance of the models. Adversarial training compels the model to learn more robust decision boundaries by building adversarial examples. The FGM technique introduces perturbations to the gradient direction of the embedding space, and is appropriately robust, and requires comparatively low computational effort. This method can be successfully used to increase the resistance of the model to input noise on text classification and named entity recognition problems. Contrastive learning learns to maximize the distance between similar samples and distance between dissimilar samples. Supervised contrastive learning makes use of label information to create positive and negative pair of samples and the NT-Xent loss function regulates the steepness of the distribution with a temperature coefficient. The approach has brought vast results in image recognition and calculation of semantic similarity, yet it is in the exploration phase when it comes to its use in text generation. The existing literature primarily focuses on the use of adversarial training or contrastive learning in isolation on a particular task, without providing a systematic solution to the issue of consistency in writing feedback, and the interaction between the two methods is not entirely researched [9-10].

III. METHODS

A. ENCODER-DECODER ARCHITECTURE DESIGN BASED ON MACBERT-LARGE

The encoder-decoder architecture formed in this paper is based on MacBERT-large as the backbone network of the encoder. This model has a 24-layer Transformer architecture that has a hidden layer size of 1024 and 16 attention heads. Academic text fragments sent to the encoder are fed back, and it is better to capture the terminology boundaries and collocation relationship in academic writing using the pre-trained weights produced using a whole-word masking mechanism. Once the input text has been tokenized using the WordPiece algorithm it is combined with the position embedding and fragment embedding to make a 512-dimensional input vector, which is then processed via multi-layer bidirectional self-attention computation to produce a context-aware semantic representation.

The decoder employs a 6-layer unidirectional Transformer architecture. This depth was selected to balance computational efficiency with generative capacity, following empirical findings that additional layers yielded diminishing returns for feedback generation tasks. To resolve the dimensionality difference between the MacBERT-large encoder (1024 dimensions) and the decoder layers, a linear projection layer is implemented to map the encoder's hidden states into the cross-attention space, ensuring seamless feature integration. The process of masking self-attention is added during decoding to avoid information leakage; every decoding layer has a self-attention sublayer, encoder-decoder attention sublayer, and feedforward network sublayer. In order to fit the generation properties of academic writing feedback, the decoder output layer applies a special vocabulary of 8000 academic words, which includes grammatical words, proposed revision words and vocabulary related to the subject matter. The decoder produces feedback word-by-word text through a greedy search or beam search algorithm, where each prediction is made based on the joint probability distribution of the encoder semantic information, and the previously generated content.

B. EMBEDDING LAYER PERTURBATION MECHANISM IN FGM ADVERSARIAL TRAINING

FGM adversarial training takes on the form of simulating subtle manipulations of the input text through the injection of specially designed perturbation vectors into the embedding layer, which induces the model to acquire a more robust feedback generation policy. Within training, once the model has done a forward and a backward propagation on the

original academic text sample, it computes the gradient of the embedding layer and retrieves the direction information. The perturbation vector is created in the direction of the gradient ascent and its magnitude is determined by the value of epsilon which is taken to be 0.5 in the present paper to balance the degree of perturbation and stability of training. In this study, epsilon is set to 0.1. This value was selected based on internal ablation tests to ensure that the perturbation is sufficient to improve robustness without distorting the underlying semantic integrity of the embedding space, which can occur at higher values (e.g., > 0.5).

The produced perturbation vectors are simply added to the word embedding representations to create the input representation of adversarial examples. Such adversarial examples keep the sequence of discrete tokens that make up the text fixed, but only create offsets in the underlying continuous embedding space, which simulates semantically equivalent transformations like synonym substitution and expression changes. The adversarial examples are re-used to compute the forward propagation again, and the feedback generation loss is calculated, and is used in the parameter updates alongside the original sample loss. Adversarial loss is introduced in such a way that the model is able to uphold the consistency of prediction in the vicinity of the embedding space. Given academic texts that have a minor variation in wording, but the error type is the same, the model can provide semantically consistent modification proposals.

C. SUPERVISED CONTRASTIVE LEARNING MODULE AND NT-XENT LOSS FUNCTION

The contrastive learning layer is an encoder-based module that learns to extract sentence-level representation vectors of each text sample on the top encoder layer and converts the hidden states of all tokens on the encoder into a fixed length 512-dimensional representation through average pooling. The module builds up positive and negative samples pairs based on the error type labels in the dataset; text pairs of the same error category are treated as positive, and text pairs of different categories as negative. In training, the anchor samples are randomly chosen among samples in each batch and the cosine similarity of the anchor samples with all other samples in the same batch is subsequently computed to create a relation matrix between samples.

The NT-Xent loss function "warps" the representation space by putting the representations of the grammatical error cluster into specific corners of the semantic space and pushing representations of other error types like lexical inappropriateness or logical inconsistencies to the other corners. The loss calculation adds a temperature coefficient of 0.07; This specific temperature was validated through a hyperparameter grid search, proving optimal for concentrating representation clusters in high-dimensional text space. For samples containing multiple error types, the primary error (the first identified category) is used for label assignment to maintain clear semantic boundaries during training; the smaller the temperature, the more sensitive the model is to small differences in representations. The different loss functions get a sample of anchors, which are then maximized over their similarity to the total of positive samples and minimized over their similarity to negative samples. Through this mechanism, maximizing intra-class similarity while minimizing inter-class correlation establishes distinct semantic boundaries in the representation space, which makes it easy to differentiate between two types of grammatical errors that the model may have difficulty distinguishing, like subject-verb disagreement and tense misuse.

D. JOINT OPTIMIZATION STRATEGY UNDER MULTI-TASK LEARNING FRAMEWORK

The multi-task learning model aggregates cross entropy loss on feedback, adversarial training loss and contrastive learning loss into one optimization objective. The cross-entropy loss is the loss of cross entropy loss on feedback, which is computed as negative log-likelihood of predicted word distribution at time is the target word; adversarial loss is to compare the prediction bias of model on perturbed samples, and restrict the model behavior by minimizing KL divergence between output distribution of clean and adversarial samples; Contrastive loss is a representation learning objective that optimizes the encoder to get tight semantic space clusters of texts of similar error type.

The weight of cross-entropy loss is 1.0, the weight of adversarial loss is 0.3 and the weight of contrastive loss is 0.2. The weight of cross-entropy loss is the largest because we aim to guarantee the accuracy of feedback generation; the weight of adversarial loss is moderate to balance the improvement of robustness and the stability of training; the weight of contrastive loss is low to avoid over-weighting on class distinction and degrading quality of generation. In addition, we use AdamW optimizer and train with learning rate of 2×10^{-5}. We use a linear warm-up schedule. Specifically, the learning rate grows linearly to its maximum up to first 1000 steps, and then decays cosine annealing to 0. There is gradient clipping threshold of 1.0 to avoid gradient explosion. We clip the gradient of an update after the four batches of gradient accumulation to simulate a larger effective batch size.

IV. RESULTS AND DISCUSSION

A. PERFORMANCE COMPARISON ANALYSIS ON THE AEW-FEEDBACK DATASET

This paper evaluates model performance on 3000 test samples of the AEW-Feedback dataset. The control groups include the baseline models BERT-base, BERT-large, MacBERT-base, and MacBERT-large, as well as the MacBERT-large+FGM variant model using only adversarial training and the MacBERT-large+CL variant model using only contrastive learning. To ensure a fair evaluation, all baseline models were implemented using the same encoder-

decoder framework described in Section 3.1. Evaluation metrics include BLEU-4, ROUGE-L, METEOR, and feedback consistency score. The results are shown in Table 1.

TABLE 1. Performance comparison of different models on the AEW-Feedback dataset

Model	BLEU-4 (%)	ROUGE-L (%)	METEOR (%)	Consistency score
BERT-base	52.1	58.3	61.2	0.743
BERT-large	55.8	61.7	64.5	0.776
MacBERT-base	56.4	62.9	65.8	0.792
MacBERT-large	58.6	64.2	67.1	0.804
MacBERT-large+FGM	62.9	67.5	70.3	0.847
MacBERT-large+CL	61.7	66.8	69.6	0.838
This artcle's model	**67.3**	**71.2**	**74.8**	**0.891**

Our model achieves 67.3% on the BLEU-4 metric, an improvement of 8.7 percentage points over the MacBERT-large baseline, with a consistency score of 0.891. Compared to variant models using adversarial training or contrastive learning alone, BLEU-4 scores are improved by 4.4 and 5.6 percentage points, respectively, validating the synergistic effect of the two techniques. Adversarial training enhances stability under input perturbations, contrastive learning strengthens error type discrimination, and joint optimization enables the model to both cope with variations in representation and accurately identify error categories, thereby significantly improving feedback generation quality and consistency performance.

B. EVALUATION OF FEEDBACK CONSISTENCY AND ERROR TYPE IDENTIFICATION CAPABILITIES

In this paper, a set of consistency tests is built to measure the stability of feedbacks of the model on semantically equivalent texts. The test set includes 500 sample pairs, each of which has academic texts, but with various expressions, yet with the same type of error. To obtain the consistency score, we compute the BLEU similarity of the feedback generated from sample pairs to measure linguistic stability. To further ensure that this consistency reflects pedagogical correctness rather than the repetition of erroneous patterns, a manual verification was conducted on a subset of consistent outputs. The results confirm that the proposed model provides both semantically stable and pedagogically accurate revision advice across paraphrased inputs. The ability to identify the type of error is tested on 1200 samples that bear four types of errors: grammatical errors, inappropriate vocabulary, logical inconsistency and formatting errors. It is measured using four metrics: accuracy, precision, recall and F1 score. The performance of our model in fine-grained error classification is compared with that of the baseline model. The data on the feedback consistency, and error identification accuracy are in tables 2 and 3, respectively.

TABLE 2. Performance Evaluation of Feedback Consistency and Error Type Identification

Model	Consistency score	Overall accuracy (%)	Accuracy (%)	Recall rate (%)	F1 score (%)
MacBERT-large	0.804	83.8	82.5	81.9	82.2
MacBERT-large+FGM	0.847	87.6	86.8	86.2	86.5
MacBERT-large+CL	0.838	89.3	88.7	88.1	88.4
This artcle's model	**0.891**	**94.2**	**93.8**	**93.5**	**93.6**

TABLE 3. Accuracy of different error types

Error Type	MacBERT-large (%)	MacBERT-large+FGM (%)	MacBERT-large+CL (%)	This paper's model (%)
Syntax error	86.3	89.7	92.4	**96.1**
Inappropriate vocabulary	82.1	85.9	88.6	**93.8**
Logical confusion	81.5	86.2	87.1	**92.5**
Formatting issues	85.7	88.8	89.2	**94.3**

Our model achieved a consistency score of 0.891, a 10.8% improvement over the baseline, with an error type identification accuracy of 94.2%. Contrastive learning improved the grammatical error identification rate to 96.1%, while adversarial training enabled the model to maintain stable output on paraphrasing. Models using FGM or CL alone achieved consistency scores of 0.847 and 0.838, respectively, with accuracies of 87.6% and 89.3%, both lower than our proposed model. Joint optimization enabled the model to adapt to input variations while accurately distinguishing error categories, achieving over 92% accuracy in identifying all four error types, validating the effectiveness of our approach.

V. DISCUSSION

Adversarial training increases the resilience of the model to variation in inputs by perturbing the embedding layer, such that the model can have consistent judgment criteria in the presence of semantically equivalent transformations like paraphrasing and variations in sentence structure. The ability to learn the contrasting types of errors is enhanced by contrastive learning to optimize the category separation of the representation space, preventing errors of grammar and related problems like inappropriate vocabulary. Combining these two methods produces a synergetic effect: The horizontal consistency of adversarial training and the vertical discriminative power of contrastive learning are complementary to each other. The multi-task learning model strikes the trade-off between the three maximization goals of feedback generation accuracy, robustness and category discriminative power by assigning reasonable loss weight allowing the model to greatly enhance consistency performance and error recognition accuracy without compromising the quality of feedback generation.

VI. CONCLUSION

To solve the problems of lack of consistency and error type confusion in academic English writing feedback generation, this paper proposes an improved MacBERT-large model based on adversarial training and contrastive learning. The FGM adversarial training forced the perturbation gradients to appear in the embedding layer of the model, which improved the robustness of the model to input data. The

supervised contrastive learning module optimized the representation space of the model based on the NT-Xent loss function and made the semantic space of the error type more distinct. The multi-task learning framework optimized the three types of loss functions and made the feedback accuracy, robustness, and discriminative loss optimal together. The experimental results verified the improvement in feedback accuracy, robustness, and error recognition ability. The current methods still depend on large amounts of labeled data and high computational costs. In the future, we will study the weakly supervised learning method to reduce the amount of labeled data, introduce the lightweight model structure to improve the inference efficiency of the model, and extend the model to the multilingual academic writing scene to improve the cross-language transferability of the model.

AUTHOR CONTRIBUTIONS

Jun Chen designed, collected and analyzed the data, and drafted the manuscript. Jun Chen conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Jun Chen participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] N. Yurko, I. Styfanyshyn, U. Protsenko, et al., "The peculiarities of academic English writing," Paradigmatic view on the concept of world science - volume 2, 2020, doi: 10.36074/21.08.2020.v2.37.

[2] V. Makarova, Z. Li, and Z. Wang, "Can ChatGPT grade non-native academic English writing? Advances in Educational Technologies and Instructional Design," 2024:97-116, doi: 10.4018/979-8-3693-1054-0.ch005.

[3] J. Chen, H. Shen, and S. Hu, "A survey study of a catechism course in academic English writing," Education Quarterly Reviews, pp. 6(3), 2023, doi: 10.31014/aior.1993.06.03.771.

[4] J. Ma, "A study of the impact of peer review on academic English writing among English majors," SHS Web of Conferences, vol. 168, pp. 5, 2023, doi: 10.1051/shsconf/202316801005.

[5] H. Li, "Blended assessment in university academic English writing course," Journal of Education and Educational Research, vol. 12, no. 1, pp. 161-164, 2025, doi: 10.54097/2wkab381.

[6] D. Perrodin D, "Assessment by Thai academic English writing teachers of the flow of given to new topic information within academic writing," Journal on English as a Foreign Language, 2021, doi: 10.23971/jefl.v11i2.2684.

[7] F. Rahman, "Fostering academic English writing through blended approach: a study of Bengali learners of English," The Dhaka University Journal of Linguistics, Art. no. 16(31_32), 2024, doi: 10.70438/dujl/163132/0010.

[8] C. Srisawat and K. Poonpon, "Revision of an academic English writing rubric for a graduate school admission test," PASAA: A Journal of Language Teaching & Learning in Thailand, pp. 65, 2023, doi: 10.58837/chula.pasaa.65.1.9.

[9] Y. Gao and Y. Lou, "A study on teacher's regulative discourse in online academic English writing course," Education Reform and Development, vol. 6, no. 7, pp. 144-150, 2024, doi: 10.26689/erd.v6i7.7737.

[10] S. Li, "Research on academic English writing teaching based on production-oriented approach," World Journal of Educational Research, vol. 11, no. 3, pp. p36, 2024, doi: 10.22158/wjer.v11n3p36.