The Application of Artificial Intelligence in the Design of English Vocabulary Query System

Mingqi Wang[1] and Jia Zhao[2,*]

[1]Department of Foreign Language, Qinhuangdao Vocational and Technical College, Qinhuangdao, Hebei 066100, China, wangmingqi@qvc.edu.cn
[2]Department of Foreign Language, Qinhuangdao Vocational and Technical College, Qinhuangdao, Hebei 066100, China, zhaojia@qvc.edu.cn

Corresponding author: Jia Zhao, zhaojia@qvc.edu.cn

Abstract. It is excluded. At the same time, try not to involve Java script; realize the normal operation of Struts application through steps such as copying files and adding elements, so as to ensure that the system can conduct English vocabulary query smoothly. Experiments show that the designed system can accurately query English vocabulary, and has the advantages of high efficiency and strong practicability. The vocabulary learning resources are divided into thirteen categories. Finally, according to the English level 4 vocabulary resources According to the construction principle of the library, the vocabulary resource library and the test question resource library are constructed. The rational design of the vocabulary resource is the basis for personalized vocabulary recommendation.

Keywords: Computer-Aided System; English Vocabulary Query; Page Cumbersomeness; Struts; Query Module; JSP Page.
DOI: https://doi.org/10.14733/cadaps.2023.S5.146-156

1 INTRODUCTION

As China blends with global cultures, public interest in learning English is growing. It is imperative to make full use of information technology to improve the English level of the people and to assist language learning through software systems. According to statistics, 93% of Chinese students have studied English, and only 3% of them can read English periodicals and books without barriers [1]. Their difficulties are mainly caused by insufficient vocabulary. Chinese students' learning habits of English words are often "meaning-driven", and learners usually only expand their vocabulary by memorizing word forms and Chinese explanations, ignoring the understanding of word application scenarios, and ignoring the development of vocabulary knowledge to depth. The current English vocabulary learning methods mainly include: memorizing vocabulary books, computer software learning and mobile APP learning. With the rapid development of mobile Internet and the popularization of smart devices, it is rare to recite words while holding English books or word books. Learning English vocabulary using mobile phones or computer software has become the main channel for learning English. Through vocabulary memory software, learners flexibly use fragmented time to memorize words through smart devices such as computers or mobile phones, which not only improves vocabulary but also increases their interest in learning [2].

Fendji et al. [3] believes that learning English scientifically and mastering vocabulary should fully consider the following points: First, language is the carrier of culture. Different cultural forms of different nationalities are embodied in language, including value orientation, artistic achievements and living customs. Ahn and Smagulova [4] believe that they have not learned articles in pure foreign languages. Context learning can't really understand foreign culture. He [5] believes that words are not only memory, but also the number of words. To really understand them, you have to put the words in the context. This requires learners to learn and consolidate words while reading. Leona et al. [6] prove that vocabulary learned in context is more unforgettable than simple memory. The improvement of English vocabulary and reading ability goes hand in hand. However, if you search for words too frequently in the reading process, and if the interpretation of the dictionary you search does not match the meaning of the words in the original text, it will not only affect the reading speed, but also affect the reading comprehension and mastering of new words. Vocabulary learning resources are divided into 13 categories. Liu et al. [7] according to the construction principles of the vocabulary bank of CET-4, it has constructed a vocabulary resource bank and a test question resource bank. The rational design of vocabulary resources is the basis of personalized vocabulary recommendation. Second, word memory should rely on scientific memory methods.

Learning requires repeated review and study, so that memory is not easy to forget. If there is no scientific memory method, it is easy to give up learning, and it is difficult to master the knowledge learned for a long time. This is why the word "software" in the market is mainly used for sudden learning and rarely used by everyone for a long time. In addition, Luo et al. [8] experiment results show that paying attention to high-frequency words can improve students' reading comprehension. In the design of students' reading materials, increasing attention to word frequency can improve students' reading comprehension. Finally, the comprehensive ability of English is reflected in all aspects of listening, speaking, reading and writing. Words learned only for preparation are not practical in actual learning and work. This requires that the source of memorized words is high-frequency words.

2 RELATED CONCEPTS AND THEORETICAL BASIS

2.1 A Computer-Aided English Vocabulary Query System Based on Struts

The approach, thereby updating and progressing tag li-brary technology is an important part of Struts framework structure [9].

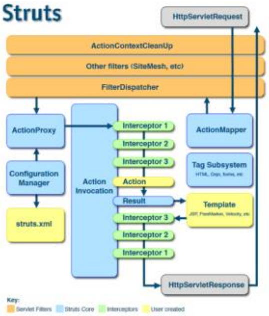

Figure 1: Struts frame structure.

As can be seen from the Struts framework in Figure 1:

1) Researchers have not reached a unified and uncontroversial consensus. Fundamentally, system of vocabulary knowledge is too complex. However, by reviewing relevant literature, it can be seen that, in general, researchers tend to the point of view

2) It is that lexical knowledge is not single-dimensional, but complex and multi-dimensional. Vocabulary knowledge should include at least two dimensions of quality and quantity. The quality and quantity of lexical knowledge represent lexical depth knowledge and lexical

3) an object of type Servlet Students take the test after each unit. If the answer is correct, they can continue to study the following content; if the answer is wrong, enter the branch program. After the students have mastered the content of the branch program [10]

2.2 Query Module Design

The user uses communicate with each other and reflecting the changes in the cooperation or sequence diagrams. The query process is shown in Figure 2.

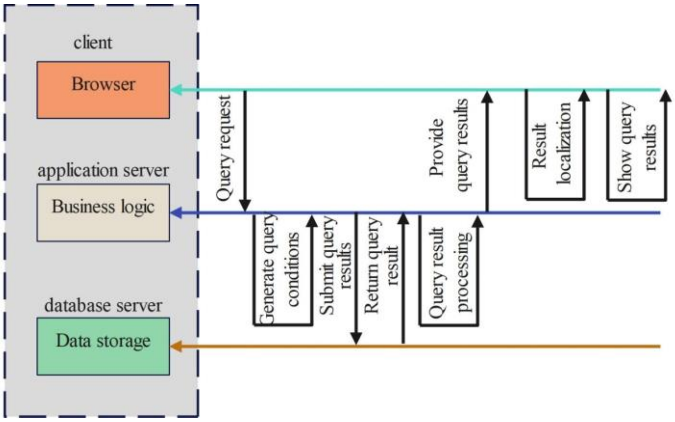

Figure 2: Sequence diagram of the query process.

The steps for query result can be recorded in the client SQLite database for multiple use in the future; the browser presents the query concisely or in detail The result is viewed by the user.

The vocabulary query system in this article by transforming the construction method. The system includes main function interfaces such as vocabulary category maintenance and vocabulary quick query, and auxiliary function interfaces.

Since the system of this article is operation of the Struts application is the key to ensure that the system of this article can query English vocabulary smoothly. jar and lib/struts *. Tld; WEB-INF/web in the web application directory. It will be added to the xml file respectively to define the controller servlet and set the URI (a string that identifies and locates any resource) to obtain interaction with the servlet; for implementing the tag library definition can Add to the JSP page using the create the WEB-INF/action.xml file; When writing the program, add the struts.jar file to the CLASS-PATH.

2.3 Theory of Program Teaching

The origin of program teaching theory can be traced back to the advent of the automatic teaching machine designed by American psychologist Plessy in 1924. However, the formation of its theoretical system and its wide application in teaching should be attributed to Skinner's work in the 1950s. A lot of work over the years. The English word-assisted learning system is designed according to Skinner's program teaching theory, so this article will focus on introducing Skinner's thought about program teaching. The vocabulary learning resources are divided into thirteen categories. Finally, according to the English level 4 vocabulary resources According to the construction principle of the library, the vocabulary resource library and the test question resource library are constructed. The rational design of the vocabulary resource is the basis for personalized vocabulary recommendation.

Learning is a behavior in which the response the syllabus, and teachers play the role of supervisors or middlemen. He believes that teaching should pay attention to operating conditions and reinforcement. Should be through a variety of stimulating conditions, to mobilize students' enthusiasm for learning, to strengthen students' learning effect.

The Skinner disc machine is a typical Skinner teaching machine. This is an infill-type teaching machine. The machine presents the question through a small window, and the student writes the answer to the question on a piece of paper at the other end of the machine. If the answer is correct, the student uses the lever to make a hole in the tape so that the question is not presented again during the next study session. When the lever is pulled back to its original position, a new frame appears. In this cycle, learn.

Skinner's application of teaching machines for learning has many advantages: teaching machines can give correct answers in time, provide students with timely feedback information, and shorten the feedback time in traditional classroom teaching; teaching machines divide teaching materials into small units. It is convenient for students to learn and enhance their self-confidence; students can determine the learning speed according to their own level; use both hands and brains during learning, which can cultivate students' self-learning ability.

Program teaching methods are generally divided into two modes: linear and branched. Skinner proposed a straight-line teaching model, that is, after students answer the first question, regardless of whether the answer is correct, the machine will then present the correct answer, and then move on to the next question. The branch teaching model was proposed by the American psychologist Claude. Students take the test after each unit. If the answer is correct, they can continue to study the following content; if the answer is wrong, enter the branch program. After the students have mastered the content of the branch program, they return to the main program and continue to learn new content.

2.4 Knowledge of English Vocabulary

Vocabulary knowledge, also called vocabulary ability by many foreign researchers, is the basis of language learning. The decisive role of vocabulary is self-evident, whether it is for foreign language exams large or small, or for communicative applications in daily life.

Regarding the definition of lexical knowledge, researchers in the field of second language vocabulary have proposed many theoretical frameworks, each with its own emphasis. So far, researchers have not reached a unified and uncontroversial consensus. Fundamentally, system of vocabulary knowledge is too complex. However, by reviewing relevant literature, it can be seen that, in general, researchers tend to the point of view that lexical knowledge is not single-dimensional, but complex and multi-dimensional. Vocabulary knowledge should include at least two dimensions of quality and quantity. The quality and quantity of lexical knowledge represent lexical depth knowledge and lexical breadth knowledge respectively. This definition method has a great influence in the academic circle, and it is also the two most basic dimensions adopted by lexical knowledge.

3 RELATED TECHNOLOGIES

3.1 Vocabulary Ability Estimation Techniques

At the beginning of the test, there is no information about the subject's ability in the system. The first question is generally a random test. If the subject answers correctly, the difficulty of the next question will increase. On the contrary, the difficulty of the next question will be increased. When the answer of the subject is right or wrong, the initial value □0 of the subject's ability can be roughly estimated, and the calculation method of the initial value is shown in formula (1).

$$\theta_0 = \ln \frac{x}{m-x} \tag{1}$$

The value is the initial ability value of the subject.

For nonlinear equations, there is no formula for finding the root, and it is very difficult to solve the exact root of the equation, so how to find the root that is most similar to the exact root is particularly important. Whether it is a nonlinear equation or a system of nonlinear equations, its solution can be obtained by the Newton-Raphson method, which is an iterative method, also known as the tangent method or the Newton method, which is an important method for solving the roots of nonlinear equations. The biggest advantage of this method is that it can converge squarely near the root.

That is, transforming complex nonlinear equations f(x)=0 into simple linear equations to solve. Let x0 be an approximate root of the nonlinear equation f(x)=0, the exact value of the root can be written as x=x0+1□, so the equation becomes f(x0+1)=0.

$$f(x) = \sum_{k=1}^{n} \left(x - x_0\right)^k \frac{f^{(k)}\left(x_0\right)}{k!} \tag{2}$$

Formula (2) is Taylor's formula, which is expanded into Taylor polynomial near x0, and formula (3) is obtained.

$$f\left(x_0 + \delta_1\right) = f\left(x_0\right) + \delta_1 f'\left(x_0\right) + \frac{\delta^2}{2} f''\left(x_0\right) \tag{3}$$

Where δ1= x-x0 , if δ is very small, the higher-order term 2 can be omitted to obtain formula (4).

$$f\left(x_0 + \delta_1\right) \approx f\left(x_0\right) + \delta_1 f'\left(x_0\right) \approx 0 \tag{4}$$

Therefore, the calculation formula (5) of the primary correction value x1 of the root of the equation is obtained. The teaching methods are generally divided into two modes: linear and branched. Skinner proposed a straight-line teaching model, that is, after students answer the first question, regardless of whether the answer is correct, the machine will then present the correct answer, and then move on to the next question. The branch teaching model was proposed by the American psychologist Claude. Students take the test after each unit. If the answer is correct

$$x_1 = x_0 + \theta_1 = x_0 - \frac{f\left(x_0\right)}{f'\left(x_0\right)} \tag{5}$$

And so on, it can be obtained that its approximate value is

$$x_2 = x_1 + \theta_2 = x_1 - \frac{f\left(x_1\right)}{f'\left(x_1\right)} \tag{6}$$

Repeat the above process to obtain a sequence of points, x0, x1, x2, ..., xn. During this iterative process, the value of the obtained approximate root can be continuously approached to the exact root. When the accuracy requirement is satisfied Stop the iteration, then the root of the nonlinear equation is found.

3.2 Bayesian Formula Derivation

The core of the Bayesian method is the Bayes theorem, which is about probabilistic reasoning. The problem of probabilistic reasoning is to estimate the probability of the conclusion when inferring or making decisions based on uncertain information. In probability, the research object of probabilistic reasoning is the formula for calculating objective probability. The probabilistic reasoning method revealed by the Bayesian formula for guiding people to study the process of understanding and processing, and to make effective decisions.

The expression of Bayes' theorem is a formula as shown in the following formula.

$$P(B|A) = P(A|B)P(B)/P(A) \tag{7}$$

Replace B with the subject's potential trait θ, and replace A with the subject's response to the test item, so formula (8) can be transformed as follows.

$$P(\theta|u) = P(u|\theta)P(\theta)/P(u) \tag{8}$$

In order to avoid confusion with item response function, in this study, P in equation (9) is replaced by f.

$$f(\theta|u) = f(u|\theta)f(\theta)/f(u) \tag{9}$$

Since θ is a continuous variable, f(θ|u) and f(θ) are both density functions, it is a constant for a given set of responses, so equation (10) can be expressed by the following equation.

$$f(\theta|u) \propto L(u|\theta)f(\theta) \tag{10}$$

For this relationship, the parameters can be estimated by steps similar to the maximum likelihood method, that is, a set of parameter values is required to make f(θ1, θ2, …, θN|u) reach the maximum value, because The likelihood function and its natural logarithm function reach the maximum value at the same time, so take the natural logarithm on both sides of formula (10) to obtain formula (11).

$$\ln f(\theta|u) = \ln L\left(u|\theta_1, \theta_2, \ldots, \theta_N\right) \tag{11}$$

Therefore, as long as the values of m and n can be determined, the equivalence between the two tests can be achieved. There are many ways to determine the equivalent constant, such as the project characteristic curve method, the regression equation method, the mean results obtained by using these methods, so the mean standard deviation method is selected in this study.

3.3 Project Parameter Estimation and Equivalent

The software version selected in this study is BILOGMG 3.0. After the fit test of the data and the model, it is known that the test data this time fits the two-parameter model better. The Bayesian expected posterior estimation method is used, the Newton-Raphson cycles is 2. By estimating the parameters of the test items according to the participants' answers, the discrimination parameter a and the difficulty parameter b of all test questions can be obtained. By checking whether the chi-square value of each question reaches a significant level, the questions that do not fit the two-parameter model are eliminated from the final deep knowledge question bank.

At this time, the last step, and also the most critical step, is to put the test questions into the database, and the equivalence of the test questions parameters is carried out. Because of the anchor test design in advance, the four sets of depth test papers involved in this study can be equivalent with the help of this design. The so-called equivalence refers to the project parameters estimated from two sets of different test data to determine the conversion relationship between the two sets of project parameters. After equivalence transformation, the item parameters of the two sets of tests can be placed on the same scale for comparison. In the anchor test design, the parameter estimates for the two sets of tests are linearly related, which exists for the two-parameter logistic model:

$$b_Y = \mathrm{m}b_X + \mathrm{n} \tag{12}$$

Equation 12 expresses the criterion for the equivalent conversion from test X to test Y. In this case, test Y is used as the reference unit, that is, test Y is the "scale test" and test X is the "original test". bX and bY represent the difficulty parameters of the test questions in test X and test Y respectively, aX and aY represent the distinguishing degree parameters of test questions in test X and test Y, m and n are conversion coefficients, also known as equivalent constants.

According to IRT, Equation 13 is established, that is, in the anchor test, the estimated value of the difficulty parameter obtained in the two tests for the same anchor question has this linear relationship, and it is easy to infer:

$$\bar{b}_Y = \mathrm{m}\bar{b}_X + \mathrm{n} \tag{13}$$

Among them, b¯X and SX represent the mean and standard deviation of the difficulty of all anchor questions in test X, respectively, and b¯Y and SY respectively represent the mean and standard deviation of the difficulty of all anchor questions in test Y. Since the values of the project parameters have been estimated in advance before the equivalence, b¯X , SX , b¯Y , SY can be easily obtained, then the unknowns in Equation 12 and Equation 13 are only m and n. After the deformation of these two formulas, it can be known that:

$$\mathrm{m} = \frac{S_Y}{S_X} \tag{14}$$

$$\mathrm{n} = \bar{b}_Y - \mathrm{m}\bar{b}_X \tag{15}$$

4 EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Accuracy Test and Comparison

As can be seen from Figure 3, in the selected English academic text corpus, the accuracy of the PLS-RF-Softmax highest, especially in the linguistics term corpus, by evaluating tense Adjustment, the overall accuracy of tense reached 98.9984%, which is very impressive. As the number of operation words increases, the accuracy of the PLS-RF-Softmax algorithm does not show a sharp decline, indicating that it has strong fitting and generalization capabilities. This is because the PLS-RF-softmax algorithm can not only effectively deal with the multicollinearity problem, but also can model under the condition that the sample size does not exceed the variables, and establish a discriminant model for evaluating the correctness of tense in English.

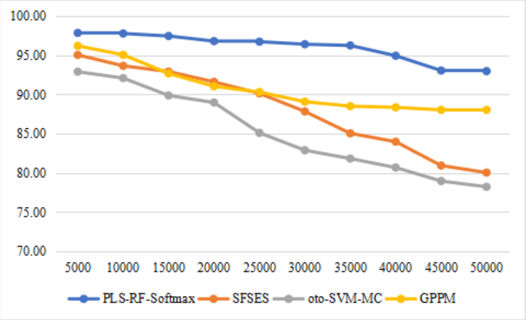

Figure 3: Accuracy table of four algorithms in English lexical text corpus.

As can be seen from Figure 4, the analysis time of the PLS-RF-Softmax algorithm is less than 100ms in each term corpus, and greater than 100ms only in the linguistic term. The effectiveness of this method is illustrated. The analysis time of the SFSES algorithm is almost always the highest, even exceeding 300ms in chemistry academic papers. The SFSES algorithm performs well in economics academic papers. In terms of analysis time, OTO-SVM-MC and GPPM perform generally. The algorithm in this paper significantly improves the interpretation degree of the model and the expression of nonlinear structure with extremely low analysis time. This is a necessary condition for English tense adjustment for human-computer interaction (HCI).

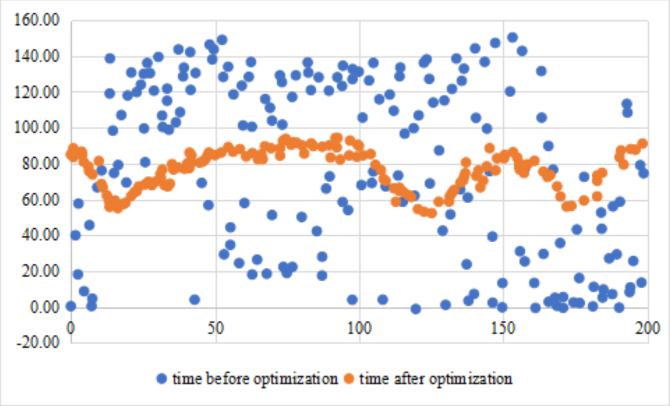

Figure 4: Analysis timetable of the four algorithms in the English lexical text corpus.

4.2 Comparison and Analysis of Multiple Algorithms

There are 3 ways in the paper:

CNN. By estimating the parameters of the test items according to the participants' answers, the discrimination parameter a and the difficulty parameter b of all test questions can be obtained.

SVM. the equivalence of the test questions parameters is carried out. Because of the anchor test design in advance, the four sets of depth test papers involved in this study can be equivalent with the help of this design.

AT-LSTM: The so-called equivalence refers to the project parameters estimated from two sets of different test data to determine the conversion relationship between the two sets of project parameters.

This paper after equivalence transformation, the item parameters of the two sets of tests can be placed on the same scale for comparison in Figure 5.

It compared followed by AT-LSTM, followed by CNN and SVM algorithms. Such teaching methods are generally divided into two modes: linear and branched. Skinner proposed a straight-line teaching model, that is, after students answer the first question, regardless of whether the answer is correct, the machine will then present the correct answer, and then move on to the next question. The branch teaching model was proposed by the American psychologist Claude. Students take the test after each unit.

Figure 6 shows the training time of each category of algorithm corpus The training efficiency of the algorithm in this paper is much lower than that of CNN and SVM, in terms of significance, compared with the other three Algorithms are fine.

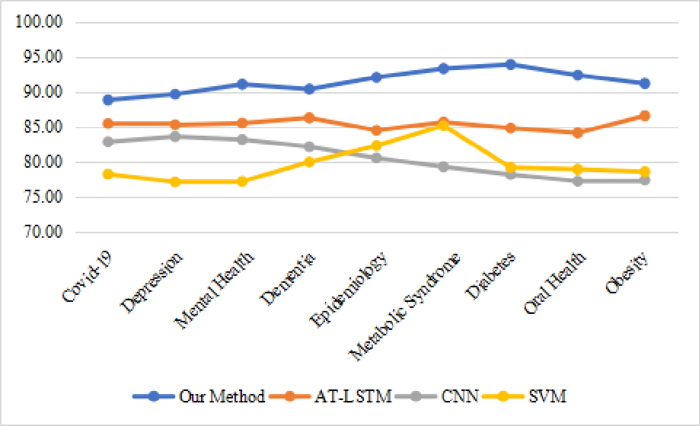

Figure 5: Accuracy comparison of the four algorithms.

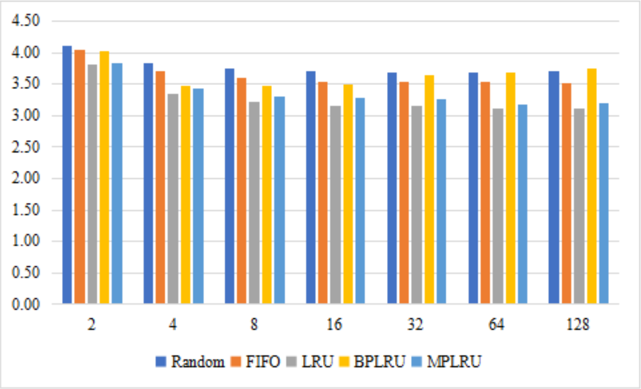

Figure 6: Training time chart of each category of algorithm corpus.

As shown in Figure 7, 57.69% of the learners and 25% of the learners believe that the system is a better tool for assisting vocabulary learning to varying degrees, and 11.54% of the learners think that the system is not very useful for vocabulary learning. To a certain extent, the system can help learners to learn vocabulary, which means that the system formulates a personalized vocabulary learning sequence for learners, which has a positive effect on learners' learning of CET-4 vocabulary. 63.46% of the learners agree that the system can improve the learners' interest in learning English vocabulary, 21.15% the interest in vocabulary learning, and a small number of learners think that the system cannot help learners to increase their interest in learning. 61.54% of the learners agree that the system can improve the enthusiasm of learning vocabulary, 21.15% the enthusiasm of learning vocabulary, a few learners think that the system can improve the enthusiasm of vocabulary learning is not high. Learning English vocabulary has a good positive effect.

5 CONCLUSION

It is mainly used by students and expects to improve the efficiency of students' English vocabulary learning. Combined with the characteristics of vocabulary resources and the learning characteristics of college students, the vocabulary learning resources are reasonably classified and organized.

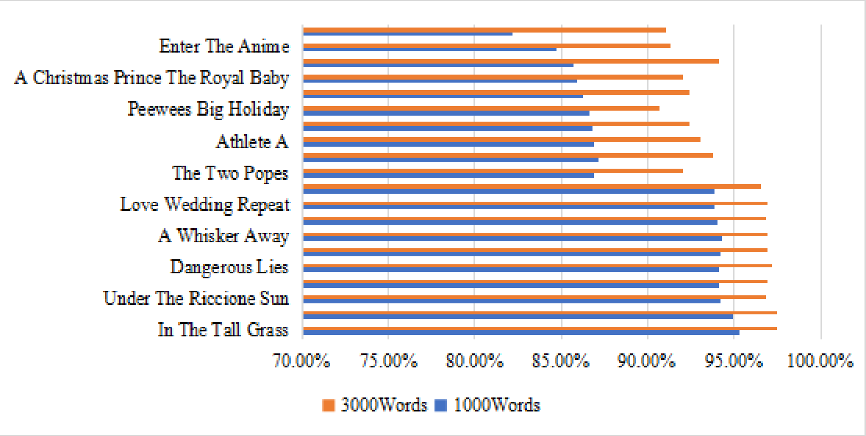

Figure 7: Study Attitude Questionnaire.

Combined with the subject characteristics of college students, the vocabulary learning resources are divided into thirteen categories. The machine will then present the correct answer, and then move on to the next question. The branch teaching model was proposed by the American psychologist Claude. Students take the test after each unit. If the answer is correct, according to the English level 4 vocabulary resources and the construction principle of the library, the vocabulary resource library and the test question resource library are constructed. The rational design of the vocabulary resource is the basis for personalized vocabulary recommendation.

Mingqi Wang, https://orcid.org/0000-0002-8923-6460
Jia Zhao, https://orcid.org/0000-0001-6328-6884

REFERENCES

[1] Ahn, E.-S.; Smagulova, J.: English language choices in Kazakhstan and Kyrgyzstan, World Englishes, 41(1), 2022, 9-23. https://doi.org/10.1111/weng.12552

[2] Cope, B.; Kalantzis, M.: Artificial intelligence in the long view: from mechanical intelligence to cyber-social systems, Discover Artificial Intelligence, 2(1), 2022, 1-18. https://doi.org/10.1007/s44163-022-00029-1

[3] Cope, B.; Kalantzis, M.; Searsmith, D.: Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies, Educational Philosophy and Theory, 53(12), 2021, 1229-1245. https://doi.org/10.1080/00131857.2020.1728732

[4] Fendji, J.-L.-K.-E.; Tala, D.-C.; Yenke, B.-O.; Atemkeng, M.: Automatic Speech Recognition using limited vocabulary: A survey, Applied Artificial Intelligence, 36(1), 2022, 2095039. http://doi.org/10.48550/arXiv.2108.10254

[5] He, B.: Application of Mobile Technology in College English Vocabulary Teaching, Journal of Mathematics, 5(52), 2022, 281-282. https://doi.org/10.1155/2022/9009008

[6] Leona, N.-L.; Koert, M.; Molen, M.: Explaining individual differences in young English language learners' vocabulary knowledge: The role of Extramural English Exposure and motivation - ScienceDirect, System, 96(66), 2020, 157-168. https://doi.org/10.1016/j.system.2020.102402

[7] Liu, C.: Corpus Design of Chinese Medicine English Vocabulary Translation Teaching System Based on Python, Journal of Information & Knowledge Management, 21(2), 2021, 54-57. https://doi.org/10.1142/S0219649222400226

[8] Luo, Y.; Wei, W.; Ying, Z.: Artificial Intelligence-Generated and Human Expert-Designed Vocabulary Tests: A Comparative Study, SAGE Open, 12(1), 2022, 1-23. https://doi.org/10.1177/21582440221082130

[9] Read, K.; Contreras, P.-D.; Rodriguez, B.: Read conmigo: The effect of code-switching storybooks on dual-language learners' retention of new vocabulary, Early education and development, 32(4), 2021, 516-533. https://doi.org/10.1080/10409289.2020.1780090

[10] Rosenberg, J.-M.; Krist, C.: Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations, Journal of Science Education and Technology, 30(2), 2021, 255-267. https://doi.org/10.1007/s10956-020-09862-4