Assessment Model for Language Learners' Writing Practice (in Preparing for TOEFL iBT) Based on Comparing Structure, Vocabulary, and Identifying Discrepant Essays

Duc Huu Pham
Department of English
International University –VNU HCMC
Ho Chi Minh City, Vietnam

Tu Ngoc Nguyen
Department of English
International University –VNU HCMC
Ho Chi Minh City, Vietnam

***Abstract*—This study aims to investigate if learners of English can improve computer-assisted writing skills through the analysis of the data from the post test. In this study, the focus was given to intermediate-level students of English taking final writing tests (integrated and independent responses) in preparation for TOEFL iBT. We manually scored and categorized the students' writing responses into five-point levels for the data to make the software. The results of the study showed that the model could be suitable for computerized scoring for language instructors to grade in a fair and exact way and for students to improve their writing performance through practice on the computer.**

***Keywords*—Computer-assisted writing skills; computerized scoring; integrated and independent responses; model; posttest**

I. INTRODUCTION

The computer has so far been used to assist the assessment of the writing ability of learners of English. This summative assessment helps language instructors to judge the success of their teaching and helps English language learners identify areas that need improvements.

In this paper, we suggest a model to help learners of English to improve their writing skills after an investigation of the Vietnamese students' English performance at a university in Vietnam. There has been significant research on how to assess foreign language students' performance [4]. However, more investigations are needed to develop computer-assisted writing skills for these learners. This study aims at exploiting language criteria with a reference to the scale of the Educational Testing Service [8] as the foundation to build a model that can help to language learners better their writing skills.

The study was carried out to present the development of a computerized assessment to enhance language learners' writing abilities. This study will lead to forming a scoring method, which is more objective and does not involve the participation of many scorers, especially when the individual human factor is always subjective. In this paper, we compared learners' responses with an answer text to find out how much they can match each other. According to [7], a text must consist of collections of clauses, and contextual coherence and cohesion (pronoun/noun reference, ellipsis, substitution). The following is the workflow of document processing.

Language learner's integrated or independent response

Fig 1a. Writing test

Introduction

The materials are concerned with the issue of whether dinosaurs were homoeothermic or poikilothermic creatures. ***The lecturer completely disagrees with the reading's position that*** they could have been homoeothermic. ***This belief is based on*** theories of hibernating patterns and body structure.

Body paragraph I

The reading suggests that because dinosaur fossils were found in the Arctic, they must have been warm-blooded homoiotherms. ***However, the professor contests this, claiming instead that*** dinosaurs were cold-blooded. ***The professor explains that*** the presence of dinosaur fossils in the arctic is a result of the dinosaurs migrating there to hibernate. ***He goes on to say that*** modern reptiles hibernate in cold weather.

Body paragraph II

Also, according to the reading, the adaptation of the dinosaurs' legs underneath their bodies is like that of a mammal or bird and does not resemble modern day poikilothermic, reptilian whose legs are on the sides of their bodies. ***In response, the lecturer says that*** dinosaurs could have adapted this way due to their size. In fact, the professor says that this adaptation was necessary in order for the dinosaur to carry its massive weight. ***This is understandable because dinosaurs*** were hundreds of times the size of modern day reptiles.

Body paragraph III

Finally, the passage claims that because dinosaurs bone structure is similar to that of a modern day homoeothermic mammal or bird, they must have been warm-blooded. ***However, the lecturer refutes this, suggesting that*** dinosaur bone histology is not a result of being warm-blooded. The professor explains that this is because the dinosaurs' rapid growth and evolution adapted their bone structure to carry their large body weight. ***Furthermore, it is noted that*** dinosaur bones would have had to be dense in order to carry their large bodies.

Fig 1b. The answer text [13] in the dataset document sample with highlighted language criteria for assessment.

Assessment: Learner's writing matched with the answer text

Fig 1c. The final stage of assessing a writing

Fig. 1. The process of assessing learner's writing skills

The comparison based on structure and vocabulary and identification of discrepant essays will contribute to

transferring the manual scoring to automatic scoring with the higher accuracy. This high precision was enhanced based on the improvement in the comparison of the documents in not only structure and vocabulary but also the whole layout. Also, this model will help raise learners' test scores.

With the language features of a text, we can design the application of the model in which foreign language learners will have their responses assessed. The model will compare and match the responses with the features of a sample answer text based on the language criteria (addressing the topic, organization, coherence and language use) given in the model. This method has the following characteristics.

- Helping learners to raise their autonomy in acquiring a certain level of foreign language,
- The fastest way of practicing language writing skills for some formatted tests,
- Being objective in the assessment of language writing ability,
- Time-saving in marking learners' writing responses in writing tests,
- Being able to be used as a model for the comprehensive automatic scoring of the written tests.

The rest of the paper is organized as follows: In Related Works, we review some literature. In Development, we present the steps of writing the software. In Results, we show how the software work and compare our method with other methods available to validate our work the results. In Discussion, Conclusion and Future work, we propose using our method as the basis to enhance the use of the software in the semantic aspect.

II. Related Works

The literature review in this study analyses some investigations of computer software programs and the relationship between the issues of computer-assisted language learning and the second language acquisition. Accordingly, theory and practice in the second language learning can be matched together by using modern technology. Also, the development of technology has led to the dispensable incorporation of this medium into the instruction process. Therefore, the computer has become an integral part of the learning activity, through which learners can learn language skills [3].

Several studies on software programs reviewed by [5] and [10] have showed that the validity of the automated writing evaluation (AWE) or the automated essay scoring (AES) system, has not been thoroughly ascertained. Though they seem to be positive in some aspects, tools to review the second language through computer technology still do not meet the requirements of the standards of educational software for written communication such as assessing writing tests [4]. There was a correlation between AES scores and instructors' numerical grades and analytic ratings, which shows the usefulness of AES programs to classroom-based formative assessments and has provided support for us to write this paper.

[18] has developed a new version e-rater v.2.0 with 12 features: 4 in identifying errors in grammar, usage, mechanics, and style, 2 in organization and development, 3 in lexical complexity, 2 in pro-specific vocabulary usage, and one in essay length. However, e-rater v.2.0 still needs improving in three ways: (1) providing more different writing aspects through the theories of writing, (2) ameliorating the model process, and (3) expanding the identification of different essays [18].

Based on the characteristics already mentioned in e-rater v.2.0, we combined the treatment of grammar errors, the set sample vocabulary, and the identification of discrepant essays. We referred to the comparison of event models for naïve Bayes text classification [2], the support vector machines [16], and text categorization algorithms [15], construction of dictionary features covering word groups relevant to semantics or n-grams for text classification [12]. Then we used vector space model [9] to classify writings.

III. Development

This study used the collection data of typical 200 responses (100 integrated writing responses in which students had to combine different skills: reading a passage, listening to a lecture, and then write down the responses, and 100 independent responses in which we gave students a topic to write). Different test takers wrote 200 responses in different exams at a university in Vietnam, which were similar to those in the TOEFL iBT. After we had marked them manually, we found out the statistical difference ($p<0.05$) that we presented in the previous work [6]. We statistically list errors (spelling, grammar, vocabulary – content words and function words) that the students made in writing. Then we carried out the process of constructing a prototype for assisting language writing skills as follows.

Step 1: We define the language criteria about [8] and [14] regarding

1) Addressing the topic: Does the essay address the subject given?

2) Organization: Does the essay have an introduction, body paragraphs (including paragraph structure), and a conclusion?

3) Coherence: Does the essay have the connectives that join or make the sentences go smooth?

4) Language use: Does the essay have spelling errors or grammar mistakes?

Step 2: We collected the data to sample the dataset including main words and phrases according to the standard definition of an integrated and independent structure with the main words and phrases in response to separate parts. Based on that, we defined the structure of the integrated dataset.

1. Integrated

$S_1 = \{I, P_I, P_{II}, P_{III}\}$

In which:

- S_1: The dataset following the standard definition of an integrated structure with the main words and phrases in Introduction part, and in every Body Paragraph I, II, III part.

- I: The dataset of common words and phrases in Introduction part.

$I = \{w_1, w_2, ..., w_n, p_1, p_2, ..., p_n\}$

- P_1: The dataset of words and phrases in Body Paragraph I.

$P_I = \{w_1, w_2, ..., w_n, p_1, p_2, ..., p_n\}$

- P_2: The dataset of words and phrases in Body Paragraph II.

$P_{II} = \{w_1, w_2, ..., w_n, p_1, p_2, ..., p_n\}$

- P_3: The dataset of words and phrases in Body Paragraph III.

$P_{III} = \{w_1, w_2, ..., w_n, p_1, p_2, ..., p_n\}$

- W: Dataset common words in Introduction part.
- P: Dataset common phrases in Introduction part

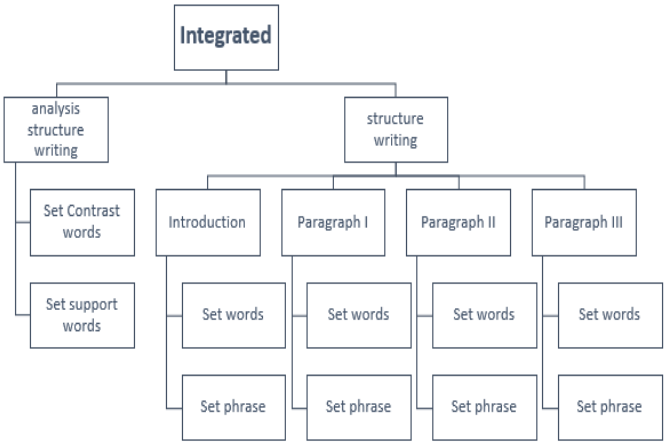

Fig. 2. The dataset structure of integrated part

2. Independent

$S_2 = \{w_1, w_2, ..., w_n, p_1, p_2, ..., p_n\}$

In which:

- W: Dataset common words in Independent.
- P: Dataset common phrases in Independent

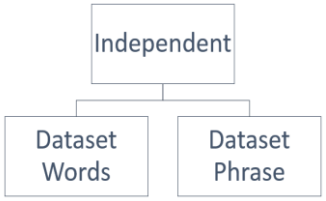

Fig. 3. The dataset structure of independent part

Step 3: Calculation of points for Integrated and Independent writings

Part 1: Comparing with words and phrases in the sample dataset.

Integrated:

Analysing the structure of integrated writing.

Based on the symbol Enter '\n' for recognizing the writing paragraph, we can construe the structure of Introduction, Body Paragraphs I, II, and III.

$W_1 = \{I_1, P_1, P_2, P_3\}$

In which:

- I_1: Introduction paragraph
- P_1: Body paragraph I
- P_2: Body paragraph II
- P_3: Body paragraph III

Checking the number that matches words or phrases in Introduction, Body Paragraphs I, II and III with the sample dataset. After that, based on number matching, we calculated the point of Part 1.

Algorithm:

Input:

- Dataset I, P_I, P_{II}, P_{III} and I_1, P_1, P_2, P_3

Output:

- R_1: Points of Part 1 user writing document **(A)**
- *T:* Array right answer

Initialization:

$R_1 \leftarrow 0; T \leftarrow 0$

// Introduction

For *i=0* **to length**(*I*) *do*

If *I[i]* **stored in** I_1 **then**

$T = T + 1$

// Body I

For *i=0* **to length**(P_I) *do*

If $P_I[i]$ **stored in** P_1 **then**

$T = T + 1$

//Body II

For *i=0* **to length**(P_{II}) *do*

If $P_{II}[i]$ **stored in** P_2 **then**

$T = T + 1$

//Body III

For *i=0* **to length**(P_{III}) *do*

If $P_{III}[i]$ **stored in** P_3 **then**

$T = T + 1$

Return *T*;

Independent :

Checking the number matching of words or phrases in Independent writing.

Algorithm:

Input:

- Dataset w, p

- D: Document Independent

Output:

- R_1: Points of Part 1 user writing document **(A)**

- *T:* Array right answer

Initialization:

$R_1 \leftarrow 0; T \leftarrow 0$

//Word

For *i=0* **to length**(*w*) *do*

If *w[i]* **stored in** *D* **then**

$T = T + 1$

Return *T*;

// Phrase

For *i=0* **to length**(*p*) *do*

If *p[i]* **stored in** *D* **then**

$T = T + 1$

Return *T*;

Following the Rule of this table below for Point of R_1 based on T

T ➔ R_1

T	R_1	T	R_1
>14	5.0	9	2.5
13	4.5	8	2.0
12	4.0	7	1.5
11	3.5	6	1.0
10	3.0	5	0.5

Fig. 4. The table of point levels

Part 2: Comparing the Integrated or the Independent writing with the sample dataset standard document writing of this topic.

We used the comparison based on document classification method [17]. Then we checked some methods classifying documents such as Naive Bayes Text Classification [2], Support Vector Machines [16], and Vector Space Model [9]. After comparing some kinds of the algorithm [15], we saw that the vector calculation is done very quickly as well as very efficiently for the algorithm to optimize the selection of models, allowing for the revenue of the decreased dimensional vector and the visualization of vector space. Also, the vector space model and its variants are still appreciated as in the field of information retrieval. We chose Vector Space Model (VSM) to present the sample documents.

First, we carried out preprocessing which is one of the main components in a typical text classification model [1]. Then we set up the following model to describe the encoding of every document and the creation of a vector for every encoded document [17]:

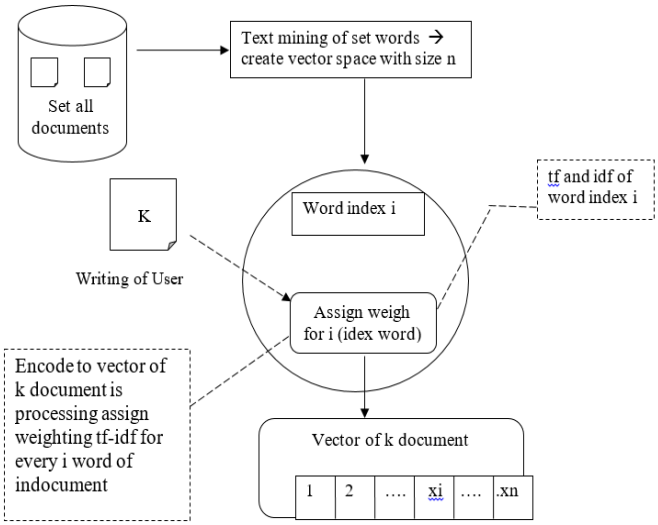

Fig. 5. The model creating vector space

- Creating vector space with size n.

$T = \{D_1, D_2, .., D_n, D_{n+1}\}$

$V = \{V_1, V_2, .., V_n, V_{n+1}\}$

In which:

- T: all documents
- $D_{i \to n}$: Every document in sample data set
- D_{n+1}: User writing document
- V: Vector set of all documents
- $V_{1 \to n}$: Vector of every document in sample dataset
- V_{n+1}: Vector of user writing document
- tf: *term frequency terms weighting*
- idf: *inverse document frequencies*

Algorithm:

Input:

- T: all documents

Output:

- R: Result of distance Vectors

- R_2: Points of Part 2 user writing document

Initialization

$V \leftarrow 0; N \leftarrow 0; R \leftarrow 0$

- N: Set of words for all documents
- S: Vector space

1. Creating vector space

For i=0 **to length** (T)

Separate words on T_i

For j = 0 **to length** T_i

N ← T_i[j]

For i = 0 to **length** (N)

num = 0;

For j = 0 to **length** (N)

If N[i] is **equal** N[j]

num=num+1

if num = 3

S ← N[i]

***2.* Creating vector for every document** [11]

For i=0 **to length** (T)

Separate words on T_i

For j = 0 **to length** T_i

tf = (T_i **words stored in** S) / S

idf = (T_i **words not stored** in S) / S

$V_i[j] = tf_{ij} * \log_n(n / df_i)$

Return V;

***3.* Comparing 2 vectors**

- Applying the calculation distance Euclide in group Minkowski

$$D_E(x,y) = \sqrt[2]{\sum_{i=1}^{n} | x_i - y_i |^2}$$

// calculating distance
For i = 0 **length to** (V-1)
R ← **Distance** (V_{n+1}, V_i)
// **The maximum of the percentage of the similarity of user writing document and the sample dataset document was presented by the minimum value of distance of both vectors.**
Double X = **Min**(R)
R_2 ← (1-X)*5(**Points**) **(B)**
In which:

Double **Distance** (vector V_{n+1}, vector V_i):
double dis = 0;
int weigh(n+1), weigh(i);
For i = 0 **length** to N
String word = S[i];
weigh(n+1) = V_{n+1}.searchHash(word);
weigh(i) = V_i.searchHash(word);
dis = dis + | weigh(n+1) – weigh(i) |^2;
dis = dis $^{1/2}$;

***4.* Total Points (A) & (B)**

TOTAL RESULT = (R_1 + R_2) / 2

IV. RESULTS

As this study was to compare the dataset and learners' responses in the posttest. The responses were the integrated and independent writings. We provided a two-box interface on the screen. The left box contained a reading text (for the integrated) or a topic (for the independent), and the right box was blank for learners to fill in their responses.

We scored both kinds of writings on the language-criterion basis. The language criteria are topic addressing, organization, coherence and language use. We provided the writing topics within a single theme or content area of language, which learners had acquired in the real world or the classroom.

The learners performed the integrated task first. They listened to a short lecture and read a passage from which they had to combine the information to give the responses. They spent five minutes reading the passage provided in the left box, and took notes. After that, they listened to a two-minute lecture, and took notes. Then they used the notes to write their answers in the right box in 30 minutes. After the learners had finished the integrated task, they went on to spend another 30 minutes on the independent writing about a given topic.

The model assessed both the writing tasks and gave the scores on the screen.

The parameters in the model:

- Input
  - Dataset
  - Integrated writing.
  - Independent writing
- Output
  - Errors
  - Scores

The assessment appeared as follows:

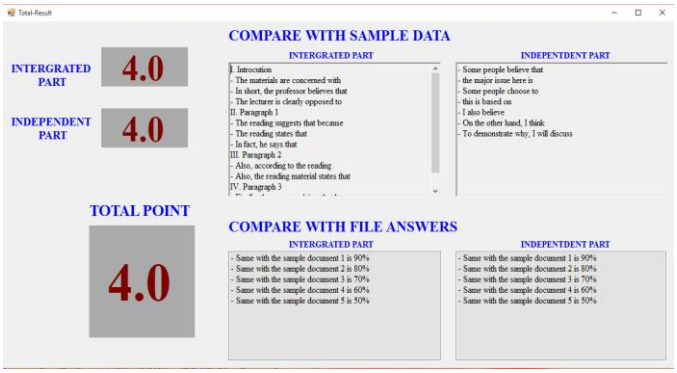

Fig. 6. The scoring model

This scoring method has some advantages over the other scoring methods which aim at the betterment of students' assignments through a continuous, iterative process of writing and revising [5] in that it can help learners to practice writing and get the results through matching words, phrases, and text documents of learners' work and the dataset. Also, this method relieves teachers of the burden of scoring essays which may involve subjective factors.

V. Discussion, Conclusion and Future Work

This performance assessment was for learners at the intermediate level of language proficiency. The design was in accordance with the Raw-to-Scale Score Conversion Tables (Converting Rubric Scores to Scaled Scores) [12] that rate writing performance based on whether it would meet the expectations, exceed the expectations, or not satisfy the expectations for the writing tasks. The performance assessment was valid and reliable according to the university requirements.

The flexible integration of both computer and humans (teacher and student) can increase students' autonomy and raise their awareness of language criteria through students' working with the software independently.

This method is a comprehensive performance assessment. The study contributes to identifying language errors and different kinds of essays to increase the language course outcomes and provide necessary feedback to work out the appropriate methods to improve English language learners' weaknesses. The proposed model can allow users with little knowledge of information technology to access the process of test performance. The software is user-friendly, which is a highly interactive between the software and the user. The analysis in this study ascertains learners' beliefs that they are competent to use computers in their choice of taking writing tests on the computer.

The model is supposed to be an open source so that language instructors can adjust their criteria to be suitable for specific requirements. Future work could use this research as the foundation to improve the implementation of this model in the direction of processing the contextual semantics of the writings for academic English proficiency.

Acknowledgement

The authors would like to thank Vietnam National University HCMC and International University - VNU HCMC for supporting Project C 2014-28-08/HĐ-KHCN, part of which is this article.

References

[1] A. K. Uysal, and S. Gunal, "The impact of preprocessing on text classification," Information Processing & Management, vol. 50.1, 2014, pp. 104-112.

[2] A. McCallum, and N. Kamal, "A comparison of event models for naive bayes text classification," AAAI-98 workshop on learning for text categorization, vol. 752, 1998.

[3] B.B. Nomass, "The impact of using technology in teaching English as a second language," English Language and Literature Studies,vol. 3.1, 2013, p.111.

[4] C. A. Chapelle and D. Douglas, Assessing language through computer technology, Ernst Klett Sprachen, Cambridge: Cambridge University Press, 2006.

[5] C. E. Chen and W. E. Cheng. "Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes," Language Learning & Technology, vol. 12.2), 2008, pp. 94-112.

[6] D. H. Pham, "A Computer-Based Model for Assessing English Writing Skills for Vietnamese EFL Learners," unpublished.

[7] E. Suzanne, Introduction to systemic functional linguistics, A&C Black, 2004.

[8] Educational Testing Service, TOEFL iBT Scores, 2005. Retrieved from http://www.hhl.de/fileadmin/texte/_relaunch/Conversion_Table_TOEFL_(PBT,CBT,iBT).pdf

[9] G. Kanaan and A. Jafar, "A comprehensive comparative study using vector space model with k-nearest neighbor on text categorization data," Asian Journal of Information Management, vol. 2.1, 2008, pp.14-22.

[10] J. Choi, "The impact of automated essay scoring (AES) for improving English language learner's essay writing (Doctoral dissertation, University of Virginia)," 2010, retrieved from http://www.researchgate.net/profile/Jaeho_Choi6/

[11] L. P. Jing , H. K. Huang, and H. B. Shi, "Improved feature selection approach TFIDF in text mining," Machine Learning and Cybernetics, 2002, Proceedings, 2002 International Conference on vol. 2, IEEE, 2002, pp. 944-946.

[12] M. Brooks, S. Amershi,B. Lee, S.M. Drucker, A. Kapoor, and P. Simard, "FeatureInsight: Visual support for error-driven feature ideation in text classification," in Visual Analytics Science and Technology (VAST), 2015 IEEE Conference, pp. 105-112, IEEE.

[13] R. Hahn, A1 TOEFL writing (iBT) (Korean edition), 2008.

[14] Raw-to-Scale Score Conversion Tables by Educational Testing Service, TOEFL iBT Scores, 2005. Retrieved from http://www.etweb.fju.edu.tw/elite/ETS%20%20ibt%20TOEFL%20Converting_Rubric.pdf

[15] S.H.I Yong-feng and Z. Yan-ping, "Comparison of text categorization algorithms," Wuhan university Journal of natural sciences, vol. 9.5, 2004, pp.798-804.

[16] T. Joachims, Text categorization with support vector machines: Learning with many relevant features. Springer Berlin Heidelberg, 1998, pp. 137-142.

[17] V. Korde and C. N. Mahender, "Text classification and classifiers: A survey," International Journal of Artificial Intelligence & Applications, vol. 3.2, 2012, p.85.

[18] Y. Attali and J. Burstein, "Automated essay scoring with e-rater V. 2.0 (ETS RR-04-45),". Educational Testing Service, Princeton, NJ, 2005.