A Study on an Intelligent Algorithm for Automatic Test Paper Generation and Scoring in University English Exams

Han Yang

Public Education Department, Anyang Vocational and Technical College, Anyang, Henan 455000, China
E-mail: r931hy@yeah.net

Received 10 June 2023; Accepted 17 August 2023;
Publication 18 November 2023

Abstract

This paper mainly studied the automatic test paper generation and scoring problems in university English exams. Firstly, an automatic test paper generation model was established. Then, an improved genetic algorithm (IGA) was designed for intelligent test paper generation, and it was also used to automatically score answers to Chinese-to-English translation questions in terms of syntax and semantics. It was found that compared with the traditional GA and particle swarm optimization algorithm, the IGA method was faster in generating test papers, with an average generation time of 25 s, and had a higher success rate (94%), suggesting higher validity, and the difficulty and differentiation degrees of the test papers were closer to the preset values. The results of automatic scoring also had a correlation of more than 0.8 with the manual scoring results. The results prove the effectiveness of the

Journal of ICT Standardization, Vol. 11_4, 391–402.
doi: 10.13052/jicts2245-800X.1144
© 2023 River Publishers

automatic test paper generation and scoring method. It can be promoted and applied in practice to enhance the security and fairness of large-scale English exams, as well as achieve objectivity and consistency in scoring.

Keywords: Intelligent test paper generation, university English, genetic algorithm, automatic scoring, semantics.

1 Introduction

Examinations are an important part of school education, traditionally involving manual creation of test papers by teachers. However, this process consumes substantial human and material resources for question setting, review, examination organization, and answer scoring. The application of computer technology can realize the automatic test paper generation and scoring [1], and related methods have become the focus of researchers. Wang et al. [2] studied two popular unit test generation tools, EVOSUITE and Randoop, and found through a empirical research that the two tools had significant improvements in code coverage and variance scores. Yang et al. [3] combined particle swarm optimization with Unified Modeling Language (UML) modeling tools to achieve intelligent test paper generation for online exams and demonstrated the performance of the method through experiments. Wang et al. [4] studied the automatic scoring of Chinese fill-in-the-blank questions. They used an improved P-means model to calculate the semantic similarity between standard answers and exam answers and found through experiments that the highest accuracy of the method reached 94.3%. Yuan [5] established an automatic scoring system for college English composition based on the multiple regression method and conducted experiments to validate its effectiveness. Zhao and Li [6] developed an examination system using Java web technology, which combines client-side programming and server-side programming to provide functions such as question management, paper generation, and online testing. Yağci and Üünal [7] designed an adaptive online examination system that can automatically determine different sets of questions for each student through interactive means in order to more effectively assess their abilities. A framework for an online examination system based on blockchain was proposed by Sadayapillai and Kottursamy [8]. It employed encryption technology to gather data, guaranteeing the integrity of candidates' answers and preventing problems like test paper leaks or answer cheating. Nevertheless, research into automated question paper generation and scoring during exams is still at an early stage, with a dearth of specific application studies conducted

in university English exams. Furthermore, the effectiveness of automated test paper generation and scoring algorithms remains uncertain. Therefore, this article conducted in-depth research on the automatic paper generation and scoring methods for college English tests. An improved genetic algorithm (IGA) was designed to achieve intelligent paper generation, and then grammar and semantics were combined to implement automatic scoring for answers to Chinese-to-English translation questions. The reliability of this method was demonstrated through experiments. This paper makes a contribution to improving the intelligence and automation of the university English examinations. The method can be applied in other examinations to promote the development of informatization in school work.

2 Automatic Test Paper Generation and Scoring for University English Examinations

2.1 A Model for Generating Test Papers for English Exams at University

Taking university English exams as an example, the test papers needs to meet conditions in terms of difficulty, coverage, and differentiation. Suppose that the test paper obtained by automatic generation is composed of N questions and every question contains M attributes, then the mathematical model of the test paper is written as:

$$A(a_{ij}) = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{bmatrix}. \tag{1}$$

In the above equation, each row represents all the attributes of a question. Based on the actual examination situation and the nature of automatic test paper generation, eight attributes are defined, as shown in Table 1.

The constraints considered in this paper in automatic test paper generation are as follows.

(1) Total score: $S_A = \sum_{i=1}^{n} a_{i3}$, where S_A is the total score of the test paper and a_{i3} refers to the score of test question i.
(2) Chapter score: $C_K = \sum_{i=1}^{n} a_{i3}c_{ik}$, where C_K is the score of chapter K and c_{ik} is the score of test question i in chapter K. When the chapter to which the knowledge point belongs $a_{i1} = K$, $c_{ik} = 1$; otherwise, $c_{ik} = 0$.

Table 1 Specific attributes in the test paper generation model

Code	Definition	Description
a_1	Chapter	The chapter related to the test content
a_2	Question type	Such as multiple choice questions, judgment questions, etc.
a_3	Score	Scores of test questions
a_4	Time	Exam duration
a_5	Difficulty	The difficulty of a test question. $a_5 = \overline{X}/A$, where $\overline{X}$ is the average score of the students on the question and A is the full score of the question.
a_6	Degree of differentiation	The role of test papers in differentiating students' abilities. $a_6 = (\overline{H} - \overline{L})/A$, where H is the top 20% of the scores, called the high score band, L is the lowest 20% of the scores, called the low scoring band, and $\overline{H}$ and $\overline{L}$ are the average scores of the high and low score bands.
a_7	Degree of coverage	The coverage of the content required by the examination syllabus in the examination paper. $a_7 = C_k/C$, where C is the number of chapters required to be covered by the syllabus and C_k is the number of chapters actually included in the test paper.
a_8	Degree of exposure	The number of times the test question has been used. If it has been used, its value is 1; otherwise, its value is 0.

(3) Question type score: $T_E = \sum_{i=1}^{n} a_{i2} t_{ie}$, where T_E is the score of question type E and t_{ie} is the score of test question i in question type E. When question type $a_{i2} = E$, $t_{ie} = 1$; otherwise, $t_{ie} = 0$.
(4) Time: $J_A = \sum_{i=1}^{n} a_{i4}$.
(5) Difficulty: $D_A = (\sum_{i=1}^{n} a_{i3} a_{i5})/S_A$, where D_A is the difficulty of the test paper.
(6) Degree of coverage: $F_A = \sum_{i=1}^{n} a_{i3} f_{ia}$, where F_A is the sum of the scores of different knowledge points and f_{ia} is the score of test question i in knowledge point F. When $a_{i3} = F$, $f_{ia} = 1$; otherwise, $f_{ia} = 0$.
(7) Degree of differentiation: $Q_A = (\sum_{i=1}^{n} a_{i6} a_{i7})/S_A$.
(8) Degree of exposure: $B_A = (\sum_{i=1}^{n} a_{i8})/N$.

2.2 Improved Genetic Algorithm-based Test Paper Generation Algorithm

The problem of test paper generation can be regarded as a solution to a multi-objective constraint problem. Although GA can solve this problem, it suffers

from issues like premature convergence and low computational efficiency [9]. Therefore, an improved version called IGA is developed. The specific steps of the IGA are as follows.

(1) Initialize the population and encode: the population is generated by randomization, and the individuals in the population consist of chromosomes, each of which is a set of test papers. The chromosomes are combinations of test questions. The segmented real-number encoding method is used. Each segment of the gene denotes a question type, and each gene denotes the serial number of that question in the test question database. For the university English test, suppose it contains four types of questions: multiple-choice, fill-in-the-blank, translation, and essay. The corresponding number of test questions for these types is 10, 5, 4, and 1, respectively. In this case, the code is:

1 4 66 85 75 74 112 546 157 10 ||201 3 35 524 168||213 451 221 15||842

(2) Determine the fitness function: $f(x) = \sum_{i=1}^{n} f_i(x)^2 \omega_i$, where $f_i(x)$ is the error between the actual value of the constraint and the target value, and ω_i is the weight of different constraints. The sum of the weights of all constraints is 1.

(3) Genetic operators: they are used to evolve the population to obtain new individuals, and the details are as follows.

① Selection operator: the best individuals are selected from the population to enter the next generation. The roulette strategy is applied to perform the selection operation. The probability of a test paper being selected is: $p_i = f_i / \sum_i^m f_i$, where m is the population size and f_i is the fitness value of test paper i.

② Crossover operator: through the genetic recombination of chromosomes, new individuals are generated. The segmented single-point crossover is used, and the value of crossover probability p_c is improved using the following equation:

$$\mathrm{p_c} = \begin{cases} 0.9 - \dfrac{0.3 \times (\mathrm{f}' - \mathrm{f_{avg}})}{\mathrm{f_{max}} - \mathrm{f_{avg}}}, & \mathrm{f}' \geq \mathrm{f_{avg}} \\ 0.9, & \mathrm{f}' < \mathrm{f_{avg}} \end{cases}, \tag{2}$$

where f' is the fitness value of the larger one among the two individuals to be crossed and f_{max} and f_{avg} are the maximum and average values of the population fitness values.

③ Mutation operator: according to the probability of mutation p_m, genes at some positions in the chromosome are changed.

The segmented single-point mutation is used, and the value of p_m is improved using the formula:

$$\mathrm{p_m} = \begin{cases} 0.1 - \dfrac{0.099 \times (\mathrm{f_{max}} - \mathrm{f})}{\mathrm{f_{max}} - \mathrm{f_{avg}}}, & \mathrm{f} \geq \mathrm{f_{avg}} \\ 0.1, & \mathrm{f} < \mathrm{f_{avg}} \end{cases}. \tag{3}$$

④ The test paper generation result is output when the preset number of iterations is reached.

2.3 Automatic Scoring Algorithm for Chinese to English Translation Questions

Chinese to English translation is a subjective question that serves as an important way to test students' English ability. Its scoring mainly involves analyzing the semantic similarity between students' translations and the standard translations; therefore, students' Chinese to English translation questions are primarily scored automatically based on grammar and semantics. The corresponding formula is:

$$S(T) = 0.25G(T) + 0.75Y(T) \tag{4}$$

where $G(T)$ is a score for grammar, which can reflect the grammatical errors in the student's translation. Its formula is:

$$G(T) = 1 - \frac{M}{N} \tag{5}$$

where M is the total number of sentences with grammatical errors in the students' translations, N is the total number of sentences, and $M \geq N$.

$Y(T)$ is a score for semantics, and its formula is:

$$Y(T, T') = \frac{1}{n} \sum_{i=1}^{n} sim(T_{s_i}, T'_{s_i}), \tag{6}$$

where $\mathrm{sim}(\mathrm{T_{s_i}}, \mathrm{T'_{s_i}})$ represents the cosine similarity between the vector of sentence i in the students' translation and the vector of sentence i in the standard translation.

3 Results and Analysis

Experiments were conducted in the MATLAB environment. Firstly, the automatic test paper generation method designed in this paper was analyzed,

and four question types were set according to the university English test. There were ten multiple-choice questions, two points each question, ten fill-in-the-blank questions, two points each question, two translation questions, 15 points each question, and an essay with a score of 30 points. The total score was 100 points.

The standard question bank used in the experiment consisted of a total of 5000 test questions, which include 1,500 multiple-choice questions and 1,500 fill-in-the-blank questions, as well as 1000 translation questions and 1,000 essay questions. The difficulty of the test paper was set at 0.65, and the degree of differentiation was 0.45. It covered all chapters of the university English course and had an examination time of 120 min. The population size of the IGA was 100, the initial values of p_c and p_m were 0.65 and 0.01, and the maximum number of iterations was 300. The IGA was compared with the traditional GA and particle swarm optimization (PSO) algorithm [10] by conducting the test paper generation experiment 200 times.

Ten out of the 200 experiments were randomly selected to compare the test paper generation times, and the results are presented in Figure 1.

It was observed in Figure 1 that the test paper generation time of the traditional GA was significantly longer. In the ten experiments extracted, the traditional GA took around 60 s to generate a test paper, while the PSO algorithm took 45–50 s, slightly less than the traditional GA. On the other hand, the IGA only required between 20 s and 30 s for generation. Calculations showed that on average, the traditional GA took 62.5 s to generate a test paper, whereas the PSO algorithm only needed 46 s and the IGA merely required

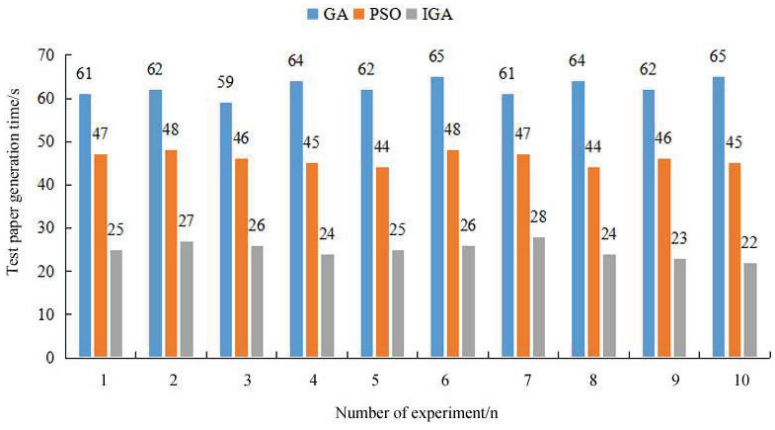

Figure 1 Comparison of the test paper generation time.

Table 2 Comparison of paper generation success rate

	GA	PSO	IGA
Times of successful generation/n	67	75	94
Times of failed generation/n	33	25	6
Success rate of paper generation/%	67	75	94

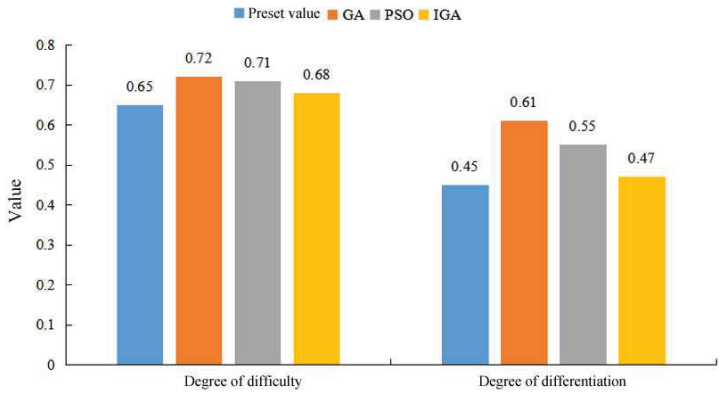

Figure 2 Comparison of test paper generation validity.

25 s – significantly shorter than both alternatives. This suggested that the IGA was more efficient.

The success rate of paper generation was compared between different algorithms in 100 experiments. The results are presented in Table 2.

From Table 2, it can be observed that among the 100 experiments conducted, the GA had the highest times of failed generation, 33, and a success rate of only 67%. The PSO algorithm had slightly fewer failures, 25, resulting in a success rate of 75%, an increase of 8% compared to the GA. However, the IGA achieved only six failures and a remarkable success rate of 94%, surpassing the PSO algorithm by an additional 19%. This further confirmed the feasibility and effectiveness of the designed approach for paper generation.

Then, the validity of the test paper generation was analyzed by comparing the degrees of difficulty and differentiation of the papers obtained by the three methods. The degrees of difficulty and differentiation of the 200 experiments were averaged, and the results are shown in Figure 2.

It was seen from Figure 2 that the preset difficulty degree of the test paper was 0.65, the average difficulty degree of the test paper obtained by the GA was 0.72, which was obviously higher than the preset value. Similarly, the

Table 3 Comparison of scoring results

	1	2	3	4	5	6	7	8	9	10
Manual scoring	14	13	9	14	14	13	10	11	14	15
Manual scoring	13.76	12.52	10.12	13.69	14.21	12.57	11.03	10.67	14.35	14.59
Error	0.24	0.48	1.12	0.31	0.21	0.43	1.03	0.33	0.35	0.41

Table 4 Comparison of score similarity

Number of translation questions	10	50	100	200
The value of r	0.87	0.88	0.88	0.89

PSO algorithm yielded an average difficulty degree of 0.71, surpassing the preset value by 0.06. On the other hand, the IGA produced a test paper with an average difficulty degree of 0.68, deviating from the preset value by only 0.03. The differentiation degree in the test paper obtained by the GA was 0.61, which was 0.16 higher than the preset value. The differentiation degree in the paper obtained by the PSO algorithm was 0.55, which was 0.1 higher than the preset value, and the differentiation degree in the test paper obtained by the IGA was 0.47, which was just 0.02 higher than the preset value. It was found that the test paper obtained by the IGA had higher quality and was closer to the requirement, more reasonable, and more effective, thus it can be used in practical exams.

Then, the automatic scoring algorithm for Chinese to English translation questions was analyzed. The scoring effect was evaluated using the Pearson correlation coefficient [11] (r value). The larger the value of r was, the better the automatic scoring effect was.

The scores of the ten translation questions (total 15 points) were taken as examples. The results of automatic and manual scoring are shown in Table 3.

It was seen from Table 3 that the difference between the scoring results was small, with a maximum error of 1.12 and a minimum error of 0.21, indicating that using the automatic scoring algorithm could yield similar results to the manual scoring. To further assess the reliability of automatic scoring, the similarity between automatic scoring and manual scoring was compared by taking 10, 50, 100, and 200 translation questions as examples. The results are shown in Table 4.

It was seen from Table 4 that there was a high similarity between the result obtained from the automatic scoring method and the manual scoring result as all values of r were above 0.8. This meant that the automatic scoring closely aligned the manual scoring, thus making it suitable for automated checking of test papers in university English exams.

4 Conclusion

This paper designed an IGA-based automatic test paper generation method and an automatic scoring method for translation questions. Through experiments, it was found that compared with the traditional GA and the PSO algorithm, the IGA-based method generated test papers faster, with a higher success rate and higher quality. Additionally, the results obtained from the automatic scoring method closely matched the manual scoring results, demonstrating the effectiveness of the IGA. Therefore, the IGA can be applied in actual examinations to enhance their level of informatization and automation.

References

[1] A. Solyman, Z. Wang, Q. Tao, A. A. M. Elhag, M. Toseef, Z. Aleibeid, 'Synthetic data with neural machine translation for automatic correction in arabic grammar', Egypt. Inform. J., 22(3), pp. 303–315, 2021.

[2] S. Wang, N. Shrestha, A. K. Subburaman, J. Wang, M. Wei, N. Nagappan, 'Automatic Unit Test Generation for Machine Learning Libraries: How Far Are We?', 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), pp. 1548–1560, 2021.

[3] B. Yang, H. Xie, K. Ye, H. Qin, R. Zu, A. Liu, 'Analysis of intelligent test paper generation method for online examination based on UML and particle swarm optimisation', Int. J. Inform. Commun. Technol., 18(3), pp. 317–333, 2021.

[4] D. Wang, Y. Zhao, H. Lin, X. Zuo, Automatic scoring of Chinese fill-in-the-blank questions based on improved P-means, J. Intell. Fuzzy Syst., 40(3), pp. 5473–5482, 2021.

[5] Z. Yuan, 'Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm', J. Intell. Fuzzy Syst., 40(2), pp. 2069–2081, 2021.

[6] Q. F. Zhao, Y. F. Li, 'Research and development of online examination system', Adv. Mater. Res., 756–759, pp. 1110–1113, 2013.

[7] M. Yağci, M. Üünal, 'Designing and Implementing an Adaptive Online Examination System', Proc. Soc. Behav. Sci., 116:3079–3083, 2014.

[8] B. Sadayapillai, K. Kottursamy, 'A blockchain-based framework for transparent, secure, and verifiable online examination system', J. Uncertain Syst., 15(03), 2022.

[9] A. Zemliak, 'A modified genetic algorithm for system optimization', COMPEL, 41(1), pp. 499–516, 2022.
[10] H. Yu, M. Zheng, W. Zhang, W. Nie, T. Bian, 'Optimal design of helical flute of irregular tooth end milling cutter based on particle swarm optimization algorithm', P. I. Mech. Eng. C. J. MEC, 236(7), 3323–3339, 2022.
[11] S. Vaziri, J. Abbatematteo, M. Fleisher, A. B. Dru, D. T. Lockney, P. S. Kubilis, D. J. Hoh, 'Correlation of perioperative risk scores with hospital costs in neurosurgical patients', J. Neurosurg., 132(3), pp. 818–824, 2019.

Biography

Han Yang, born in 1982, has received the master's degree of education from He'nan Normal University in December 2014. She is a lecturer and is working in Anyang Vocational and Technical College. She is interested in English language teaching.