Constant Evaluation of L2 Students' English Writing Ability

Li Li
Shanghai University of Political Science and Law
Shanghai, China
e-mail: lily2211@126.com

***Abstract*—Writing teaching is an indispensible part of college English teaching in China. Compared with L1 students, the writing teaching of L2 students is much more challenging. Recent years, Automatic Evaluation System (AES) has been more frequently employed to score students' essays by teachers. However, AES cannot replace teachers for the following shortcomings. 1) The essay evaluation is not very precise; 2) It cannot truly reflect a student's real writing ability based on independent scores. To solve the problems, this paper proposes a method of constantly evaluating L2 student's real writing ability, which evaluates, compares and analyzes several essays written by a student within a certain period. First, the framework of constant evaluation of writing ability is proposed, which consists of a single essay evaluation and a timeline-based evaluation. Next, several aspects of automatic evaluation of a single essay are improved. Then, a timeline-based writing ability evaluation method is proposed based on the knowledge graph. Finally, the experiments are conducted, the results of which show that the proposed method is effective in evaluating a student's real writing ability.**

Keywords-writing evalution; AES; auto scoring; L2 students; timeline-based evalution; knowledge graph

I. INTRODUCTION

Writing teaching is an indispensible part of college English teaching in China. Compared with L1 students, the writing teaching of L2 students is much more challenging. It is apparent that L2 students' English proficiency is lower than L1 students'. Each basic element of an article may be a barrier to writing for L2 students. That is to say, there will be problems of the choice of words, sentences, paragraphs or discourses. Therefore, it has been widely acknowledged that English writing teaching is a rather tough and challenging task. In order to increase the efficiency of evaluating students' essays and decrease the workload of teachers, Automatic Evaluation System (AES) has been more frequently employed in the writing teaching to help teachers to evaluate students' essays and turned out to be effective[1].

However, AES can't provide a precise evaluation because of the limitation of text understanding technologies, so it is impossible for AES to replace teachers to evaluate students' essays. What's more, the score of a single essay can't truly reflect a student's real writing ability. In the long run, if students depend much on it, AES may have a negative impact on their learning plan. In effect, a student's writing ability is influenced by diverse factors, apart from their own knowledge. For example, in different conditions, a student will present obviously different writing abilities. In a good condition, the choice of words, sentence-making and sentence variety will be at the high level; otherwise, they will be poor, which shows the volatility of one aspect of his/her abilities in writing. It shouldn't become a focus of teachers if a student shows that his/her abilities decrease temporarily in one aspect. On the contrary, it will have a negative effect on the student's learning initiative if teachers highlight this problem. Especially in China, there are a large number of students, it is rather hard for teachers to follow every student and constantly evaluate students' writing ability.

To solve the problems, this paper proposes a method of constantly evaluating L2 student's real writing ability, which evaluates, compares and analyzes several essays written by a student within a period. First, the framework of constant evaluation of writing ability is proposed, which consists of a single essay evaluation and a timeline-based evaluation. Next, several aspects of automatic evaluation of a single essay are improved. Then, a timeline-based writing ability evaluation method based on knowledge graph is proposed. Knowledge graph of writing ability is constructed to make a comprehensive description of a student's writing ability, including use of words, choice of sentences, coherence, logic etc. By analyzing the changes of knowledge graphs with time going on, the real writing ability of a student can be tested and obtained. Finally, the experiments are conducted and the experimental results show that the proposed method is effective in evaluating a student's real writing ability.

The rest of this paper is organized as follows: section II introduces related work, and section III puts forward a framework of constantly evaluating one's writing ability. Section IV and section V discusses the method of evaluating a single essay and the method of constantly evaluating one's writing ability respectively. In section VI, the experiment is performed to verify the proposed method and the last section draws a conclusion.

II. RELATED WORK

AES is defined as the computer technology that evaluates and scores the written prose [2]. With the growing development of computer technology, AES systems have been improved a lot and are being improved. In order to make the large-scale essay scoring process more practical and effective, Project Essay Grader (PEG) was developed by Ellis Page upon the request of College Board [3]. It utilizes proxy measures to evaluate the quality of essays. But it has been criticized for ignoring the semantic aspect of essays and focusing more on

DOI reference number: 10.18293/SEKE2017-149

the surface structures [4][5]. With the advance of computer technology, more AES systems, such as Intelligent Essay Assessor (IEA), the Electronic Essay Rater (E-Rater) were developed to meet the requirements. IEA analyzes and scores an essay using a semantic text analysis method called Latent Semantic Analysis (LSA). It is claimed that unlike other AES systems, IEA's main focus is more on the content-related features rather than the form-related ones. However, this doesn't mean that IEA offers no feedback on formal aspects, i.e., grammar and punctuation, in an essay. However, the system doesn't evaluate the creativity and reflective thinking. E-Rater was developed by the Educational Testing Service (ETS) to evaluate the quality of an essay by identifying linguistic features in the text [6]. E-Rater uses natural language processing (NLP) techniques, which identify specific lexical and syntactic cues in a text, to analyze essays [5]. Later, artificial Intelligence (AI) was introduced to the development of AES systems. IntelliMetric, developed by Vantage learning, is known as the first essay-scoring tool that was based on AI. Like, E-Rater, IntelliMetric relies on NLP, which determines "the meaning of a text by parsing the text in known ways according to known rules conforming to the rules of English language" [7]. Another AES system, named My Access, is known as the instructional application of IntelliMetric. My Access is a web-based writing assessment tool that relies on Vantage Learning's IntelliMetric automated essay scoring system. The main purpose of the program is to offer students a writing environment that provides immediate scoring and diagnostic feedback, which allows them to revise their essays accordingly and motivates them to continue writing on the topic to improve their writing ability. ETS' Criterion, a web-based instructional writing tool, uses the E-Rater engine to provide both scores and targeted feedback. It allows students to improve their writing skills while working independently with immediate, detailed feedback on grammar, spelling, mechanics, usage, and organization and development.

III. FRAMEWORK OF CONSTANTLY EVALUATING STUDENTS' WRITING ABILITY

According to the above analysis, it is a gradual process to enhance students' writing ability and there exist ups and downs in the writing quality with individual factors and changes of the external surroundings outside. L2 students will show different features of changes in their abilities of different writing elements with the time passing. For example, there remains greater influence on diction (choice and use of words). As to different themes and styles, there are great differences. If they haven't written an essay of the same subject for a long time, students' ability of using the words of the subject will obviously decrease. Comparatively speaking, the ability of writing arrangement is rather stable. As long as they keep practicing writing, students can maintain this ability and are likely to enhance it. So considering such factors as fluctuation of writing ability, the evaluation of a single essay can't reflect a student's real writing ability objectively and comprehensively.

The paper proposes the method of constantly observing and evaluating a student's writing ability, whose framework is shown in Fig. 1 and mainly consists of two modules.

1) Single essay evaluation module

This module realizes the evaluation of a single essay, which is similar to the function of AES. The difference between them is that this module also provides additional data for the writing ability evaluation module so as to build the forgetting curve of the student.

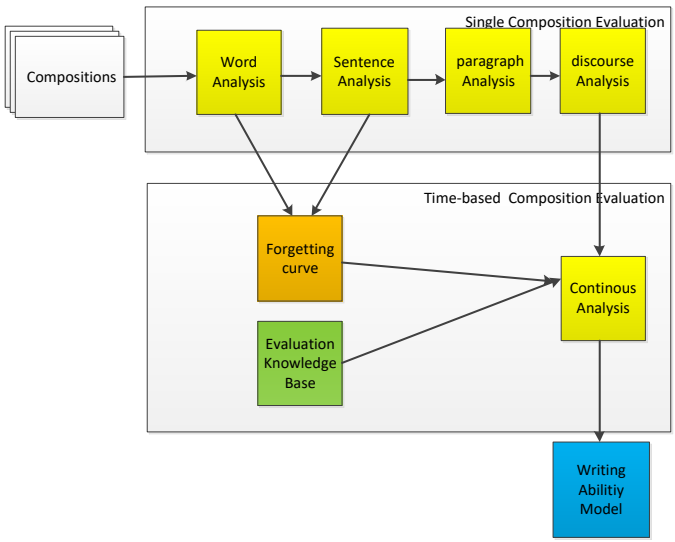

Figure 1. Framework of constant evaluation of writing ability

This module is composed of word evaluation, sentence evaluation, paragraph evaluation, discourse evaluation and an overall evaluation by summarizing these evaluations of writing elements. The overall score of an essay gives a direct and straightforward evaluation to students, while the evaluation of writing elements can present students their weak points in certain aspect, which has a much more instructive effect on writing teaching.

2) Writing ability evaluation module

This module is to evaluate students' writing ability comprehensively within a certain period. Namely, based on evaluations of several essays within a certain period, the timing analysis is adopted to analyze different writing ability evaluation indexes and then the knowledge graph of writing ability is built.

The indexes of writing ability are related to each other, so it is not likely to improve the overall level of the students effectively only by intensifying the training of one element. For instance, a student leaves much room to improve his/her ability of making a sentence. If he/she only practices making a sentence, he/she can't really improve his/her overall writing ability, for how to make a sentence has much to do with such elements as words, phrases, grammar, sentence patterns.

Knowledge graph is a model used to show the relation between knowledge points, which can offer both the overall view and the detailed view. Therefore, the writing ability is modeled by knowledge graph, which can also show the writing ability of a student from both the overall and the detailed aspects.

IV. SINGLE ESSAY EVALUATION

Compared with teachers' evaluation, AES has the following shortcomings when just used to evaluate a single essay: 1) the precisions of some evaluation indexes are low; 2) some evaluation indexes cannot be assessed by AES. However, the distinguishing advantage is its high speed, which is suitable for the heavy task of constantly evaluating a large number of essays of the students by using our method.

A. Evaluation indexes

Taking both the difficulties in the implementation and the precision of AES into consideration, we select some feasible and practicable evaluation indexes to evaluate a single essay evaluation, which are shown in Table I.

TABLE I. ESSAY EVALUATION INDEXS

One class index	Two class index	Weight
1.Word & phrase	1.1 Spelling	0.3
	1.2 Grammar	0.3
	1.3 Vocabulary	0.5
2.Sentence	2.1 Punctuation	0.2
	2.2 Sentence structure	0.3
	2.3 Sentence grammar	0.3
	2.4 Sentence pattern	0.5
3.Pragraph	3.1 Sentence coherence	0.6
	3.2 Topic relevance	0.8
4. Discourse	4.1 Ideas	0.7
	4.2 Organisation	0.6
	4.3 Pragraph coherence	0.6
	4.4 Theme relevancy	0.8

B. Evaluation method

Among the above-mentioned evaluation indexes, some can be easily realized by means of computer algorithm. For example, the evaluation of spelling and grammar can reach high precision with the support of the dictionary and grammar corpus. But evaluations of coherence, unity and transition are not easy to carry out. Some evaluation indexes in our method are discussed as follows.

1) Vocabulary: To vocabulary, we focus on the evaluation of the breadth and depth of the use of words. Taking into account the actual situation of Chinese students, we select ten levels of vacabularies as the standards, shown in Table II, whose difficulties increase gradually.

The ability of choosing and using the words is calculated based on the distribution of words in the essay, which is calculated as

$$v = \sum_{i=1}^{n} (\frac{l_i}{m} \times w_i) = \frac{1}{m} \sum_{i=1}^{n} (l_i \times w_i), \quad (1)$$

where, m is the number of words in the essay, l_i is the number of words in the level i vocabulary, and $\frac{l_i}{m}$ is the ratio of level i in the essay. The ability of using the words, v, is the weighted sum of the use of words at all levels. It is valuable to know the changes of a student's v within the given time for the purpose of evaluating his ability of using the words.

TABLE II. DIFFICULTY LEVELS OF VOCABULARY

Level #	Vocabulary	Weight
1	College English Test Band 1	0.1
2	College English Test Band 2	0.2
3	College English Test Band 3	0.3
4	College English Test Band 4	0.4
5	College English Test Band 5	0.5
6	College English Test Band 6	0.6
7	Test for English Majors Band 4	0.7
8	Test for English Majors Band 8	0.8
9	TOFEL	0.9
10	GRE	1.0

2) Sentence Pattern

The variety of sentence patterns is viewed as a main evaluation index. If different sentence patterns are employed to express one's opinions in an essay, the whole will be richer. A succession of simple sentences may be jerky and choppy, a succession of loose sentences relaxed or even slovenly and a succession of periodic sentences formal, stiff and difficult to follow. Too many sentences of the same pattern following one another are at least monotonous.

According to different standards, sentence can be divided into different types. According to their use, sentences are declarative, interrogative, imperative or exclamatory. According to their structure, sentences are simple, compound, complex or compound-complex. From the rhetorical point of view, sentences are loose, periodic and balanced.

In the English writing, the method of evaluating the variety of sentence patterns is used to calculate the distribution of different sentence patterns in an essay. The calculation method is

$$s = \frac{sp}{sn} - \frac{1}{sm} \sum_{i=1}^{m} \mathrm{n}sp_i \ , \quad (2)$$

where, $\frac{sp}{sn}$ denotes the ratio of sentence patterns used in the writing, sn and sp denotes the number of the variety of sentence patterns and the number of the variety of sentence patterns in the writing respectively; $\frac{1}{sm} \sum_{i=1}^{m} \mathrm{N}spi$ measures

repetition of sentence patterns, *sm* is the number of specific sentence patterns, nsp_i is the frequency of repeating the sentence pattern I, *sm* is the number of sentences in the writing.

3) Coherence

The coherence of an essay includes the text coherence and the semantic coherence.

Text coherence is a literary technique that refers to the meaningful connections that readers perceive in a written text. In other words, it is a well-written piece that is not only consistent and logical, but also unified and meaningful. It makes sense when read as a whole. The structure of a coherent paragraph could be general to particular and particular to general or any other format. In order to achieve the effect of coherence, proper transitions have to be employed which are used to make a connection clear [9]. The local coherence and global coherence of an essay can be evaluated by analyzing the transitional words and phrases.

Semantic coherence refers to the coherence of content, namely, the association of the whole passage from the beginning to the end. Given that L2 students don't usually write a long essay and each paragraph consists of several sentences, it is not easy to calculate the semantic coherence of sentences within a paragraph. This paper mainly focuses on the semantic coherence among paragraphs, the algorithm of which is described in Algorithm 1.

Algorithm 1: Semantic coherence evaluation of an essay
Input: C // a student's essay
Output: sc // semantic coherence

1) For each p_i in C do // p_i is the i[th] paragraph of C
2) extract keywords of p_i using TF-IDF
3) denote $p_i = \{k_1, k_2, ..., k_n\}$
4) End for
5) Mine association rules based on paragraphs and get the association rules set *ARs*
6) For i=1 to n do
7) Calculate the association degree sc_i between p_i and p_{i+1}
8) End for
9) $sc = \frac{1}{n-1}\sum_{i=1}^{n-1} sc_i$

In the algorithm, each paragraph is denoted by VSM (from step 1 to 4). And then, each paragraph can be seen as a transaction and the essay can be seen as a transaction set. Each keyword of a paragraph can be seen as an item. As a result, the association rule mining algorithm can work on it. In this step we can get the association rule set *ARs* in this paragraph. Step 7 finds out all the association rules in ARs which can bridge the neighboring paragraphs. Weighs of all the selected association rules can be summed to get the sc_i. In the end, the average of the semantic coherence between each pair of neighboring paragraphs is calculated and then viewed as the total semantic coherence of the essay.

4) Theme relevancy

What students write must center about a given theme, otherwise they will have to face the danger of straying away from the point. By calculating the similarity between each paragraph and the theme, we can obtain a value of theme relevancy. The method similar to Algorithm 1 is employed to show each paragraph's VSM, namely, $p_i = \{k_1, k_2, ..., k_n\}$. Meanwhile, the requirements of the writing can be represented by VSM, $theme = \{t_1, t_2, ..., t_n\}$.

The similarity between p_i and *theme* can be calculated by COS method

$$sim(p, theme) = \frac{\sum_{i=1}^{n}(k_i \times t_i)}{\sqrt{\sum_{i=1}^{n}(k_i)^2} \times \sqrt{\sum_{i=1}^{n}(t_i)^2}} \quad (3)$$

And the total theme relevancy can be gotten by calculating the average of similarity of all paragraphs.

V. COMPREHENSIVE EVALUATION OF WRITING ABILITY BASED ON TIMING ANALYSIS

A. Knowledge Graph of writing abliltiy

To offer students their global evaluation and specific evaluation on their writing ability, the paper introduces knowledge graph to present students' writing ability, shown in Figure 2. In the figure, circle refers to evaluation indexes at different levels. Yellow circle(dotted box) indicates level 0, which is students' global evaluation; green one(dashed box) refers to level 1, which is similar to the first row of Table I. Gray one(solid box) refers to lever 2, similar to the second row of Table I. The size of the circle shows the weight of the nodes, here referring to the score a student gains on the index. In order to directly demonstrate the continuous changes of students' writing ability, the model employs numbers from 0 to 100 to show the weight of nodes.

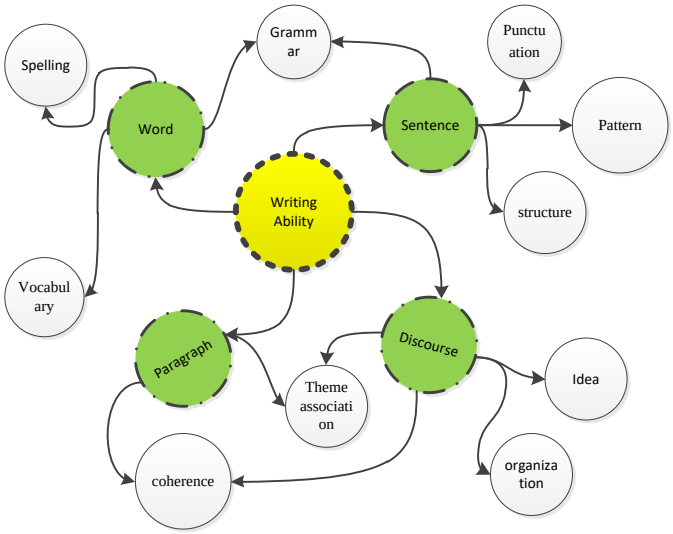

Figure 2. Knowledge graph of writing ability

In Figure 2 drawn from Table 1, the arrow shows the relevancy exists between the two evaluation indexes. As to

different students, their knowledge graph of writing ability is different, is listed as follows:

a) The size of nodes: The size of nodes indicates that students have different abilities in a certain aspect of writing. According to the size of nodes, students can do some targeted and specific training, which contributes to improving their overall writing ability.

*b) The length of sides:*The length of *edge* indicates that students have different connections between the indexes of writing ability. If he/she gets a low score on *an* index, he/she has to intensify his training of this index and relevant indexes, and then they will bear fruit.

B. Build the Knowledge Graph of writing abliltiy

Because we are still keeping working on mining the relation between evaluation indexes, here we mainly discuss the method of calculating the node weight of knowledge graph of writing ability.

According to the above analysis, the evaluation of a single essay cannot be used to judge a student's writing ability. We must eliminate ill effects of some factors including the surroundings on students' writing ability from the perspective of time. Therefore, the model of a student's writing ability needs to save the historical evaluation data. The model of a student's writing ability is represented by several knowledge graphs with time label. Figure 3 presents the first level of evaluation index of a student's knowledge graph in the time sequence. This Figure shows the changes in his /her writing ability and provides historical data for us to calculate his/her next knowledge graph as well.

When analyzing a single index from the aspect of time, we will find the influence of such factors as the surroundings will be singular points in the writing ability curve, such as the fourth node shown in Figure 4. If the cycle is short, these singular points can be effectively eliminated by means of linear fitting. If the cycle is long, Detrended Fluctuation Analysis (DFA) will be adopted to eliminate the detrended fluctuation [8]. Detrended fluctuation analysis (DFA) is a method of determining the statistical self-affinity of a signal.

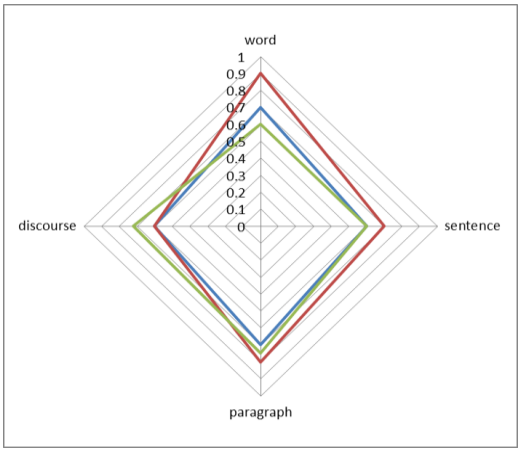

Figure 3. Time-based writing ability model

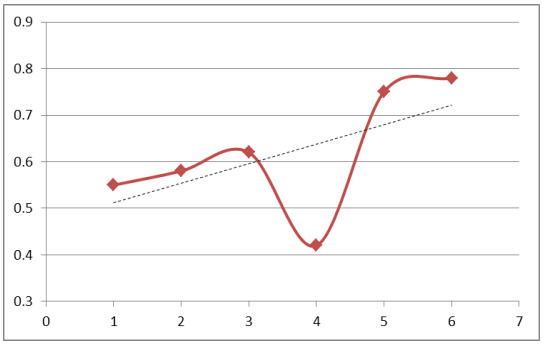

Figure 4. Time-based analysis of writing ability index

VI. EVALUATION

To verify the effect of the proposed method, the comparative experiment is conducted to compare the precision of evaluation of teachers with that of the proposed method.

A. Participants

Participants are involved in ten students and ten teachers. Twenty essays written by each in two semesters are regarded as experimental data. Ten teachers fall into two groups, five teachers in each. Among them, one is experimental group, the other is expert group. The experimental group mainly evaluates students' essays and their writing ability. While the expert group judges the evaluation results of teacher and the proposed method.

B. Procedures

1) Teachers in the experimental group adopt a method of global scoring to evaluate each essay in the sequence of time when students complete them. Meanwhile, they have to evaluate them based on the first level and second level indexes respectively.

2) Teachers in the experimental group require timing analysis of twenty essays of each student. Every four essays is seen as an observation point and builds a knowledge graph of writing ability, so there are five knowledge graphs totally.

3) The first two steps are repeated by computer based on the proposed method.

4) The teachers in expert group compare the evaluations of teachers in the experimental group with those given by computer.

C. Experimental Results and Analysis

The experimental results are shown in Figure 5 and Figure 6. Figure 5 shows the comparative experiment on the precision of writing evaluation. It must be pointed out that each essay should be evaluated by five teachers and finally take the average. The data in Figure 5 is the average of twenty essays of each student.

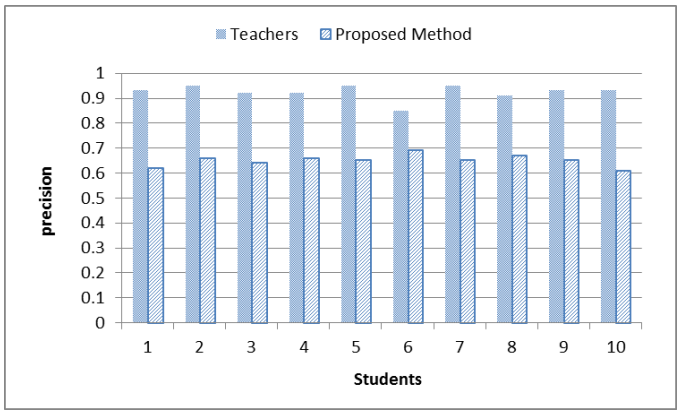

Figure 5. Comparison of scoring precision between teachers and the proposed method

Figure 6 shows the result of comparative experiments on the construction of knowledge graph. The comparative item is the precision of construction method. To make the results more general, the final knowledge graph of each student is constructed on the average of knowledge graphs constructed by five teachers independently.

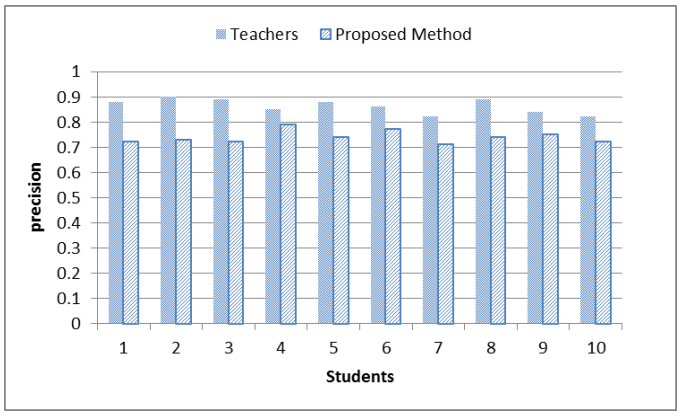

Figure 6. Comparison of knowledge graphs constructed by teachers and the proposed method

The results of comparative experiment show that the precision of teachers' evaluation is better than the proposed method, in terms of evaluation of a single essay or comprehensive evaluation of writing ability. But the results also indicate that the gap between machine scoring and manual scoring narrows when they conduct a comprehensive evaluation.

The experimental process still shows that the speed of teachers' evaluation is much slower than machines'. What's more, in the large-scale evaluation, the precision of manual evaluation will decrease. The results indicate that the precision of machine evaluation is bigger than 0.7, and the error is less than 0.15, compared to the accuracy of teachers. So in the large scale evaluation, machine can take the place of teachers.

VII. CONCLUSIONS

As to the problem of evaluation of L2 students' English writing ability, the method of constantly evaluating their writing ability is proposed and timing knowledge graph is constructed to reflect students' overall writing ability so as to have the knowledge of students' real writing ability. The experimental results show that the proposed method is effective in evaluating a student's real writing ability.

The future work is involved in carrying out a deep research on mining the relation of the indexes of knowledge graph. At the same time, the research on how to work out learning plan automatically by means of timing evaluation results is going to be carried out.

ACKNOWLEDGEMENTS

This research was financially supported by 2017 Scientific Research Project of Shanghai University of Political Science and Law.

REFERENCES

[1] Mcnamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. Assessing Writing, 23, 35-59.

[2] M. D. Shermis, and J. C. Burstein, "Automated essay scoring: A cross-disciplinary perspective", Routledge, 2003.

[3] E. B. Page, "Project essay grade: PEG," Automated essay scoring: A cross-disciplinary perspective, pp.43-54, 2003.

[4] G. K.Chung, and H. F. O'Neil, Methodological approaches to online scoring of essays. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles,1997.

[5] K. Kukich, "Beyond automated essay scoring," IEEE intelligent systems, vol.15, no.5, pp.22-27,2000.

[6] J. C. Burstein, D. Marcu, S. Andreyev, and M. Chodorow, "Towards automatic classification of discourse elements in essays," In Proceedings of the 39th annual meeting of the Association for Computational Linguistics, pp.90-92, France, 2001.

[7] S. Elliot, "IntelliMetric: from here to validity. Automated essay scoring: A cross-disciplinary perspective," Automated essay scoring: A cross-disciplinary perspective, pp. 71-86, 2003.

[8] Peng, C.K.; et al. (1994). "Mosaic organization of DNA nucleotides". Phys. Rev. E. 49: 1685–1689. doi:10.1103/physreve.49.1685

[9] Duane H. Roen, "Coherence." Encyclopedia of Rhetoric and Composition: Communication From Ancient Times to the Information Age, ed. by Theresa Enos. Taylor & Francis, 1996