Recommendation for English Multiple-choice Cloze Questions Based on Expected Test Scores

Tomoharu Iwata
NTT Communication Science Laboratories
iwata@cslab.kecl.ntt.co.jp

Tomoko Kojiri
Graduate School of Information Science, Nagoya University
kojiri@is.nagoya-u.ac.jp

Takeshi Yamada
NTT Communication Science Laboratories

Toyohide Watanabe
Graduate School of Information Science, Nagoya University

Abstract

When students study for multiple-choice cloze tests as the Test of English for International Communication (TOEIC), they tend to repeatedly tackle questions of the same type. In such situations, students can effectively solve questions related to their incorrectly answered questions. However, since they need several different kinds of knowledge and a large vocabulary to derive answers, it is inappropriate to statically define the relations among questions from various viewpoints beforehand. In this paper, we propose a recommendation algorithm for English multiple-choice cloze questions that maximize students' expected improvements of test scores based on the learning log data of other students. Effective questions may be identical for most students who incorrectly answered the same questions. Therefore, in our approach, relations among questions in tests and questions studied during tests are determined based on the change from incorrect to the correct answers of the test questions. Questions that maximize the expected test scores, which are calculated based on the input test scores using regression models, are recommended for future students. Based on this method, students can acquire higher test scores with better learning efficiency. Experimental results show that our method yields major improvements in performance compared with random material recommendation method.

1 Introduction

As network and database technologies continue to grow rapidly, e-Learning systems have become widely used in various domains. English multiple-choice cloze questions are one popular learning material in e-Learning. Such questions are applied in various tests, namely, the Test of English for International Communication (TOEIC) or university entrance examinations in Japan. A large number of such questions are already stored in e-Learning systems, and automatic question generation systems have also been developed to generate new questions [14].

In e-Learning systems, providing suitable learning materials that teach unacquired knowledge is crucial. In the e-Learning of English multiple-choice cloze questions, students tend to repeatedly tackle the same type of questions to heuristically acquire the knowledge asked in the questions. Since students need various kinds of

English knowledge to solve questions, such as grammar and a large vocabulary, determining the important knowledge for answering each question is difficult. Moreover, because there are many questions, it is impractical to statically define the relations among all questions beforehand.

This research assumes students who are studying to acquire English knowledge within the given tests for English multiple-choice cloze questions. They take English multiple-choice cloze tests that contain a specific number of questions and study using the same type of questions to acquire the knowledge of the incorrectly answered questions. The objective of our research is to provide effective questions in the studying phase to increase the scores of subsequent tests taken after the studying phase. Questions can be selected in various ways based on such supporting policies as giving many questions so that students deeply understand the unacquired knowledge and giving a minimum number of questions so that they can quickly pass the test. Our research focuses on increasing test scores with great efficiency by giving a small set of questions that are adequate for gaining good test scores.

Question answers reflect the acquired/unacquired knowledge of students. If students answered incorrectly the same question, their unacquired knowledge probably is the same. Therefore, questions that help students grasp the knowledge of questions may also be effective for other students who failed the same question. To provide appropriate questions based on the tendency of student understandings, this research proposes a recommendation algorithm for questions based on implicit relations among questions acquired by students' learning log data.

Although many personalized learning material recommendation algorithms have been proposed, they do not directly support learning efficiency. Instead, they consider student preferences or/and understanding levels [6, 5, 12, 13, 19, 20]. This research focuses on situations where students take the same test before and after studying to quantify the expected improvement in test scores; appropriate questions must be recommended to maximize such expected improvement.

In the research field of recommendation systems, since the cognitive load on users to assign explicit ratings is heavy, gathering enough appropriate ratings is difficult [16, 2]. Therefore, in our approach, test scores are used to determine the relations between questions and student explicit ratings are not applied.

Our method recommends questions by which students can provide correct answers to questions that they answered incorrectly before studying. Effective questions for incorrectly answered questions may be identical for most students. If particular questions are studied between tests and student test scores are increased after such studying, they may provide knowledge about questions whose answers were changed and became correct. The expected improvement of test scores based on questions is calculated by logistic regression models [9]. By training the logistic regression models using student test scores and their learning log data, namely, the studied questions, questions that improve test scores can be automatically extracted.

The remainder of this paper is organized as follows. In Section 2, we briefly review related work. In Section 3, we explain our approach for providing effective questions. In Section 4, we present our recommendation algorithm, and our method is evaluated using learning log data and test scores in Section 5.

2 Related Work

Traditional intelligent tutoring systems (ITS) evaluate the learning situations of students and give learning materials that supply the unacquired knowledge of students [13, 10, 11, 1, 4]. In such systems, metadata about materials are attached beforehand, including the difficulty, topic, prerequisites, and the relationship between materials. The learning materials to be provided are determined by traversing the databases of learning materials based on student answers, the metadata of the learning materials, and the relations among them. This method may be appropriate if the number of learning materials is not so large and their relations and metadata can be defined uniquely. Our research focuses on situations where plenty of questions exist whose relations are not defined beforehand. Moreover, the unacquired knowledge of students is not uniquely determined by the metadata. Therefore, attaching metadata to questions and determining questions based on it is impractical. Instead of statically preparing metadata of the learning materials, our method grasps the implicit relations among questions based on the learning log data of the students who tackle the same tests and provides questions that may affect the incorrectly answered questions.

Many methods for recommending learning materials, which are also called curriculum sequencing, adaptive tutoring, or personalized learning path guidance, have been proposed [6, 5, 12, 13, 19, 20]. For example, the Personalized E-Learning system based on Item Response Theory (PEL-IRT) [6] automatically adjusts the difficulty of learning materials according to the student's level of understanding based on item response theory [8] to provide appropriate learning materials. However, since PEL-IRT does not consider questions that students need to tackle in the test, it does not necessarily provide learning materials that can improve test scores.

Recommendation systems for books, music, and movies are used in many on-line stores [18], where collaborative filtering [17] is a common method. Collaborative filtering recommends products that are purchased by users with similar preferences. E-Learning systems that reflect student preferences have also been proposed [13]. However, if students only use their favorite materials their test scores may not improve.

Our method is related to the scheme described in [5] because both involve a test for each student prior to learning. The method in [5] recommends materials considering their difficulty and incorrectly answered questions. However, the objective of this method is to make students understand unacquired knowledge. Since this method does not consider learning efficiency, students need to cope with many questions before understanding the questions that they answered incorrectly. Our research increases specific test scores previously tackled by students with a limited number of questions.

3 Approach

This research assumes that the student, who took a couple of tests in the past, is studying a set of multiple-choice cloze questions in the *studying phase* to take a new test after the studying phase, where each test consists of a number of questions. While studying, students tackle different questions from those in the tests. The aim of this research is not to give a large volume of questions. Instead, it provides a limited number of suitable questions that can help increase test scores. The increase of test scores corresponds to situations where students can successfully acquire the knowledge on which the questions focus in the tests.

Students need various grammar and vocabulary knowledge for answering English multiple-choice cloze questions. Of course, important knowledge exists to determine answers. Effective questions for maximizing test scores have the same important knowledge as incorrectly answered questions. If the important knowledge is identical for different questions, the explanations for deriving the answer for one question can be applied to other questions. Such questions can be inferred by student learning log data. When two questions share important knowledge and students cannot answer one of the questions in the tests, they may answer correctly after studying with other questions. Of course, some students may not notice the relations between the questions. Effective questions are those that give hints for deriving incorrectly answered questions. Therefore, studied questions are detected as effective for test questions whose answers are changed from incorrect to correct answer by many students.

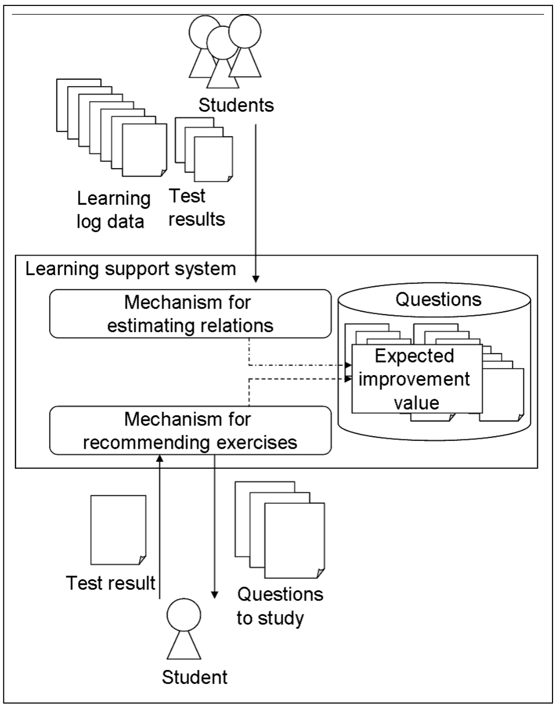


Figure 1: Framework of learning support system

Figure 1 shows the framework of our learning support system. The mechanism for recommending questions selects effective questions from the question database and provides appropriate study questions based on the input test results of students. To evaluate the questions, the expected improvement values are introduced that represent the expected test scores when individual questions are posed for studying. The expected improvement values are calculated based on the learning log data and the test results of many other students. In the mechanism for estimating the relations between questions, the changes of the test results and the studied questions are investigated and set as expected improvement values.

4 Proposed Method

4.1 Problem setting

The goal of our method is to select questions to be used during the studying phase from a set of questions to enhance learning efficiency, which is quantified by the expected improvement in test scores. In our approach, the implicit relations between test questions and studied questions are learned.

Let z_j be a variable that represents whether question j is recommended in the studying phase as follows:

$$z_j = \begin{cases} 1 & \text{if question } j \text{ is studied} \\ & \text{in the studying phase,} \\ 0 & \text{if question } j \text{ is not studied.} \end{cases} \tag{1}$$

The studied questions are represented by vector $\boldsymbol{z} = (z_j)_{j \in \boldsymbol{M}}$, where $\boldsymbol{M}$ represents a set of all candidate questions to study.

Let x_i and y_i be variables that represent whether question i is correctly or incorrectly answered before and after the studying phases, as follows:

$$x_i = \begin{cases} 1 & \text{if question } i \text{ is correctly answered} \\ & \text{before the studying phase,} \\ -1 & \text{if question } i \text{ is incorrectly answered} \\ & \text{before the studying phase,} \\ 0 & \text{if question } i \text{ is not answered,} \end{cases} \tag{2}$$

$$y_i = \begin{cases} 1 & \text{if question } i \text{ is correctly answered} \\ & \text{after the studying phase,} \\ -1 & \text{if question } i \text{ is incorrectly answered} \\ & \text{after the studying phase,} \\ 0 & \text{if question } i \text{ is not answered,} \end{cases} \tag{3}$$

The results of the set of test questions $\boldsymbol{V}$ before and after the studying phase are represented by vectors $\boldsymbol{x} = (x_i)_{i \in \boldsymbol{V}}$ and $\boldsymbol{y} = (y_i)_{i \in \boldsymbol{V}}$, respectively.

The improvement in the test scores expected from recommended questions $\boldsymbol{z}$ is written as follows:

$$E(\boldsymbol{z}) = \sum_{i \in \boldsymbol{V}} S_i P(i) P(x_i = -1) P(y_i = 1 | x_i = -1, \boldsymbol{z}), \tag{4}$$

where S_i represents the score allocated to question i, $P(i)$ represents the probability that question i is asked in future test $P(i) + \bar{P}(i) = 1$ in which $\bar{P}(i)$ represents the probability that question i is not asked in future tests, $P(x_i = -1)$ represents the probability that question i is incorrectly answered before the studying phase, and $P(y_i = 1 | x_i = -1, \boldsymbol{z})$ represents the probability that question i is correctly answered after the studying phase when question i is incorrectly answered before the studying phase and questions $\boldsymbol{z}$ are recommended in the studying phase.

If the test result before studying phase x is given, the expected improvement in the test score can be simplified as follows:

$$E(\boldsymbol{z} | \boldsymbol{x}) = \sum_{i : x_i = -1} S_i P(i) P(y_i = 1 | x_i = -1, \boldsymbol{z}). \tag{5}$$

We use Eq. (5) as the expected improvement in test scores in the following sections.

4.2 Recommendation algorithm

Appropriate questions may differ according to previously studied questions from the studying phase. Our method sequentially selects a question that maximizes the expected improvement from questions that have not yet been recommended as follows:

$$\hat{j} = \arg \max_{j : z_j = 0} E(\boldsymbol{z}^{+j} | \boldsymbol{x}), \tag{6}$$

where $\boldsymbol{z} = (z_j)_{j \in \boldsymbol{M}}$ represents the currently studied questions and $\boldsymbol{z}^{+j}$ represents the studied questions when question j is newly recommended, or $z_{j'}^{+j} = 1$ if $j = j'$ and $z_{j'}^{+j} = z_{j'}$ if $j \neq j'$. Table 1 shows our method's question recommendation procedure. Examples of end conditions include those where the number of studied questions, the expected improvement, or the time period of the studying phase exceeds a certain threshold.

We use a greedy algorithm to determine a set of questions to be recommended as described above. If the number of questions is fixed and known before the studying phase, we can identify the set of questions that maximizes the improvement scores by calculating Eq. (5) for all possible combinations of the fixed number of questions. A greedy algorithm is used because it is fast and does not require the number of studied questions to be fixed beforehand.

Table 1: Question recommendation procedure with proposed method.

1. Input the test result before studying $\boldsymbol{x}$;
2. Initialize the studied question vector: $\boldsymbol{z} = (0, \cdots, 0)$;
3. Select question $\hat{j}$ which was recommended by Eq. (6);
4. Update the studied question vector: $\boldsymbol{z} = \boldsymbol{z}^{+\hat{j}}$;
5. Return to step 3 unless an end condition is satisfied.

4.3 Improvement model

When recommending questions, our method requires improvement model $P(y_{ni} = 1 | x_{ni} = -1, \boldsymbol{z}_n)$, which is the probability of the improvement of answering question i with study questions $\boldsymbol{z}$. We model the improvement using logistic regression [9] as follows:

$$P(y_{ni} = 1 | x_{ni} = -1, \boldsymbol{z}_n) = \frac{1}{1 + \exp\left(-(\mu_i + \boldsymbol{\theta}_i^\top \boldsymbol{z}_n)\right)}, \tag{7}$$

where μ_i and $\boldsymbol{\theta}_i = (\theta_{ij})_{j \in \boldsymbol{M}}$ are unknown parameters. Although the logistic regression is widely used for binary classifiers, it is the first attempt to model the improvement of scores, to our knowledge. Intuitively, μ_i represents the ease with which the answer to question i is improved, and θ_{ij} represents the influence of question j on the improvement in the answer to question i. Unknown parameters $\boldsymbol{\Theta} = \{\mu_i, \boldsymbol{\theta}_i\}_{i \in \boldsymbol{V}}$ can be estimated by maximizing the following log likelihood, which consists of the learning log data and the test results for set of students $\boldsymbol{N}$:

$$\begin{aligned}
& L(\boldsymbol{\Theta}) \\
= & \sum_{n \in \boldsymbol{N}} \sum_{i \in \boldsymbol{V}} \Big(I(x_{ni} = -1 \wedge y_{ni} = 1) \\
& \times \log P(y_{ni} = 1 | x_{ni} = -1, \boldsymbol{z}_n) \\
& + I(x_{ni} = -1 \wedge y_{ni} = -1) \\
& \times \log P(y_{ni} = -1 | x_{ni} = -1, \boldsymbol{z}_n) \Big) \\
= & \sum_{n \in \boldsymbol{N}} \sum_{i \in \boldsymbol{V}} \Big(I(x_{ni} = -1 \wedge y_{ni} = 1)(\mu_i + \boldsymbol{\theta}_i^\top \boldsymbol{z}_n) \\
& - I(x_{ni} = -1) \log\big(1 + \exp(\mu_i + \boldsymbol{\theta}_i^\top \boldsymbol{z}_n)\big) \Big), \quad (8)
\end{aligned}$$

where x_{ni} and y_{ni} indicate whether question i was correctly or incorrectly answered by student n before and after the studying phases and

$$\begin{aligned}
& P(y_{ni} = -1 | x_{ni} = -1, \boldsymbol{z}_n) \\
= & 1 - P(y_{ni} = 1 | x_{ni} = -1, \boldsymbol{z}_n) \\
= & \frac{1}{1 + \exp(\mu_i + \boldsymbol{\theta}_i^\top \boldsymbol{z}_n)}, \quad (9)
\end{aligned}$$

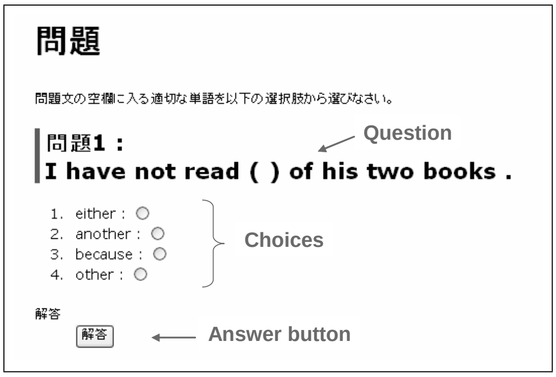
represents the probability that question i is incorrectly answered by student n when question i is incorrectly answered before the studying phase and questions $\boldsymbol{z}$ are recommended. In the experiments, we used a quasi-Newton method [15] for the optimization and Gaussian priors with zero means for the unknown parameters to avoid overfitting [7]. The Hessian of the log likelihood function with Gaussian priors with respect to the parameters is positive definite. Therefore, the log likelihood is a convex function, and the global optimum solution is guaranteed.

5 Experiments

5.1 Setting

We evaluated our method using TOEIC multiple-choice cloze questions in which students select appropriate words from four options with the correct grammar for the blank in the sentence.

We implemented a web-based e-Learning system for the evaluation. In the experiment, students take tests before and after the studying phase, which we call pre-test and post-test, to measure the effect of studying on improving test scores. One question is presented on one web page, and students answer each question in a series.



問題

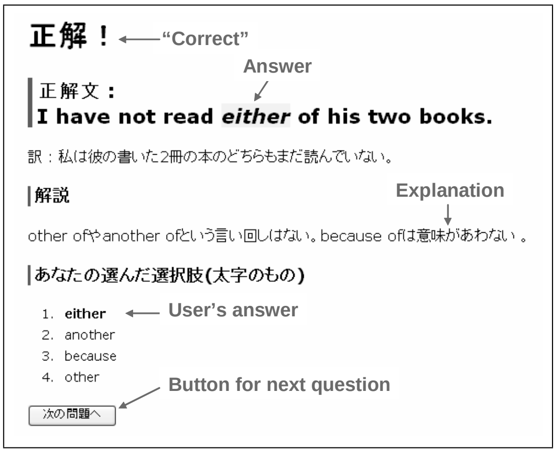
問題文の空欄に入る適切な単語を以下の選択肢から選びなさい。

問題1：
I have not read ( ) of his two books .

1. either : ○
2. another : ○
3. because : ○
4. other : ○

解答

Figure 2: Web page posing a question generated by our e-Learning system



正解！

正解文：
I have not read *either* of his two books.

訳：私は彼の書いた2冊の本のどちらもまだ読んでいない。

解説

other ofやanother ofという言い回しはない。because ofは意味があわない 。

あなたの選んだ選択肢(太字のもの)

1. **either**
2. another
3. because
4. other

Figure 3: Web page showing the solution and explanation generated by our e-Learning system

The 40 questions in the pre- and post-tests are identical, $|\boldsymbol{V}| = 40$.

The recommended questions in the studying phase are provided with solutions and explanations. One question is recommended to each student on one web page for studying, and the solution and explanation are presented on another web page after the student has answered the question. Note that the students are not supplied with solutions and explanations in the pre- and post-tests. There are 80 candidate questions for the studying phase, $|\boldsymbol{M}| = 80$, and 40 are recommended to each student in the studying phase. The recommended questions are different from the questions in the pre- and post-tests. However, about half are related to the test questions, for example, they involve questions about identical idioms and grammatical rules. They were heuristically selected by the authors.

Figures 2 and 3 are the web pages generated by our e-Learning system. Fig. 2 is an web page that displays a question sentence and a choice of fill in the blank answers. The student selects one answer and pushes the answer button. Fig. 3 is an web page showing the solution and explanation for the studying phase. The student's answer is evaluated and its correctness is displayed at the top. After the student pushes the button, the next question appears.

5.2 Evaluation of improvement models

Our method requires improvement model $P(y_i = 1|x_i = -1, \boldsymbol{z})$. We constructed and evaluated improvement models using the log data of 52 students, such as $|\boldsymbol{N}| = 52$, with random material recommendations.

The proposed improved model (7) estimates the improvement probability by utilizing relationships among questions. To evaluate the effect of the relationships, we compared the proposed model with a model that does not consider relationships among questions or the Bernoulli model as follows:

$$P(y_i|x_i = -1, \boldsymbol{z}) = \phi_i^{\frac{1+y_i}{2}} (1 - \phi_i)^{\frac{1-y_i}{2}}, \qquad (10)$$

where ϕ_i represents the probability that question i is correctly answered in the post-test when question i is incorrectly answered in the pre-test. The Bernoulli model assumes that the improvement does not depend on recommended materials $\boldsymbol{z}$. Parameter ϕ_i can be estimated based on the maximum likelihood as follows:

$$\hat{\phi}_i = \frac{\sum_{n \in \boldsymbol{N}} I(y_{ni} = 1 \wedge x_{ni} = -1)}{\sum_{n \in \boldsymbol{N}} I(x_{ni} = -1)}. \qquad (11)$$

For the evaluation measurement, we used the AUC [3] of the problem to predict whether questions that were incorrectly answered in the pre-test are correctly answered in the post-test. AUC is the area under the Receiver Operating Characteristic (ROC) curve, where the ROC curve is a graphical plot of the true positive rate versus the false positive rate. A higher AUC represents better predictive performance. We computed AUC using leave-one-out cross-validation. We used 52 evaluation data sets, in each of which one student's data were used for the evaluation and the data of the other 51 students were used for training. Table 2 shows the AUC, and Fig. 4 shows the ROC curve. The AUC of

Table 2: AUC of improvement models based on Bernoulli distribution and logistic regression

Bernoulli distribution	Logistic regression
0.556	0.592

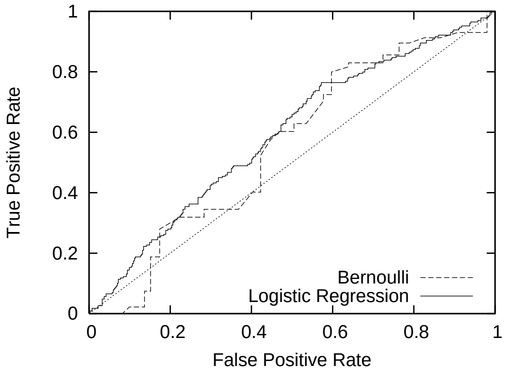


Figure 4: ROC curves for improvement models based on Bernoulli distribution and logistic regression

the improvement model based on the logistic regression is higher than that of the Bernoulli model, which implies that the recommended questions are important for predicting improvements in the scores and that we can predict them with the logistic regression model.

The computational time for learning the parameters of the proposed improvement model was 2.19 second when using a computer with a Xeon5160 3GHz CPU.

5.3 Analysis of improvement models

The highest and second highest θ_{ij} were $\theta_{i_1j_1} = 0.388$ and $\theta_{i_2j_2} = 0.208$, where questions in test i_1 and studying phase j_1 contain the idiom 'stop by' and questions in test i_2 and studying phase j_2 contain the idiom 'across the street.' This result is natural because recommendations of questions about identical idioms can improve test scores. Even though our method does not use information related to question contents, it automatically extracts the relationship between questions using the learning log data and the test results.

We analyzed the relationship between questions in the tests and in the studying phase using questions i_1 and j_1 that contain the idiom 'stop by' as an example. In the pre-test, 32 students answered question i_1 incorrectly. The student results for the studying phase and the post-test are shown in Table 3. The probability of students answering questions i_1 correctly in the post-test when question j_1 was recommended in the studying phase was $\hat{P}(y_{i_1} = 1|x_{i_1} = -1, z_{j_1} = 1) = 15/17$. In contrast, the probability when question j_1 was not recommended in the studying phase was $\hat{P}(y_{i_1} = 1|x_{i_1} = -1, z_{j_1} = 0) = 2/15$. This result indicates that the recommendation of question j_1 effectively improved the responses to question i_1.

Table 3: Students who incorrectly answered question x_{i_1} that contained idiom 'stop by'

pre-test	studying phase	post-test	# of
x_{i_1}	z_{j_1}	y_{i_1}	students
-1	1	1	15
-1	1	-1	2
-1	0	1	2
-1	0	-1	13

Algorithm 1 Simulation of student n with our recommendation

```
1: Set $\boldsymbol{z} \leftarrow \mathbf{0}$, $t \leftarrow 1$
2: while $t \leq T$ do
3:     $\hat{j} = \arg\max_{j:z_j=0} E(\boldsymbol{z}^{+j}|\boldsymbol{x}_n)$
4:     Set $\boldsymbol{z} \leftarrow \boldsymbol{z}^{+j}$
5:     Set $t \leftarrow t + 1$
6: end while
7: Output $E(\boldsymbol{z}|\boldsymbol{x}_n)$
```

5.4 Evaluation of recommendation algorithms by simulations

We examined the effectiveness of our recommendation algorithm by simulation. Student behavior was simulated using the improvement model that was estimated using the log data of the 52 students described above. The function of Algorithm 1 is to generate an expected improvement of the test scores of student n, where T is the number of recommendations set as 40.

We compared our algorithm with the following three algorithms: Random, Mistakable, and Level. Random randomly recommends a question. Mistakable recommends a question that is mistaken by many students. Recommendation question $\hat{j}$ is determined as follows:

$$\hat{j} = \arg\max_{j:z_j=0} \sum_{n \in \boldsymbol{N}} I(z_{nj} = -1), \qquad (12)$$

where $z_{nj} = -1$ represents that student n incorrectly

Table 4: Average percentage of expected improvement rates and standard deviations with simulations

Random	Mistakable	Level	Proposed
36.4 ± 4.9	38.6 ± 4.4	38.1 ± 4.1	**52.7** ± 7.6

Table 5: Average percentage improvement rates and standard deviations with actual students

Random	Proposed
15.6 ± 22.9	**27.5** ± 20.6

Table 6: Average scores and standard deviations in pre-test, studying phase, and post-test

	pre-test	studying phase	post-test
Random	72.8 ± 15.7	73.1 ± 14.4	77.3 ± 14.1
Proposed	70.1 ± 13.2	69.8 ± 12.7	78.2 ± 13.1

answered question j in the studying phase. Level recommends a question whose difficulty is appropriate to the student's understanding level by selecting recommendation question $\hat{j}$ as follows:

$$\hat{j} = \arg\min_{j:z_j=0} |d_j - l_n|, \tag{13}$$

where $d_j = \frac{1}{|\boldsymbol{N}|}\sum_{n\in\boldsymbol{N}} I(z_{nj} = -1)$ is the difficulty of question j and $l_n = \frac{1}{|\boldsymbol{V}|}\sum_{i\in\boldsymbol{V}} I(x_{ni} = -1)$ is the level of student n.

We evaluated the recommendation algorithm by the expected improvement rate, which is the expected number of correctly answered questions after studying to the number of incorrectly answered questions before studying as follows:

$$S_n = \frac{E(\boldsymbol{z}|\boldsymbol{x}_n)}{\sum_{i\in\boldsymbol{V}} I(x_{ni} = -1)} \times 100. \tag{14}$$

Here, we assumed that questions correctly answered before studying were also correct after studying. Table 4 shows the expected improvement rates for each recommendation algorithm. The proposed algorithm outperformed the others for improving the expected test scores. The Mistakable and Level algorithms also improved the scores more than the Random algorithm, although the effect was smaller than that of our algorithm because they did not directly maximize test scores.

5.5 Evaluation of recommendation algorithms by actual students

We evaluated the learning efficiency of our recommendation algorithm with actual students. Because the evaluation of recommendation with many actual students costs much, we compared the proposed method with the most basic random method. 38 students studied questions recommended randomly, and 49 studied questions recommended by our method. The students included members of the graduate and undergraduate schools and the staffs of Nagoya and Kansai Universities. All had previously studied the basic English knowledge for the provided questions in high school.

We evaluated the recommendation algorithm by the improvement rate, which is the correctly answered questions after studying to the incorrectly answered questions before studying as follows:

$$R_n = \left(1 - \frac{\sum_{i\in\boldsymbol{V}} I(y_{ni} = -1)}{\sum_{i\in\boldsymbol{V}} I(x_{ni} = -1)}\right) \times 100. \tag{15}$$

Table 5 shows the average improvement rates. Our algorithm corrected 27.5% of the mistakes and provided statistically significant increases compared with the random recommendation method (one-tailed t-test, $p < 0.007$). The rates of real questions in Table 5 were smaller than those in the simulations in Table 4 because we assumed that questions correctly answered before learning were not incorrectly answered after studying in the simulations.

Table 6 shows the average scores obtained in the pre-test, the studying phase, and the post-test. The maximum score is 100, which means the score assigned to one question is 2.5. Although the average pre-test score with our method is lower than that with the random recommendation method, the average post-test score with our method is higher than that with the random recommendation method. Moreover, since the deviation with our method is smaller than that with the random recommendation method, many students with our system successfully increased their test scores. This result indicates that our method is superior to random recommendation methods.

Figure 5 shows the average improvements in the test scores in relation to the pre-test scores. The improvement is in inverse proportion to the pre-test score. Our method is superior to the random recommendation

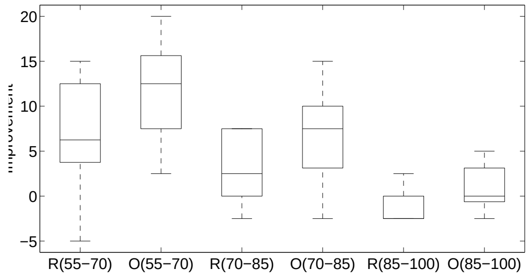


Figure 5: Average improvements in test scores with random recommendation 'R' and our method 'O' in relation to pre-test scores. Values in parentheses are pre-test scores

Table 7: Students who were/were not recommended related questions when questions that contain identical idioms were incorrectly answered in pre-tests

idiom	recommended	not recommended
stop by	30	0
due to	14	1
across the street	15	2

Table 8: Questionnaire answers: "Did the questions in the studying phase help in the post-test?"

answer	frequency
Yes	42
No	7

method regardless of the pre-test score.

We also analyzed whether our method recommended questions that can improve test scores when students incorrectly answered questions in tests that included the same idioms. Table 7 shows the number of students who were/were not recommended related questions when questions about idioms were incorrectly answered in the pre-test. This result shows only few cases where the related questions were not recommended with our method.

After studying, students who learned with our method received questionnaires about the recommended questions. Table 8 shows the answers to the question "Did the questions help you in the post-test?" Since the questions were helpful for about 86% of the students, our method successfully provided questions that related to the test questions. Table 9 shows a list of words and idioms in the recommended questions cited by students as useful for deriving the answers in the post-test and the numbers of students who cited them. The list contains six idioms and five words from various parts of speech. This result indicates that our method can extract relationships from various parts of speech.

Table 9: Useful words/idioms in questions in studying phase and number of students who cited them

word/idiom	# of students
stop by	7
across the street	3
the day after tomorrow	2
due to	2
launch	2
although	1
despite	1
follow in one's footsteps	1
resign	1
in order to	1
among	1

The computational time for recommending a material with the proposed method was 0.04 second. The computational efficiency indicates that the proposed method can be used for a real time recommender system.

6 Conclusion

We proposed a method for recommending questions for English multiple-choice cloze questions that maximizes learning efficiency based on the expected improvement in test scores. The experimental results suggest that our question recommendation approach is promising and will become a useful tool for e-Learning.

Although we modeled the expected improvement in test scores with simple logistic regression using only the learning log data and test results to simplify our framework's novelty, we could also use more information about the questions and student attributes for the modeling.

Currently, we are focusing on English multiple-choice cloze questions. However, this recommendation method is not specific to such questions. It can also be applied to mathematics and physics, for example, because implicit relations among questions for understanding may exist in other fields. In addition, learning materials tackled by students in the studying phase can also be other types of materials. Students learn not only with questions but also with textbooks. Our approach does not depend on questions; it can be applied to other types of materials. We would like to apply our proposed method to other courses of learning and other types of learning materials. We plan further verification of our proposed method by comparing other methods with actual students.

References

[1] M. E. S. A, E. Aimeur, and C. Frasson. Towards a case-based intelligent tutoring system for student modelling. In *ICCE '98: Proceedings of the International Conference on Computers in Education*, volume 1, pages 528–535, 1998.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. In *IEEE Trans. on Knowledge and Data Engineering*, volume 17, pages 734–739, 2005.

[3] A. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. In *Pattern Recognition*, volume 30, pages 1145–1159, 1997.

[4] C. M. Buff and M. A. Williams. An intelligent tutoring system architecture based on knowledge representation and reasoning. In *ICCE '99: Proceedings of the International Conference on Computers in Education*, volume 1, pages 736–743, 1999.

[5] C.-M. Chen, C.-M. Hong, and M.-H. Chang. Personalized learning path generation scheme utilizing genetic algorithm for web-based learning. *WSEAS Transactions on Information Science and Applications*, 3(1):88–95, 2006.

[6] C.-M. Chen, H.-M. Lee, and Y.-H. Chen. Personalized e-learning system using item response theory. *Comput. Educ.*, 44(3):237–255, 2005.

[7] S. F. Chen and R. Rosenfeld. A Gaussian prior for smoothing maximum entropy models. Technical report, CMUCS–99–108, 1999.

[8] R. K. Hambleton, H. Swaminathan, and H. J. Rogers. *Fundamentals of Item Response Theory*. Sage Publications, Newburg Park, 1991.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, New York, 2001.

[10] R. Hübscher. Logically optimal curriculum sequences for adaptive hypermedia systems. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 121–132, 2000.

[11] G. F. Knolmayer. Decision support models for composing and navigating through e-learning objects. In *HICSS '03: Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, Washington, DC, USA, 2003. IEEE Computer Society.

[12] M.-G. Lee. Profiling students' adaptation styles in web-based learning. *Comput. Educ.*, 36:121–132, 2001.

[13] X. Li and S. K. Chang. A personalized e-learning system based on user profile constructed using information fusion. In *Proceedings of the 11th International Conference on Distributed Multimedia Systems*, pages 109–114, Banff, Canada, Sep. 2005.

[14] Y.-C. Lin, L.-C. Sung, and M. C. Chen. An automatic multiple-choice question generation scheme for english adjective understanding. In *ICCE2007 Workshop Proceedings of Modeling, Management and Generation of Problems / Questions in eLearning*, pages 137–142, 2007.

[15] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Programming*, 45(3):503–528, 1989.

[16] D. W. Oard and J. Kim. Implicit feedback for recommender systems. In *AAAI Technical Report WS-98-08*, pages 81–83, 1998.

[17] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: an open architecture for collaborative filtering of netnews. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 175–186, 1994.

[18] J. B. Schafer, J. A. Konstan, and J. Riedl. E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, 5:115–153, 2001.

[19] M. Stern and B. P. Woolff. Curriculum sequencing in a web-based tutor. In *ITS '98: Proceedings of the 4th International Conference on Intelligent Tutoring Systems*, pages 574–583, London, UK, 1998. Springer-Verlag.

[20] T. Y. Tang and G. Mccalla. Smart recommendation for evolving e-learning system. In *Proceedings of the 11th International Conference on Artificial Intelligence in Education, Workshop on Technologies for Electronic Documents for Supporting Learning*, pages 699–710, 2003.