Studies in English Language Teaching
ISSN 2372-9740 (Print) ISSN 2329-311X (Online)
Vol. 12, No. 4, 2024
www.scholink.org/ojs/index.php/selt

Original Paper

An Investigation into the Feedback of Automated Writing Evaluation System on EFL Learners' Writing

Zhuolin Tang

School of Foreign Language, Tianjin Normal University, Tianjin, China

Received: August 3, 2024 Accepted: September 12, 2024 Online Published: September 18, 2024

doi:10.22158/selt.v12n4p1 URL: http://dx.doi.org/10.22158/selt.v12n4p1

Abstract

Computer technology has developed at a rapid pace in recent decades. The combination of artificial intelligence and writing teaching has become an important development trend. For better utilization of Automated Writing Evaluation System in teaching, the quality of its feedback on various types of errors in compositions is one of the issues worth exploring. In order to investigate this issue, this study compares the diagnosis of EFL learners' compositions by Automated Writing Evaluation System with the actual errors in the compositions, and calculates the Precision and Recall of the iWrite system's feedback. Unlike previous studies, this study synthesizes macro discourse elements such as propositional logic, across-sentence logic, and discourse structure, and categorizes the actual errors in learners' compositions into seven kinds. The results show that the iWrite system performs excellent in detecting three kinds of errors, which are lexical errors, grammatical errors, and other detail-type errors, and is moderately effective in detecting pragmatic errors, while it fails to provide feedback on propositional logic, discourse cohesion and coherence, and paragraph structure, which are the high-level error types, indicating that iWrite system is completely unable to capture errors in terms of macro-discourse elements. These findings provide a strong basis for improving Automated Writing Evaluation System, and at the same time provide English teachers with a reference for writing teaching, which helps to better serve the field of foreign language education.

Keywords

Automated Writing Evaluation System, iWrite writing system, the quality of feedback, EFL learners

1. Introduction

Writing is a critical aspect of language proficiency, and as such, writing instruction has long been a central focus in language education. Over the past few decades, the rapid development of computer technology has significantly expanded the methods of foreign language instruction and enhanced its

efficiency. The integration of computer technology into English writing instruction has become an important trend. In this context, Automated Writing Evaluation (AWE) systems have emerged. Domestically developed AWE systems, such as PiGai, iWrite, Juku and Bingguo English, have been incorporated into writing instruction, gaining increasing attention from Chinese scholars regarding their functions and effectiveness.

After the introduction of AWE systems into writing instruction, questions have arisen: How accurate are the scores provided by these systems? What is the quality of the feedback they generate? These issues have attracted the attention of teachers and researchers alike. This study focuses on the iWrite system, a newly emerging AWE system in China, conducting a classification analysis of the errors targeted by the system and comparing its error diagnosis with the actual errors present in the compositions of second language learners. Existing research has not sufficiently considered issues such as propositional logic, cross-sentence logic, and discourse structure in the writing of second language learners. This study will systematically analyze these aspects to obtain more comprehensive data covering grammar, vocabulary, and discourse. The findings of this research can provide a strong foundation for the improvement of AWE systems, promoting their development toward becoming more comprehensive and accurate to better meet educational needs. Additionally, the results offer valuable insights for foreign language teachers, enabling this system to better serve foreign language instruction.

2. The Application and Related Studies of Automated Evaluation Systems

2.1 The Application of Automated Writing Evaluation Systems

The first experiment on computer-assisted essay evaluation was proposed in 1966. In the 21st century, more complex Automated Writing Evaluation (AWE) systems were applied to writing courses, such as the Criterion Online Writing Service, IntelliMetricSM essay scoring system, and the e-rater scoring engine abroad, as well as iWrite, Juku PiGai, and Bingguo English in China. AWE systems are an innovative application based on natural language processing technology, aimed at automatically evaluating student writing. Early AWE systems abroad were mostly applied to standardized writing exams and academic writing assessments, such as TOEFL and GMAT, to reduce labor costs and improve grading efficiency. However, AWE systems in China have not yet been used in large-scale exams like the Gaokao or CET-4 and CET-6 in China. Instead, they are primarily used in writing instruction, leading research in China to focus on their effectiveness and efficiency in teaching.

The iWrite writing instruction and evaluation system, involved in this study, is a commercial online program developed by the largest foreign language publisher and university press in China, Foreign Language Teaching and Research Press. The iWrite system provides real-time overall scoring and reports on four language features: language (fluency, accuracy, and complexity), content (relevance and coherence), discourse structure (organization and discourse markers), and details (spelling and punctuation). Final scores and feedback can be generated based on either automatic evaluations or teacher-modified evaluation reports. Teachers can also create their own writing tasks, with either

automatic or manual scoring. Students are required to complete assigned writing tasks, with word limits, online either during or outside class. Additionally, students can complete their writing tasks within a set time or without a time limit and submit multiple drafts as needed according to teacher requirements.

2.2 Related Studies

2.2.1 Overview of Automated Writing Evaluation Systems

These systems are capable of quickly grading essays and providing efficient and immediate feedback, thereby reducing teachers' grading burden and supporting writing instruction. Teachers can also design more targeted writing exercises based on the feedback, offering more effective writing skill guidance (Wilson & Czik, 2016). Research on AWE systems' grading functions has primarily focused on examining the reliability of the system's grading, often comparing the consistency of system-generated scores with human grading (Ramineni & Williamson, 2018; Li Yanling & Tian Xiachun, 2018). In contrast, research on written feedback provided by AWE systems is the main focus, and studies in this area can be categorized as follows:

(1) Exploring the impact of machine feedback on students' revisions during the writing process. These studies primarily examine changes in essay content, the effectiveness of revisions (good, medium, or poor), and the number of revisions made after receiving machine feedback (Wu Yong & Zhang Wenxia, 2016; Bai & Hu, 2017; Ranalli et al., 2017).

(2) Comparing the impact of machine feedback with other feedback types on English writing, such as investigating how AWE feedback and model essay feedback affect the linguistic complexity of college students' English essays (Bai Yun, Han Jiying, & Wang Junju, 2022), comparing the effect of teacher feedback versus AWE feedback on English writing (Wang Mengjun, 2020), and comparing the effectiveness of AWE feedback, teacher oral feedback, and teacher written feedback on improving argumentative essay writing skills (Zhang Ya & Zhao Yonggang, 2020).

(3) Investigating the impact of machine feedback on students' classroom English writing. Several studies abroad have shown that AWE systems can help improve writing quality (Palermo & Thomson, 2018) and enhance students' ability to revise their essays effectively (Knight et al., 2020). Additionally, these systems can increase students' writing motivation (Wilson & Czik, 2016), writing self-efficacy (Wilson & Roscoe, 2020), or change their attitudes towards writing (Roscoe et al., 2018). Studies in China primarily focus on the impact of AWE systems on university students' second-language writing in terms of vocabulary, grammar, and syntax (Shi Xiaoling, 2012; Yang Xiaoqiong & Dai Yuncai, 2015). Given that second language learners often vary in their language proficiency levels, both Chinese and international researchers have explored how AWE systems affect learners' writing abilities at different proficiency levels (Huang Shaoying & Zhang Li, 2015; Yang Ling, 2013; Jahangard, 2020).

(4) Exploring users' attitudes and perceptions of AWE systems. Research in this area is typically conducted through surveys and interviews. Several studies have indicated that machine feedback can positively contribute to writing improvement, but it cannot entirely replace teacher and peer feedback in writing instruction and should be used as a supplementary tool (Lai Y.H., 2010; Yang Ling, 2013; Li

Yihua, 2015).

(5) Investigating the precision and comprehensiveness of machine feedback. Research in this area can be divided into two types. The first compares the feedback generated by AWE systems with teacher feedback on certain aspects of essays to assess the precision of AWE systems. For instance, Dikli (2010) compared feedback from three perspectives: form (word, sentence, and paragraph levels), type (grammar, mechanics, usage, and style), and function (problem identification, clarification, elaboration, and suggestion). Shi Xiaoling (2012) found that teacher feedback was superior to machine feedback in areas like ideas and structure, with teacher feedback being clearer and more targeted. The second type of research uses quantitative data to evaluate the accuracy and comprehensiveness of machine feedback, primarily focusing on two indicators: precision and recall (Bai Lifang & Wang Jian, 2019). Some studies measured only precision (Bai & Hu, 2017; Ranalli et al., 2017), while others measured both precision and recall. However, these studies tended to focus more on lexical, grammatical, and collocational errors, without thoroughly classifying other types of errors (Han et al., 2006; Tetreault & Chodorow, 2008). Therefore, it is necessary to investigate macro-discourse elements such as propositional logic, discourse coherence, and structure in student essays. Additionally, most of these studies have focused on the Criterion software developed by the Educational Testing Service (ETS) in the U.S., which may not be well-suited for Chinese second-language learners, as it cannot detect errors unique to "Chinglish." Thus, it is important to focus research on automated evaluation systems developed in China.

2.2.2 Review of Studies on the iWrite System

iWrite is an emerging AWE system developed in China, and thus, there has been little research on it abroad. Research in China on the system, however, has been concentrated since 2017. In 2017, He Zhouchun and Gong Yanzhi conducted a case study on iWrite, summarizing its strengths and weaknesses through interviews with teachers and students, and proposing improvement strategies. In the same year, Zhang Yu (2017) explored the impact of the iWrite system on students' writing accuracy and found that it improved academic writing skills and helped teachers provide personalized feedback. He also emphasized that most students expected both automatic feedback and teacher feedback to receive more effective guidance on content and structure. Moreover, Liu Yingliang and Liu Jiaying (2018) examined the effects of error feedback, comparing students' first and final drafts before and after receiving system feedback, and concluded that automated feedback should be combined with other feedback methods to achieve better results.

Some researchers have also explored the application of iWrite in teaching. For example, Zhang Zhicai (2018) studied the advantages and disadvantages of traditional writing instruction versus automated evaluation systems, proposing innovative teaching approaches to integrate iWrite into college English writing courses. In recent years, studies on iWrite have increasingly focused on investigating how it can be combined with other teaching methods in writing instruction (e.g., Shi Lin, 2021; Liu Yingliang & Liu Shenglan, 2022).

In summary, the application of AWE systems in foreign exams and teaching in China has developed along different lines, and research on feedback quality has gradually increased. However, limited attention has been given to iWrite, a newly emerging AWE system in China. Moreover, existing research still lacks sufficient discussion on macro-discourse elements such as propositional logic, discourse coherence, and structure in student essays. This provides a starting point for further exploration in this study.

Building upon the aforementioned research, this study will further investigate how the iWrite system provides feedback on various types of errors, offering valuable insights for improving automated writing evaluation systems. This study seeks to answer the following questions:

1. What types of errors are present in the compositions of second language learners?
2. What are the characteristics of the feedback provided by the iWrite system on different types of errors found in the compositions?

3. Methodology

3.1 Participants and Materials

The participants in this study were second-year English major students from a university in Tianjin. These students were aged between 18 and 21, with the majority being female (n=31). A total of 35 students participated in four writing tasks, with the topics related to their daily lives and studies. The genre of the essays was argumentative, and the word count was no less than 120 words. The submission of the writing tasks was not limited in the number of attempts, only by a final deadline. After submitting their first drafts online, students received immediate scores, overall evaluations, and sentence-level revision suggestions from the iWrite system. Students could revise and resubmit their drafts multiple times until they were satisfied with their work.

Two researchers ranked the average scores of the students' first drafts from the four tasks as evaluated by the iWrite system, from highest to lowest. The top 10 students based on the average scores were classified as the high-proficiency group, while the bottom 10 were classified as the low-proficiency group. An independent samples t-test was conducted on the writing scores of the two groups, and the results indicated a significant difference in writing ability between the high- and low-proficiency groups ($t = 9.129$, $p < 0.05$). Since students in the low-proficiency group made a wider range and larger number of errors, their essays were more representative for error analysis. Thus, 10 students from the low-proficiency group were selected as the subjects for this study, and their 10 lowest-scoring first drafts from the four writing tasks were chosen as the materials for analysis.

3.2 Data Collection and Analysis

In this study, the types of errors made by students in their essays were classified into seven major categories: lexical errors, grammatical errors, errors in propositional logic, errors in cross-sentence logic, errors in paragraph structure, pragmatic errors, and other detail-related errors. The specific classification of these errors is shown in Table 1. Only the seven major error categories were counted,

not the subcategories.

The study aimed to count the frequency of errors in the students' essays based on the following three aspects: (1) the number of errors identified by the researchers; (2) the number of errors identified by the iWrite system; and (3) the number of errors correctly flagged by the iWrite system. If the same error appeared multiple times in one or more students' essays, it was included in the frequency count. The researchers used tools such as the Cambridge Dictionary, Collins Dictionary, and the COCA corpus to assist in identifying errors, and conducted multiple evaluations and codings of the 10 students' essays to ensure accuracy in error classification.

According to Leacock et al. (2010), precision of automatic error detection was evaluated as the number of errors correctly identified by the system divided by the total number of errors identified by the system. Recall was calculated as the number of errors that the system was able to find divided by the total number of errors identified by the researchers. Therefore, when the iWrite system generated a false positive (i.e., incorrectly flagged a grammatically correct sentence as erroneous), the precision value would decrease. Similarly, when the automated evaluation system failed to detect an error identified by the researchers, the recall value would decrease.

Table 1. Classification of Errors

Categories	Subcategories	Examples of Each Error Subcategory (corrected forms are in parentheses)
a. Lexical errors	1. Errors in the use of singular and plural forms	grade ***point*** (points) (Chen)
	2. Errors in word class usage	***excessively*** (excessive) concern (Jin)
	3. Omission or misuse of articles	have (a) better mental state. (Shen)
	4. Errors in the use of personal pronouns	***he*** *is* (they are) unnecessary (Guo)
	5. Others*	In this study, these types of errors occurred infrequently.
b. Grammatical errors	1. Errors in syntactic structure	(1) At the same time, ***in the human society has more than seventy billion people today***, exams are also the most efficient way. (Shen) (2) They tell us what we need to improve, and then we are more ***possible*** (able) to make progress. ("possible" cannot be used with a human subject) (Duan)
	2. Others* (e.g., subject-verb disagreement,tense/voice errors)	In this study, these types of errors occurred infrequently.
c. Errors in propositional	1. Errors or lack of logic in the propositional relationship of sentences	(1) Examination is a way to ***evaluate teaching*** (+effectiveness) and students' learning. (Wu) (2) At the same time, in the human society has more

logic		than seventy billion people today, exams are also ***the most efficient way*** (+of identifying talented/excellent people). (Li)
	2. Sentences with meaningless propositions	Examination is normal in our life,***every person can*** have their different ideas about the examination. (Guo)
d. Errors in cross-sentence logic	1. Cross-sentence logical inconsistency	(1) ***996 work culture is harmful*** to our family relationship and social economic development. We ***should cultivate the values of hard work***, and we should increase our efficiency rather than extending our work hours. (Duan) (2) Students may play a disorder in this examination, ***so*** they will be dejected. ***So*** students should come out of this state and be better prepared for the next round of learning.(Yang)
	2. Absence of logical connectors between paragraphs	(1) The body of the essay lacks logical connectors between two consecutive paragraphs presenting parallel ideas. (2) The conclusion is not introduced with a logical connector (e.g., “In conclusion”), but instead directly summarizes the content.(Lin)
e. Errors in paragraph structure	1. Incomplete structure of the introduction, body, and conclusion in the overall essay organization	The essay lacks a necessary conclusion. (Chen)
	2. Incomplete introduction	The introduction lacks an introductory statement and only provides a thesis statement expressing the writer’s opinion. (Wen)
	3. Unclear structure of body paragraphs	The body paragraphs only explain the thesis statement in different ways without providing evidence to support it. (Wen)
f. Pragmatic errors	1. Use of imperative mood in argumentative writing	Although the examination is very important, but it is not the only standard to measure students, ***please*** treat the examination reasonably. (Ma)
	2. Errors in phrase collocation	If we check the schedule, we can find most of football matches are played ***in*** (at) midnight due to time difference. (Shen)
	3. Chinglish errors	most of people ***have a very tense state of mind.*** (Chen)

		better play their strength (Guo)
g. Other detail-related errors	1. Spelling errors	If we want to watch every match, we have to stay up late for many times, which will harm our health ***psycially*** (physically) and ***metally***(mentally). (Shen)
	2. Errors in capitalization	All in all, ***In*** this social trend ... (Duan)
	3. Punctuation errors	In most cases, although the examination only measures people's scores, it is also because it only measures your grades, not your background**,** (.) ***f***or most people, it is a fair chance to compete with others and change their fate**,** (.) at the same time... (Wen)

Note. Each error category includes additional subcategories of errors. However, since these subcategories occurred infrequently in the sample, they are all classified under "Others*".

4. Results

Figure 1 presents the number of errors identified in the students' essays by the researchers and the iWrite system, as well as the number of errors correctly flagged by the system.

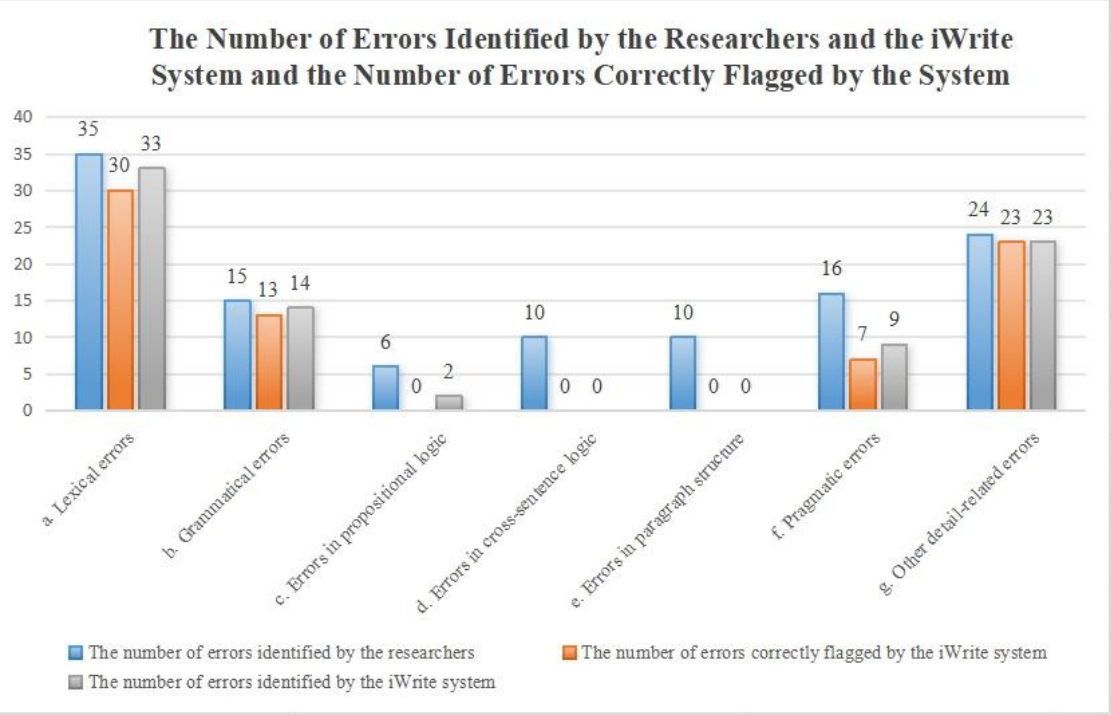

Figure 1. The Number of Errors Identified by the Researchers and the iWrite System and the Number of Errors Correctly Flagged by the System

Using the formula mentioned above, we calculated the precision and recall for the seven categories of

errors (see Figure 2). The results showed that for most error categories, the precision was higher than the recall. This could be attributed to the design of the iWrite system, which is programmed to only identify certain types of errors and may not detect as many errors as human evaluators. Notably, the iWrite system's precision and recall rates for detecting higher-level error types, such as propositional logic, discourse coherence and cohesion, and paragraph structure, were both 0%, indicating that the system was completely unable to detect errors related to macro-discourse elements.

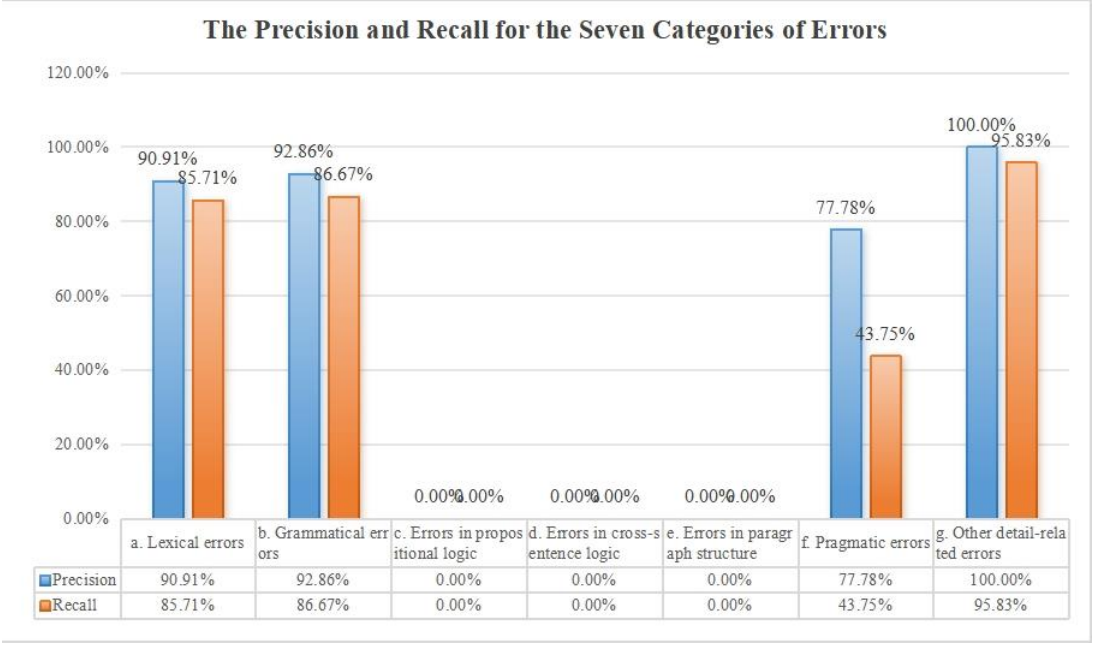

Figure 2. The Precision and Recall for the Seven Categories of Errors

Based on Figures 1 and 2, it is evident that the iWrite system demonstrates varying detection capabilities across different writing aspects, which can be categorized into three main types: high-quality detection, acceptable detection, and detection failure.

High-quality detection includes the detection of lexical errors, grammatical errors, and other detail-related errors, in which the system performs exceptionally well. The detection of detail-related errors achieves the best results, with a precision rate of 100% and a recall rate of 95.83%. This indicates that the iWrite system is capable of providing highly accurate feedback. The detection of lexical and grammatical errors follows closely, with both categories achieving relatively high precision and recall rates, ranging between 90% and 93%.

Acceptable detection focuses on pragmatic error detection. The precision and recall rates for detecting pragmatic errors show the greatest discrepancy, with a precision rate of 77.78% and a recall rate of only 43.75%, indicating a significant number of false positives and false negatives (as shown in Table 3). This suggests that while the system can accurately identify some pragmatic errors, it fails to capture all errors in this category, such as Chinglish (Examples 1 and 2) and stylistic errors (Example 3).

Example 1: Most of people ***have a very tense state of mind***.* (Chen)

Example 2: Student with better psychological quality can ***better play their strength***.* (Guo)

Example 3: Although the examination is very important, but it is not the only standard to measure students, ***please*** treat the examination reasonably.* (Ma)

Detection failure focuses on errors related to propositional logic, discourse coherence and cohesion, and paragraph structure, where the system fails to detect most errors. Among the 26 errors in these areas, the iWrite system only identified 2 propositional logic errors (Table 2), and the feedback for both was incorrect (Examples 4 and 5).

Example 4: ***In the one side***, students should know how are they (Note 1) (grammatical error) doing in this course... ***In the other side***, examination can temper students' mental state.* (Guo)

System feedback: Misuse of preposition: “In” should be corrected to “On”. Misuse of noun: “side” should be corrected to “hand”. Phrase explanation: “On the one hand... On the other hand” expresses two contrasting or opposing views, situations, or factors.

According to the Cambridge Dictionary definition, the phrase “on the one hand... on the other hand” is used to describe two contrasting or contradictory aspects of an issue. In the student's essay, the intention was to discuss two positive aspects of exams, so the appropriate phrase should have been “for one thing... for another”. Thus, the feedback provided by the iWrite system was incorrect.

Example 5: Students may play a disorder (collocational error) in this examination, so they will be ***dejected***.* (Guo)

System feedback: Misuse of verb: “dejected” should be corrected to “rejected”.

The original sentence, which logically follows that students may feel dejected (or disappointed) after failing an exam that tests their mental resilience, is reasonable in terms of propositional logic. The suggested correction, “rejected”, creates a logical inconsistency. Examples 4 and 5 demonstrate that the iWrite system does not provide feedback based on propositional logic or paragraph coherence but instead offers mechanical suggestions that closely resemble the errors made by the student.

In terms of detecting errors related to discourse coherence and cohesion, as well as paragraph structure, the system completely fails. Even obvious errors are not detected. For instance, in discourse coherence, the system fails to identify cross-sentence logical inconsistencies (Example 6) and missing logical connectors between paragraphs (Example 7). In terms of paragraph structure, the system does not detect the absence of necessary conclusions or the lack of supporting evidence for topic sentences, where students merely restate the topic sentence in different words.

Example 6: In most cases, ***although*** the examination only measures people's scores, it is also ***because*** it only measures your grades, not your background, for most people, it is a fair chance to compete with others and change their fate.* (Wen)

Example 7: (paragraph 2) It is harmful to people's family relationships. Most of people's time is spent on work...

(paragraph 3) It is harmful to social economic development. (Chen)

5. Discussion

5.1 Consistency of Research

A total of 116 errors were identified in the essays of the 10 participants, while the iWrite system detected 81 errors, of which 73 were accurate, resulting in an overall precision rate of 90.12% and a recall rate of 62.93%. In previous research on the overall precision of AWE systems, Dikli & Bleyle (2014) and Lavolette et al. (2015) found that the Criterion system had precision rates of 57.5% and 75%, respectively, while Bai & Hu (2017) reported a precision rate of 45.77% for PiGai. In terms of recall rates, Ho áng, G.T.L. & A.J. Kunnan (2016) found a recall rate of 30% for the MY Access system, and Liu, S. & A.J. Kunnan (2016) reported a recall rate of 18.7% for the WriteToLearn system.

Compared to these studies, the iWrite system shows relatively higher precision and recall rates. This may be attributed to the classification of error types in the present study. The iWrite system failed entirely in detecting higher-level errors such as propositional logic, discourse cohesion, and paragraph structure, with precision and recall rates of 0%. However, for surface-level linguistic errors (e.g., lexical, grammatical, and other detail-related errors), the iWrite system successfully identified most errors, with precision and recall rates both above 90% and 85%, respectively. Another factor might be the sample size in this study. Only 10 representative students were selected, and the relatively small sample size resulted in fewer errors in the analyzed materials, which may have inflated the system’s overall precision and recall rates. Moreover, the classification of detail-related errors in this study is similar to the "technical errors" (spelling, spacing, punctuation, and capitalization) identified in Bai & Hu's (2017) research, where PiGai achieved a 98% precision rate for technical errors. This finding aligns with the iWrite system’s performance in detecting detail-related errors in the current study.

5.2 Feedback Failure

The iWrite system demonstrated a complete failure in providing feedback on higher-level aspects of student essays, such as propositional logic, discourse cohesion, and paragraph structure. It neither detected these types of errors nor provided appropriate feedback. This supports the findings of Shi Xiaoling (2012), who noted that teacher feedback surpasses machine feedback in areas related to ideas and discourse structure, as it is more intuitive, clear, and targeted. The failure to provide feedback on macro-discourse elements may be due to the following limitations of AWE systems:

(1) Difficulty in understanding context. Understanding the propositional logic and coherence of a text requires consideration of its broader context. AWE systems may struggle to establish global contextual relationships when processing longer essays, leading to an inability to capture logical issues throughout the text.

(2) Limitations in semantic understanding. Semantic understanding in natural language processing is complex. While AWE systems can detect obvious grammatical errors, their ability to identify deeper semantic errors, such as propositional logic or coherence issues, is limited.

(3) Limited knowledge base. AWE systems typically rely on pre-existing data and knowledge bases for evaluation. If the machine's knowledge base is limited, or if the training data lacks examples of specific

types of errors, the system may fail to comprehend certain domain-specific or specialized errors.

(4) Irreplaceability of human learning and experience. For errors related to discourse cohesion and structure, human intuition and experience are, to some extent, irreplaceable. Machines often lack the intuitive judgment and comprehensive understanding that humans possess in interpreting the overall logic and structure of a text.

5.3 Implications for Teaching

The limitations exhibited by the iWrite system highlight the importance of human evaluation and teacher feedback in writing instruction. This study found that while AWE system have certain advantages in detecting spelling and grammatical errors, they fall short in identifying higher-order writing skills such as propositional logic and discourse structure. Therefore, professional evaluation and feedback from teachers remain indispensable. This confirms the necessity of combining automated evaluation systems with human assessment, as suggested by Lai, Y.H. (2015), Yang Ling (2013), and Li Yihua (2015).

Additionally, teachers should guide students to be aware of the limitations of AWE systems, particularly in areas such as propositional logic, discourse cohesion, structure, and Chinglish. They should encourage students to critically analyze system feedback and provide them with alternative resources (e.g., grammar books, online dictionaries, and corpora) to verify the accuracy of the feedback (Dikli, 2010).

Furthermore, in writing instruction, teachers can place greater emphasis on propositional logic, discourse cohesion, and structure, encouraging students to articulate clear ideas, build strong logical structures, and maintain cohesive discourse. Appropriate tasks and assessment criteria can help students construct coherent transitions and effective paragraph structures, thereby improving their overall writing skills.

6. Conclusion

This study comprehensively evaluated the precision and recall of the iWrite system across seven categories of errors. The results indicate that the system performs well in detecting detail-related errors, lexical errors, and grammatical errors, while its performance in detecting pragmatic errors is relatively weaker. The system exhibited the poorest performance in detecting higher-order errors such as propositional logic, discourse cohesion, and paragraph structure, revealing a complete feedback failure in these areas. These findings further corroborate the conclusions of many researchers: while AWE systems are effective at identifying surface-level errors (e.g., lexical and grammatical errors) and can improve writing to some extent, they have significant limitations in assessing content and discourse structure (Shi Xiaoling, 2012; Yang Xiaoqiong & Dai Yuncai, 2015; Huang Shaoying & Zhang Li, 2015). This is because evaluating logic and discourse requires advanced cognitive abilities that have yet to be developed in machines. The limitations of AI also offer insights for teaching: teachers should consider combining AWE systems with other forms of feedback (e.g., peer and teacher feedback) to

compensate for the shortcomings of automated systems. However, when deciding to implement an AWE system, teachers must be fully aware of its strengths and limitations and integrate it with teaching methods accordingly.

It is worth noting that this study has certain limitations, such as the small sample size, incomplete classification of errors in the essays, and a lack of detailed analysis of error subcategories within the seven main types. Future research could expand the sample size and include precision and recall measurements for subcategories within the seven error types.

References

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology*, *37*(1), 67-81.

Bai, Yun, Han, Jiying, & Wang, Junju. (2022). Effects of feedback modes on linguistic complexity of EFL writing. *Foreign Language Learning Theory and Practice*, *177*(01), 111-121.

Dikli, S. (2010). Nature of automated essay scoring feedback. *CALICO Journal*, 28(1), 99-134.

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.

Han, N. R., Chodorow, M., & Leacock, C. (2006). Detecting errors in English article usage by non-native speakers. *Natural Language Engineering*, *12*(2), 115-129.

He, Zhouchun, & Gong, Yanzhi. (2017). Case study of application of iWrite English Writing Evaluation System 2.0. *Journal of Chengdu Aeronautic Polytechnic*, *3*, 29-32.

Ho áng, G. T. L., & Kunnan, A. J. (2016). Automated essay evaluation for English language learners: A case study of MY Access. *Language Assessment Quarterly*, *13*(4), 359-376.

Huang, Shaoying, & Zhang, Li. (2015). The effects of Juku automated writing evaluation system on English writing ability of students at different language proficiency. *Foreign Languages and Translation*, *22*(04), 70-76+4.

Jahangard, A. (2020). *An online system's effect on Iranians' EFL academic writing performance across different proficiency levels*. Doctoral dissertation, University of Tabriz.

Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P., Sutton, N., et al. (2020). Acawriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research*, *12*.

Lai, Y. H. (2010). Which do students prefer to evaluate their essays: Peers or computer program? *British Journal of Educational Technology*, *41*(3), 432-454.

Lavolette, E., Polio, C., & Kahng, J. (2015). The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology*, *19*(2), 50-68.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1-134.

Li, Yanling, & Tian, Xiachun. (2018). An empirical research into the reliability of iWrite 2.0. *Modern Educational Technology*, (2), 75-80.

Li, Yihua. (2015). A comparative study of three types of feedback in EFL writing based on the dynamic assessment theory. *Foreign Language World*, (3), 59-67.

Liu, S., & Kunnan, A. J. (2016). Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *CALICO Journal*, *33*(1), 71-91.

Liu, Yingliang, & Liu, Jiaying. (2018). The impact of online automated writing evaluation system on the correction of English learners' writing errors: An empirical study based on iWrite. *Foreign Language Education in China (Quarterly)*, (2), 67-74.

Liu, Yingliang, Liu, Shenglan, & Yang, Jincai. (2022). Man-machine cooperative teaching and its application from the sociocultural activity theory perspective: A case of iWrite-assisted English writing instruction. *China Educational Technology*, (11), 108-116.

Palermo, C., & Thomson, M. M. (2018). Teacher implementation of self-regulated strategy development with an automated writing evaluation system: Effects on the argumentative writing performance of middle school students. *Contemporary Educational Psychology*, *54*, 255–270.

Ramineni, C., & Williamson, D. (2018). Understanding mean score differences between the e-rater automated scoring engine and humans for demographically based groups in the GRE General Test. *ETS Research Report Series, 2018*(1), 1-31.

Ranalli, J., Link, S., & Chukharey-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, *37*(1), 8-25.

Roscoe, R. D., Allen, L. K., Johnson, A. C., & McNamara, D. S. (2018). Automated writing instruction and feedback: Instructional mode, attitudes, and revising. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *12*, 23-27.

Shi, Lin. (2021). "iWrite+" integrated practice of teaching, learning, and evaluation of English writing in large classes. *Journal of Mudanjiang University*, *30*(02), 109-114.

Shi, Xiaoling. (2012). A tentative study on the validity of online automated essay scoring used in the teaching of EFL writing: Exemplified by http//www.pigai.org. *Modern Educational Technology*, (10), 67-71.

Tetreault, J. R., & Chodorow, M. (2008). The ups and downs of preposition error detection in ESL writing. In D. Scott & H. Uszkoreit (Eds.), *Proceedings of the 22nd International Conference on Computational Linguistics* (pp. 865-872). Association for Computational Linguistics.

Wang, Mengjun. (2018). *A comparative study of the effects of teacher feedback and automatic scoring system feedback on senior high school English writing*. Doctoral dissertation, Huaibei Normal University.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, *100*, 94–109.

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, *58*, 87–125.

Wu, Yong, & Zhang, Wenxia. (2016). Impact of automated writing evaluation system and teacher feedback on students' writing revision. *Foreign Language Education in China*, (1), 12-19.

Yang, Ling. (2013). On the application of AWE system in high-level students' EFL writing learning. *Modern Educational Technology*, *23*(05), 73-77.

Yang, Xiaoqiong, & Dai, Yuncai. (2015). An empirical study on college English autonomous writing teaching model based on www.pigai.org. *Technology Enhanced Foreign Language Education*, (3), 17-22.

Zhang, Ya, & Zhao, Yonggang. (2020). Efficacy of AEE system feedback on English writing from the perspective of New Literacy Studies Theory. *Foreign Language World*, *199*(04), 88-96.

Zhang, Yu. (2017). *A study on students' perceptions and experiences of computerized feedback by iWrite*. Master's thesis, Central China Normal University.

Zhang, Zhicai. (2018). The integration of iWrite 2.0 with college English writing teaching. *Journal of Hubei Open Vocational College*, (16), 179-183.

Note 1. The above examples selected from the subjects' compositions may also have other aspects of error at the same time, and here each example only shows the feedback given by the iWrite system for the category of error under discussion.