Personalized Automatic Quiz Generation Based on Proficiency Level Estimation

Yi-Ting Huang[a*], Meng Chang Chen[b], Yeali S. Sun[a]
[a] *Department of Information Management, National Taiwan University, Taiwan.*
[b] *Institute of Information Science, Academia Sinica, Taiwan.*
*d97008@im.ntu.edu.tw

Abstract: Recent years have seen increased attention given to computer-aided question generation for language student testing and evaluation. However, this approach often directly provides examinees with exhaustive questions. This is inappropriate, because these questions are not designed for any specific testing purpose. In this work, we present a personalized automatic quiz generation model that generates multiple-choice questions at various difficulty levels and categories, including grammar, vocabulary, and reading comprehension. We combined this model with a quiz strategy for estimating examinee proficiency and question selection. The proficiency is estimated using Exponential Moving Average, combining the test responses with a student's past history. The results show that the subjects in the experimental group corrected their mistakes more frequently as well as answered more difficult questions than the control group. The experimental group also demonstrated the most progress between the pre-test and post-test. In addition, most of subjects agree the quality of the generated questions in the questionnaire analysis.

Keywords: Computer-aids question generation, personalized learning, adaptive test.

1. Introduction

Recent years have seen increased attention given to computer-aided question generation in the field of computer-assisted language learning. A growing number of studies are now available for designing different question types, such as multiple-choice test items [8-10], and cloze tests [5]. The first computer-aided question generation was proposed by Mitkov and Ha [10], and generated multiple-choice items by term frequencies or predefined syntactic patterns. Lin, Sung, and Chen [8] improved the design of Mitkov and Ha [10] by investigating the semantics of words and presenting vocabulary items, including collocation, antonym, synonym and similar word questions. The Sakumon system [5] developed a cloze test as another approach to automatically assess examinees' vocabulary and grammar skills. Beyond vocabulary assessment, Chen, Ko, Wu and Chang [3] focused on automatic grammar quiz generation. Lastly, the MARCT system [13] investigated reading comprehension and designed three question stem templates for generating questions. These studies propose methods for reducing the labor of manual question generation, instead automatically generating a list of all possible questions when given a document. However, this exhaustive list of questions is inappropriate for language learning, because it can lead to redundant, overly-simplistic test questions that are unfit for evaluating student progress. Moreover, the characteristic of items generated from these studies is insufficient. It is hard to facilitate meaningful test purpose and maximize examinees' learning outcomes.

In contrast to this approach, Item Response Theory (IRT) with computerized adaptive testing has steadily developed question selection from the relationship between an

examinee's proficiency and the properties of the questions [4]. In order to create an adaptive test, IRT requires parameters, such as item difficulty parameter or item discrimination parameter, which can be determined using the items for a large number of samples first (as pre-calibration) and then manually derive the question parameters. However, their approach is more time-consuming than more automatic alternatives. In addition, in applying IRT, an examinee's ability can be obtained by observing responses during a test and then estimating using maximum likelihood estimation, maximum a posteriori or expected a posteriori [4]. While effective however, these methods only consider test responses at the time of testing, rather than incorporating this testing history.

While the studies above all investigate computer-aided question generation, little research has discussed the role of question difficulty during question generation. As a result, this study specifically examines three question types, each with various difficulty settings, relative to an examinee's proficiency level. Moreover, unlike previous research, this study incorporates an examinee's test history when estimating the proficiency. Together these improvements set this paper's model apart from the approaches briefly outlined above.

2. Automatic Quiz Generation

Figure 1 illustrates the traditional system architecture of the multiple-choice item generation process. Given the target learning material, items are created from the quiz generation process, composed of the stem generation, the answer determination, and the distractor determination. The stem generation forms a direct question or incomplete statement using predefined templates. The answer determination decides the correct answers of the question. Lastly, the distractor determination selects the plausible wrong alternatives as distractors from external resources, in order to discriminate good students from poor ones.

As shown in Figure 2, four questions (also called items) are generated from a document describing the origins of Halloween. In the document, the bolded sentences are selected by the stem generation to produce question stems, while the bold and underline word in the bolded sentences are decided by the answer determination. In the four questions, the bolded words represent stems, the bold italics are called answers, and the other plausible choices in the questions are distractors.

In this study, we generate three question types with difficulties based on the same system architecture. Those types are multiple-choice items, including vocabulary, grammar and reading comprehension. A stem is usually generated by predefined templates. In the following section, both the answer determination and the distractor determination are presented.

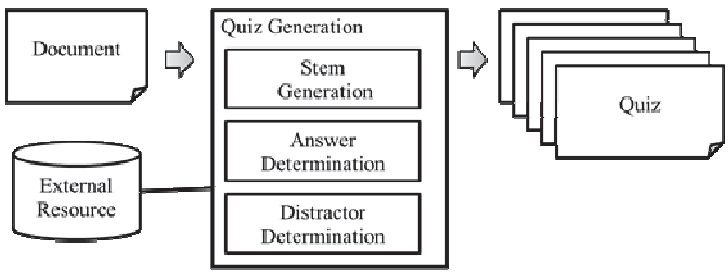

Figure 1. Overview of Traditional Automatic Quiz Generation.

2.1 Vocabulary Items

In this study, vocabulary items are generated according to an examinee's proficiency ability. The difficulty of a vocabulary question is based on the difficulty of the correct answer. We

assume that if a student selects the correct answer, it is probable that he or she understood the question stem, and distinguished the correct answer from distractors. Here, word difficulty is determined by a grading word list which made by experts (http://www.ceec.edu.tw/research/paper_doc/ce37/5.pdf). When given the vocabulary proficiency level of an examinee, words with the same difficulty level in the given document are selected as the basis to form test questions.

The majority of research on vocabulary assessment extracts plausible distractors from various resources, such as a thesaurus [8] or words from the same document [10], and then designs algorithms to select the most suitable distractors for a given question. We consult the word list to ascertain vocabulary word difficulty, and then select distractor candidates of equal difficulty, part-of-speech, similar character length and small edit distance.

The first question in Figure 2 is an example of a vocabulary question. The question stem is composed from a manual template, the correct answer is found within an original document and taught in the fourth grade, and the difficulty level of the distractors is the same as the answer.

Document

Halloween, which falls on October 31, is one of the most unusual and fun holidays in the United States. It is also one of the scariest! **It is associated with ghosts, skeletons, witches, and other scary images.** …**Many of the original Halloween traditions have developed today into fun activities for children.** The most popular one is "trick or treat." On Halloween night, children dress up in costumes and go to visit their neighbors. When someone answers the door, the children cry out, "trick or treat!" What this means is, "Give us a treat, or we'll play a trick on you!" …This tradition comes from an old Irish story about a man$_1$ named **Jack**$_2$ who was very stingy. ... But he$_3$ also could not enter hell, because he$_4$ had once played a trick on the devil$_5$. **All he could do was walk the earth as a ghost, carrying a lantern**...

Quiz

1. **In the sentence "It is _____ with ghosts, skeletons, witches, and other scary images.", the blank can be:**
   (1) distributed (2) ***associated*** (3) contributed (4) illustrated
2. **In the Sentence, "Many of the original Halloween traditions _____ today into fun activities for children.", the blank can be filled in:**
   (1) ***have developed*** (2) have developing (3) is developed (4) develop
3. **The word "he" in this sentence "All he could do was walk the earth as a ghost, carrying a lantern" refer to:**
   (1) ghost (2) devil (3) witch (4) ***Jack***
4. **Which of the following statement is TRUE?**
   (1) On Halloween night, neighbors dress up in costumes and go to visit their children.
   (2) What this means is, "Give us a trick, or we'll play a treat on you!"
   (3) But the devil also could not enter hell, because he had once played a trick on the witch.
   (4) Jack was so stingy that he could not enter heaven when he died.

Figure 2. An Example in Automatic Quiz Generation.

2.2 Grammar Items

In this study, we manually predefine grammar patterns and distractor templates to generate grammar items. A set of 44 grammar patterns and distractor templates are identified from language learning textbooks. These grammar patterns are then implemented in the form of Tgrep2 patterns [7]. The difficulty of a grammar item is similar to the vocabulary item and determined by the difficulty of the correct answer. Unfortunately however, there is no predefined grammar difficulty measure available, similar to the aforementioned word list. Thus, we assigned the difficulty of a grammar pattern based on the textbook grade in which it frequently appears, which represents the age of grammar acquisition.

The second question in Figure 2 is an example of a grammar question. The target test purpose in the second question is "present perfect tense," taught in the first grade. The distractors refer to a grammar textbook to generate distractor templates about "tense." This approach helps clarify the difference between advanced students and non-advanced students. It distinguishes advanced learners who understand the implicit purpose of the

question and identify the mistakes within the distractors from non-advanced learners who fail to comprehend the meaning of the question and choose the grammatical plausibility of the incorrect alternatives.

2.3 Reading Comprehension Items

In this work, we capture the relation between sentences to generate two kinds of meaningful reading questions based on noun phrase coreference resolution. Similar to Mitkov and Ha [10], who extracted nouns and noun phrases as important terminology in reading material, we also focus on the interaction of noun phrases as the test purpose. The purpose of noun phrase coreference resolution is to determine whether two expressions refer to the same entity in real life. An example is excerpted from Figure 2 (This tradition…on the devil$_5$). It is easy to see that *Jack*$_2$ means *man*$_1$ because of the semantic relationship between the sentences. The following *he*$_3$ and *he*$_4$ are more difficult to judge as referring to *Jack*$_1$ or *devil*$_5$ when examinees do not clearly understand the meaning of the context in the document. This information is used in this work to generate reading comprehension questions, in order to examine whether learners really understand the relationship between nouns in the given context.

There are two question types generated in the reading comprehension item. One type is an independent referential question for the single concept test purpose, while the other follows one of the frequent question templates in [13], "which of the following statement is (not) true," as the overall comprehension test purpose. When given a document, the coreferential relations are identified by the coreference system [11]. In the first type, noun phrases in the same coreference chain are selected as the correct answer, and noun phrases in the other coreference chains are determined as the distractors. In the second type, the correct answer is generated by replacing one noun phrase with another in the same coreference chain, and the distractors are composed by replacing the noun phrases with ones in the other coreference chain.

The difficulty of the reading comprehension questions is based on the reading level of the reading materials themselves. We assume that an examinee correctly answers a reading comprehension question because he or she could understand the whole of the story. The reading level estimation of a given document in recent years has increased noticeably. In this study, we adopt the measure of reading difficulty estimation developed by [6] to identify the difficulty of reading materials, as a difficulty measure for the reading comprehension questions.

The third question in Figure 2 is an example of an independent referential question, which assesses the concept of one entity involved in sentences. This question is answered correctly when examinees understand the reasonable semantics of the concept in the document. The fourth question in Figure 2 is an example of the overall referential question, which contains more than one concept that needs to be answered. This approach further examines the connected concepts of the given learning material.

3. PERSONALIZED QUIZ STRATEGY

In this section, the personalized quiz strategy based on automatic quiz generation is presented. This personalized quiz strategy aims to achieve the following two purposes: first, generating items depending on the proficiency level of an examinee, in order to provide an adaptive test; second, designing a suitable quiz in order to improve an examinee's proficiency.

3.1 Proficiency Level Estimation

Let P represent an examinee's proficiency level. Proficiency level is categorized as vocabulary ability level l_v, grammar ability level l_g and reading comprehension ability level l_r, so that an examinee's proficiency level is represented as $P = \{ l_v, l_g, l_r \}$. The variables in this formula respectively represent each examinee's proficiency level, consisting of vocabulary ability, grammatical ability, and reading ability. For a given current proficiency level $P_t = \{l_{v,t}, l_{g,t}, l_{r,t}\}$ where t represents an examinee's proficiency level in time t.

To assign an examinee's proficiency level, an examinee's current proficiency score is calculated first. This score is transformed by the following formula:

$$Y_t = \sum_{i=1}^{n} u_i / n, \quad u_i = \begin{cases} 1, & \text{if an examinee correctly answered an item } i \\ 0, & \text{if an examinee incorrectly answered an item } i \end{cases} \quad (1)$$

where Y_t is the initial proficiency score in time t, I is a set of questions in an exam, i represents the i_{th} question in the exam, n represents the number of questions in the exam and u_i represents the responses of the learner in the exam. This formula represents the percentage of n items an examinee answers correctly.

We also consider an examinee's performance history and employ exponential moving average (EMA) [2] to combine it with the current initial proficiency score, transformed by the following formula: $S_t = \alpha \times Y_t + (1-\alpha) \times S_{t-1}$ (2) where S_t is the final proficiency score in time t after the combination with EMA, S_{t-1} is the past proficiency score in the time $t-1$ as history records, $\alpha=2/(m+1)$ is a constant represented as a weight, and m represents the length of the moving window. The expectation proficiency in each grade level is also measured: $E_l = \sum_{i=1}^{n} p_{i,l} / n$ (3) where $p_{i,l}$ represents the percentage of the proficiency level in which l examinees correctly answered the question i. This formula represents the average probability that the proficiency level l examinees correctly answered the test.

An examinee's proficiency level is assigned to the closest expected proficiency in grade level: $\hat{l}_t = \arg\min |S_t - E_l|, \quad l_t \in \{l_{v,t}, l_{g,t}, l_{r,t}\}$ (4) where $\widehat{l_t}$ represents an estimated examinee's proficiency level in one of the proficiency categories in time t, S_t is a learner's proficiency score in (2) and E_l is the expected proficiency score in (3).

3.2 Quiz Strategy

This section presents the quiz strategy. When given a learner's ability l, it is critical to determine how to best form a test from a series of questions which match their ability. In [1], the researchers selected history-based questions consisting of the recently used questions and correctly answered questions. In this study, a test is composed of not only fit questions (a question's level is equal to a learner's level) and history-based questions (a question's level is easier than a learner's level) but also challenging questions (a question's level is more difficult than a learner's level). Like Barla, et al. [1], we define probability values to assign questions in a test. Here, the percentage of history-based questions, fit questions and challenging questions are 20%, 60% and 20%, respectively. When fit questions are answered incorrectly, they are stored in the system. During the next iteration of the test, if there is any similar question based on the same concept, this question will be selected first. The goal of this design is to enhance learners' impression and improve their proficiency.

4. PERSONALIZED QUIZ STRATEGY

4.1 Experimental Design

The proposed methods are developed from the AutoQuiz project [8][9], which provides English language learners with automatic quiz generation. AutoQuiz is implemented on the IWiLL learning platform, which offers learners an online English reading and writing environment. The reading interface and the test interface are given in Figure 3. A total of 2,481 items, composed of vocabulary, grammar, and reading comprehension, were automatically generated based on 72 new stories as reading materials. The news articles were collected from several global and local online news websites: Time For Kids, Student Times, Voice of America, CNN, China Post Online and Yahoo! News.

The participants in this study were high school students in Taiwan, divided into two groups: a control group with general automatic quiz generation, and an experimental group with personalized automatic quiz generation. 33 students participated within the control group, while 123 students participated in the experimental group. 21 and 72 subjects in the control group and the experimental group respectively completed all phases of the research.

The experiment was held from July 1st to September 30th, 2011. During the experiment, the subjects were asked to participate in twelve activities, consisting of reading an article and then taking a test. Each test was composed of ten vocabulary questions, five grammar questions, and three reading comprehension questions. In addition, there was a pre-test and post-test for evaluating changes in learner proficiency, each with a similar degree of difficulty. The proficiency level in this study is defined from one to six, corresponding to the six semesters of Taiwanese senior high school. Finally, 30 subjects in the experimental group volunteered to fill out a questionnaire that elicited information concerning the examinee experience and the quality of the generated questions. Questions in the questionnaire were taken from [12]. A five-point Likert scale was employed.

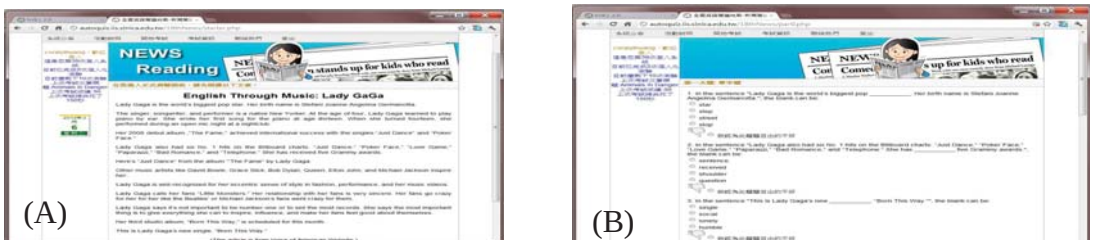

Figure 3. Snapshots of the system: (A) An example of a given reading materials from new online website; (B) An example of vocabulary items.

4.2 Experimental Results

The aim of the quiz strategy is to enhance student understanding of concepts they find unclear. We measured the rate at which students successfully corrected their mistakes on repeated concepts (denoted as the rectification rate) in the experimental group and control group, to determine the effect of generating items with repeated concepts. To make comparisons, the independent-samples t-test and the Mann-Whitney U test were both performed. Ideally, the distribution between the two groups is a normal distribution, and thereby uses a t-test. However, because of unequal sample sizes, the nonparametric method is complementary. The results suggest that the rectification rate in the experimental group was on average significantly higher than in the control group ($t=6.60$, $p<0.001$ in the independent-samples t-test and $Z=-5.97$, $p<0.001$ in the the Mann-Whitney U test). Moreover, the subjects in the experimental group ($M=0.54$, $SD=0.29$) were more than half as likely to correct unclear concepts and answer similar questions correctly. This indicates that a personalized quiz strategy would help the learners correct previous mistakes.

To further understand the influence of a personalized automatic quiz generation, the normalized score (normalized from zero to one) in the post-test between the experimental

group and control group were calculated and compared in the parametric and nonparametric analysis. The results of an independent T-test ($p=0.80$ in the pre-test and $p=0.46$ in the post-test) and the Mann-Whitney U test ($p=0.99$ in the pretest and $p=0.59$ in the post-test) showed no significant effect on the post-test between the experimental group and the control group. However, the paired sample T-test and the Wilcoxon signed-rank test showed a significant effect of the pre-test and the post-test in the experimental group ($p<0.01$), while the performance of the control group had no statistically significant effect ($p>0.05$). This indicates that the personalized automatic quiz generation within the experimental group still effectively improves their own learning.

To study the performance in each difficulty level between the pretest and post-test, the number of correctly answered questions among the six difficulty levels in the pretest and the post-test were computed. The tests are comprised of 28 items among six difficulty levels (six, three, six, three, seven and three questions per respective level, corresponding to levels one through six). A Chi-Square test for homogeneity of proportions was conducted to analyze the proportion between the pre-test and post-test. Table 1 presents two contingency tables respectively in the control group and the second graders of the experimental group. The results of the experimental group ($\chi 2(5)=16.24$, $p<0.01$) show the significant different proportions between the pre-test and post-test, while the control group ($\chi 2(5)=7.46$, $p>0.05$) has a similar percentage among the six difficulty levels. This change reveals that the personalized test affects the ability of the students in the experimental group. To further investigate the difference in the experimental group, a posteriori comparison reveals that the number of correctly answered questions with level two and level six in the post-test were statistically higher than those in the pre-test, whereas the number of questions with level one and level four in the post-test were significantly lower than those in the pre-test. This suggests that the number questions with higher difficulty level that were correctly answered increased after the personalized quiz strategy.

Table 1. Contingency tables for the number of correctly answered questions per difficulty level in the pretest and post-test

Difficulty Level		1	2	3	4	5	6
The number of questions		6	3	6	3	7	3
Control group	Pretest	69 (23.8%)	27 (9.3%)	63 (21.7%)	36 (12.4%)	68 (23.4%)	27 (9.3%)
	Post-test	73 (21.3%)	50 (14.6%)	72 (21.0%)	33 (9.6%)	71 (20.7%)	44 (12.8%)
Experimental group	Pretest	**248 (24.8%)**	99 (9.9%)	209 (20.9%)	**129 (12.9%)**	206 (20.6%)	108 (10.8%)
	Post-test	234 (20.5%)	**147 (13.1%)**	253 (22.6%)	106 (9.5%)	236 (21.1%)	142 (12.7%)

Table 2. Questionnaire results

	Items	**Mean**	**SD**
1	The reading interface is easy to use.	3.89	0.99
2	The test interface is easy to use.	3.86	0.95
3	Taking the quiz has helped me to evaluate my strengths and weaknesses..	4.00	0.67
4	Taking the quiz has helped me to identify areas of knowledge that need improvement.	4.03	0.64
5	Taking the quiz is useful preparation for exams.	3.89	0.7
6a	I clearly understood the vocabulary questions on the quiz.	3.27	0.99
6b	I clearly understood the grammar questions on the quiz.	3.46	0.99
6c	I clearly understood the reading comprehension questions on the quiz.	3.38	0.95

In terms of evaluating the performance of the automatic question generation, six questions in the questionnaire concerning the subjects' perception were investigated. Table 2 displays the detailed questions and shows their mean score and standard deviation. From the results, the quality of the interface and the functionality of the generated questions have high agreement. Most subjects agreed that the adaptive question selection strategy could

help them identify strengths and weaknesses, so that they could improve their skills and prepare well for exams. This data supports the performance of the proposed automatic question generation and represents the usefulness of the generated questions.

5. Conclusion

This paper presents a personalized automatic quiz generation model, which generates multiple-choice questions based on various difficulty levels and categories. This quiz generation technique is then paired with a quiz strategy to estimate an examinee's ability and to select suitable questions. Compared to the control group, the results show that the experimental group corrected their mistakes more frequently, answered more difficult questions correctly, and showed significant improvement between the score of the pre-test and post-test. In addition, the questionnaire results suggest most subjects support the functionality and quality of the proposed personalized automatic quiz generation.

Acknowledgment

This work was partially supported by National Science Council, Taiwan, with Grant No. 100-2511-S-008-005-MY3 and NSC97-2221-E-001-014-MY3.

References

[1] Barla, M., Bielikova, M., Ezzeddinne, A. B., Kramar, T., Simko, M. and Vozar, O. (2010). On the impact of adaptive test question selection for learning efficiency. *Computer & Education*, 55(2), 846-857.

[2] Brown, R. G. (2004). *Smoothing, Forecasting and Prediction of Discrete Time Series*. Dover Publications: Dover Phoenix Ed edition.

[3] Chen, C. Y., Ko, M. H., Wu, T. W. and Chang, J. S. (2005). FAST：Free Assistant of Structural Tests. *In Proceedings of the ROCLING 2005*.

[4] Embertson S., and Resise, S. (2000). *Item response theory for psychologists.* New Jersey, USA: Lawrence Erlbaum.

[5] Hoshino, A. and Nakagawa, H. (2007). Sakumon: An assistance system for English cloze test. *In Society for Information Technology & Teacher Education International Conference*.

[6] Huang, Y. T., Chang, H. S., Sun Y. and Chen M. C. (2011). A Robust Estimation Scheme of Reading Difficulty for Second Language Learners. *In Proceedings of the 11th IEEE International Conference on Advanced Learning Technologies*, 58-62.

[7] Levy, R. and Andrew G. (2006). Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In Proceeding of 5th International Conference on Language Resources and Evaluation.

[8] Lin, Y. C., Sung, L. C. and Chen, M. C. (2007). An Automatic Multiple-Choice Question Generation Scheme for English Adjective Understanding. *In Proceedings of the 15th International Conference on Computers in Education*, 137-142.

[9] Lin, Y. T., Chen, M. C. Sun, Y. S. (2009). Automatic Text-Coherence Question Generation Based on Coreference Resolution. *In Proceedings of the 17th International Conference on Computers in Education*.

[10] Mitkov, R. and Ha, L. A. (2003). Computer-Aided Generation of Multiple-Choice Tests. *In Workshop on Building Educational Applications Using Natural Language Processing*.

[11] Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D. and Manning, C. (2010). A Multi-Pass Sieve for Coreference Resolution. *In Procceding of EMNLP2010*.

[12] Wilson, K., Boyd, C., Chen, L. and Jamal, S. (2010). Improving student performance in a first-year geography course: Examining the importance of computer-assisted formative assessment. *Computers & Education*, 57 (2), 1493–1500.

[13] Yang, Y. C., Yang, C. F, Chang, C. M. and Chang, J. S. (2005). Computered-aid Reading Comprehension Automatic Quiz Generation. *In Proceedings of the ROCLING 2005*.