An Intelligent Testing Strategy for Vocabulary Assessment of Chinese Second Language Learners

Wei Zhou[1,3] **Renfen Hu**[1,2,⊠] **Feipeng Sun**[1,3] **Ronghuai Huang**[1,3]

{zhouwei, irishu, huangrh}@bnu.edu.cn
sunfeipeng@blcu.edu.cn

[1] National Engineering Laboratory for Cyberlearning and Intelligent Technology,
Beijing Normal University
[2] Institute of Chinese Information Processing, Beijing Normal University
[3] Faculty of Education, Bejing Normal University

Abstract

Testing is an important tool to monitor learning effects. However, it usually costs a large amount of time and human labor to build an item bank and to test large number of students. In this paper, we propose a novel testing strategy by combining automatic item generation (AIG) and computerized adaptive testing (CAT) in vocabulary assessment for Chinese L2 learners. Firstly, we generate three types of vocabulary questions by modeling both the vocabulary knowledge and learners' writing error data. After evaluation and calibration, we construct a balanced item pool with automatically generated items, and implement a three-parameter computerized adaptive test. We conduct manual item evaluation and online student tests in the experiments. The results show that the combination of AIG and CAT can construct test items efficiently and reduce test cost significantly. Also, the test result of CAT can provide valuable feedback to AIG algorithms.

1 Introduction

Vocabulary is one of the most important parts of language competence (Cook, 2016). Testing of vocabulary knowledge is central to research on reading and language (Brown et al., 2005). However, it usually costs a large amount of time and human labor to build an item bank and to test large number of students.

To enhance the testing efficiency and convenience, we propose a novel testing strategy by combining automatic item generation (AIG) and computerized adaptive testing (CAT). Based on this strategy, we build an online testing system to evaluate vocabulary knowledge of Chinese second language learners: http://test.aihanyu.org. The pipeline of our method is illustrated in Figure 1:

Step 1. Generate vocabulary questions automatically by modeling both the vocabulary knowledge and learners' writing error data.

Step 2. Construct a balanced item pool by sampling questions from different difficulty levels, and implement an online vocabulary test with these items.

Step 3. Conduct student tests in which students with different language proficiencies take both the online AIG test and a traditional student placement test developed by experts.

Step 4. Build an improved three-parameter CAT model with these items, and estimate the students' abilities.

In the experiments, the student tests demonstrate desirable results. Firstly, the scores of the online AIG test are strongly correlated with that of the placement test (ρ=0.8395). Secondly, the student abilities estimated by our CAT model reaches even stronger correlation with the placement test (ρ=0.8715). Meanwhile, the average test length decreases greatly by 81% (from 140 to 26).

The experiments show that our strategy can construct test items efficiently and reduce test cost significantly for both test developers and test takers. Also, the test result of CAT can provide valuable feedback to question generation and selection algorithms.

2 Related Work

2.1 Automatic Item Generation

Automatic item generation (AIG) is a promising approach to reduce the cost of test development. AIG methods have been used in generating different types of questions, such as reading comprehen-

Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 21–29
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics

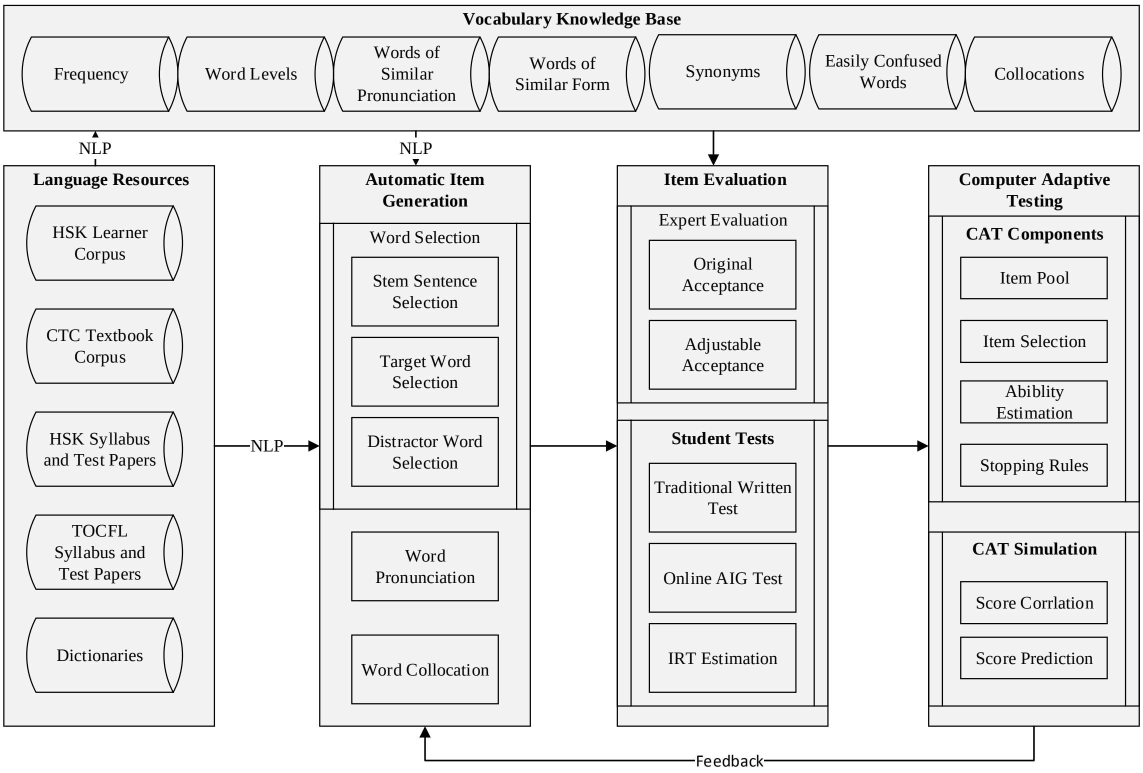


Figure 1: The pipeline of our testing strategy that combines AIG and CAT.

sion (Rus et al., 2007; Mostow et al., 2017) and vocabulary assessment (Mitkov et al., 2006, 2009; Aldabe and Maritxalar, 2014). Due to its high efficiency and controllability, automatic item generation has been used to create solutions and rationales for Computerized Formative Testing (Gierl and Lai, 2018).

For vocabulary testing, researchers have made a lot of efforts in generating vocabulary questions for ESL (English as a second language) learners (Mitkov and An Ha, 2003; Singh Bhatia et al., 2013; Correia et al., 2010; Takuya et al., 2010). It is well known that lexical knowledge vary a lot among different languages. For example, Chinese is a typical analytic language that lacks inflection. It mainly uses function words and word order to express grammatical information.

In the area of Chinese item generation, some methods have been proposed to generate factual questions and character questions (Liu et al., 2017; Ding and Gu, 2010; Liu et al., 2018). Different from existing work, this paper focuses on the generation of vocabulary questions, and utilizes them in vocabulary assessment of CSL (Chinese as a second language) learners. To enhance the test efficiency, we also integrate these automatically generated items into a computerized adaptive testing (CAT) model.

2.2 Computerized Adaptive Testing

With the development of language testing technologies, computerized adaptive testing (CAT) has attracted considerable attention in language testing area and has been successfully applied to large-scale standardized language tests, such as GRE and GMAT (Chang, 2015). Instead of giving all the examinees the same fixed test, CAT selects items that are tailored to each examinee's ability. Compared with traditional computer based or paper-pencil based tests, CAT can greatly shorten the test length by 50% while maintaining good test reliability and increasing the test security (Wainer, 2000; Weiss and Kingsbury, 1984).

However, one of the main challenges in CAT is the item pool development which requires not only large numbers of high-quality test items, but also a careful calibration of these items. In this study, we propose to construct the item pool with automatically generated questions. It can reduce the test cost significantly for both test developers and test takers.

3 Automatic Generation of Vocabulary Questions

To test the vocabulary knowledge of CSL learners, we generate three types of multiple-choice questions which account for different dimensions of vocabulary knowledge. The question examples can be seen in Figure 2.

(1) Word selection: Select a word that can fill in the blank of the sentence. It involves the knowledge of word form, meaning and how it is used in the context.

(2) Word pronunciation: Select a word that has an incorrect pinyin label. It focuses on the pronunciation part.

(3) Word collocation: Select a word that can collocate with the given word. It addresses the syntactic behaviors and collocational knowledge of words.

The generation of the vocabulary questions involves two stages: (1) Build a vocabulary knowledge base by extracting features from learner corpus, textbook corpus, test papers and dictionaries. (2) Generate different types of questions via stem selection, target word selection and distractor selection.

3.1 Vocabulary Knowledge Base

The knowledge base contains totally 8,400 word entries, which are collected from the syllabuses of two official Chinese language proficiency tests: HSK[1] and TOCFL[2]. We build a list of attributes for each entry in the knowledge base, and the attribute values are automatically extracted from large-scale language resources with multiple natural language processing (NLP) methods:

- Word frequency: It is calculated from CTC[3], a text corpus for Chinese L2 learners.
- Word level: The 8400 target words are scaled to 14 difficulty levels according to their frequencies in CTC, i.e. 600 words at each level.
- Words of similar pronunciation: They are extracted with the pronunciation similarity model proposed by Hu (2013).
- Words of similar form: If two words are of equal length in Chinese characters (hanzi) and have at least one same character, we count them as words of similar form.
- Synonyms: They are retrieved from Yang and Jia (2005)'s synonym dictionary.
- Easily confused words: They are extracted from leaners' writing error, as collected and manually labeled in HSK learner corpus[4]. If word *a* is involved in word selection error for at least 10 times in the learner corpus, and it is mistakenly used as word *b* for over 20% of the error cases, we identify word *b* as an easily confused word of *a*.
- Collocations: Nine types of collocations are retrieved from the collocation knowledge base built by Hu et al. (2016)[5].

3.2 Item Generation

3.2.1 Word Selection Question

The model generate the word selection questions via four steps: preprocessing, stem sentence selection, target word selection and question generation.

Firstly, all the texts in CTC are preprocessed via word segmentation, POS tagging and dependency parsing with LTP-Cloud (Che et al., 2010), a Chinese NLP toolkit. We obtain 2.4 million words and 154,023 dependency trees after the preprocessing.

Secondly, sentences are selected based on the NLP preprocessing results if they can satisfy multiple conditions, including sentence length, sentence independence and difficulty levels. We limit the sentence length to 10-30 words. For independence analysis, we target at sentences whose meanings are context independent, i.e. a complete declarative sentence which is not from a dialogue, and does not involve a pronoun that refers to someone or something in the previous context. We compile 3 rules based on POS tags and dependency relations to exclude unqualified sentences. For difficulty levels, we check if each word of the sentence is in our 8400-word vocabulary for L2 learners, and the percent of OOV (out-of-vocabulary) words should not exceed 10%.

Thirdly, we locate candidate target words in the stem sentences. Each candidate word should appear only once in the sentence and have at least three distractors in the vocabulary knowledge base. The distractors include words of similar pronunciation and form, as well as easily confused words. If more than one candidate target words are

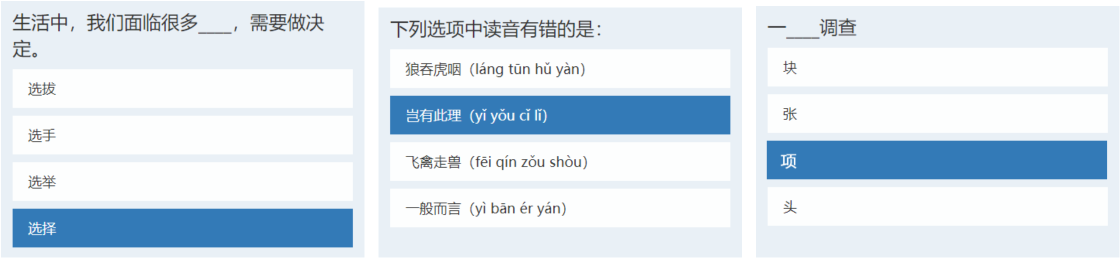
[1] http://www.chinesetest.cn/godownload.do

[2] http://www.tw.org/tocfl/

[3] http://www.aihanyu.org/basic_v2/index.html

[4] http://bcc.blcu.edu.cn/hsk

[5] http://cca.xingtanlu.cn/



生活中，我们面临很多___，需要做决定。

选拔

选手

选举

选择

(a)

下列选项中读音有错的是：

狼吞虎咽（láng tūn hǔ yàn)

岂有此理（yǐ yǒu cǐ lǐ)

飞禽走兽（fēi qín zǒu shòu)

一般而言（yì bān ér yán)

(b)

一___调查

块

张

项

头

(c)

Figure 2: Examples of automatically generated items as they shown in the online testing application. (a) Word Selection, (b) Word Pronunciation, (c) Word Collocation. The highlighted option is the correct answer.

retrieved, we choose the one with higher difficulty, i.e. lower frequency. If a target word has more than three distractors, we choose the distractors that have the most similar difficulty levels with the target words.

At last, the target word is removed to generate a fill-in-blank question. Three distractors and the target word are shuffled to construct four options.

3.2.2 Word Pronunciation Question

A target word is firstly selected if one of its characters has an easily confused pronunciation determined by the pronunciation similarity model (Hu, 2013). We replace the correct pinyin with an easily confused one, and choose three other words from the same difficulty level that have correct pinyin labels and the same length. The item stem is *"Select the word that has an incorrect pinyin label"*.

3.2.3 Word Collocation Question

For word collocation question, we firstly retrieve the collocations of frequency > 3 and mutual information > 0 for each target word. Given a target word and its collocation, we obtain candidate distractors from the vocabulary knowledge base. To ensure there is only one correct answer in the multiple-choice question, we replace the target word with each candidate distractor to constitute a new combination. If the new combination does not appear in our collocation data, this candidate distractor is accepted. If more than three distractors are accepted, we choose the ones that have the most similar difficulty levels with the target word. At last, the target word is removed and we generate the question similarly to the word selection question.

Three types and totally 93764 test items are successfully generated with our method, including 75689 items for word selection, 6697 items for word pronunciation and 11378 items for word collocation. After that, we sample questions for manual evaluation. The results will be discussed in Section 5.

4 Computerized Adaptive Testing

This paper aims at building a CAT model to evaluate vocabulary knowledge of CSL learners. We use the automatically generated questions for item calibration. The advantage is we can directly sample questions from different difficulty levels, so as to build a balanced item bank. In this study, item response theory (IRT) with three-parameter is used for calibration.

4.1 Theoretical Basis

Let $p_i(\theta)$ be the probability of a correct response to item i from a examinee with ability θ, thus $q_i(\theta) = 1 - p_i(\theta)$ is the probability of a incorrect response.

Let $u = (u_1, u_2, ..., u_n), u_i \in \{0, 1\}$ is the responses of n items. The likelihood function L is given by Equation 1.

$$L(u|\theta) = \prod_{i=1}^{n} p_i(\theta)^{u_i} q_i(\theta)^{1-u_i} \qquad (1)$$

Equation 2 gives the probability of a correct response to item i, where a_i is discrimination parameter, b_i is difficulty parameter, and c_i is the guessing parameter.

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}} \qquad (2)$$

Solving $L'(\theta) = 0$ can find the value of $\hat{\theta}$ that maximize the likelihood function L. To simplify, we transform it to a log-likelihood function $l(u|\theta) = ln(L(u|\theta))$ as shown in Equation 3. The logarithm function could convert the product of factors to a sum of log factors, which makes it much easier to get the derivative.

$$l(u|\theta) = \sum_{i=1}^{n}(u_i \ln p_i(\theta) + (1-u_i)\ln q_i(\theta)) \quad (3)$$

Thus, to find the $\hat{\theta}$ that maximize L, it is equivalent to solve $l'(\theta) = 0$. It can be computed by the Newton-Raphson method: $\theta_{t+1} = \theta_t - \frac{l'(\theta)}{l''(\theta)}$, which is an iterative algorithm with termination criterion ϵ, t_{max} s.t. $\Delta = \theta_{t+1} - \theta_t < \epsilon \vee t > t_{max}$. A simplified iterative formula is given by Equation 4 (Baker, 2001).

$$\theta_{t+1} = \theta_t + \frac{\sum_{i=1}^{n} a_i(u_i - p_i(\theta_t))}{\sum_{i=1}^{n} a_i^2 p_i(\theta_t) q_i(\theta_t)} \quad (4)$$

The information function is given by Equation 5. $I_i(\theta)$ is the amount of information for item i at ability θ.

$$I_i(\theta) = a_i^2 \frac{(p_i(\theta) - c_i)^2}{(1-c_i)^2} \frac{q_i(\theta)}{p_i(\theta)} \quad (5)$$

The test information function is given by Equation 6. It is the sum of information for all items in the test.

$$TI(\theta) = \sum_{i=1}^{n} I_i(\theta) \quad (6)$$

The standard error function is given by Equation 7. A higher test information TI implies the higher precision of estimated ability which can not be observed directly. Thus, the smaller SE is, the better estimation is. A threshold of SE acts as a termination criteria in the test.

$$SE(\theta) = \frac{1}{\sqrt{TI(\theta)}} \quad (7)$$

4.2 Adaptive Algorithm

There are four important components in an adaptive testing algorithm: the item pool, the item selection, the ability estimation and stopping rules (Weiss and Kingsbury, 1984).

Item Pool. The success of CAT is highly dependent on the item pool with sufficient items of different difficulty levels. Our AIG method enables the system to select as many items as the test needs from different levels. Thus, a balanced item pool can be easily sampled for calibration.

Item Selection. The item selection is to select an item with the highest information I at the estimated ability $\hat{\theta}$. The test normally starts with an item of medium difficulty. And items can not be repeated during the test.

Ability Estimation. After each item is answered, the examinee's ability is estimated and used by the item selection to select the next item. The most commonly used estimation method is maximum likelihood estimation (MLE). Maximum likelihood is asymptotically unbiased, but it can not provide an ability estimate for a homogeneous set of responses (all correct or all incorrect) (Weiss and Kingsbury, 1984). To address this problem, we set a bound of ability $[\theta_{min}, \theta_{max}]$ to enable Newton-Raphson method to convergence to the bound. The iterative ability estimation process is shown in Algorithm 1.

Stopping Rules. After the ability is estimated, the standard error SE is calculated to determine whether a new item must be selected or the test should be terminated. We implement three stopping rules: the test reaches the maximum length n_{max}, the ability reaches the boundary $[\theta_{min}, \theta_{max}]$ for five consecutive questions when more than 15 items are administrated, or the examinee's standard error SE falls below the threshold s.

5 Experimental Analysis

We evaluate our method via three experiments: (1) Evaluate the automatically generated items manually. (2) Conduct student test with both an online AIG test and a traditional written test developed by CSL teachers. (3) Use CAT model to estimate the students' abilities.

5.1 Expert Evaluation of AIG

To assess the students' vocabulary knowledge, we generate three types and totally 93764 test items. After that, we randomly sample 100 items for each type of question, resulting in 300 items in total. These questions are used for manual evaluation. Original Acceptance Rate (OAR) and Adjustable Acceptance Rate (AAR) are calculated. An item can be originally accepted if two professional CSL teachers both agree that this item can be directly used in a vocabulary test. And it can be an adjustable item if the teachers both agree that it only needs a few simple modifications, i.e. the replacement or deletion of less than 2 words.

The evaluation results are shown in Table 1. The question generation method performs well with the average OAR of 53% and the AAR of 81.67%.

It is noteworthy that the acceptance rate varies a lot among three types of questions. Word pro-

Algorithm 1 Estimate($B, \theta_0, s, n_{max}, \theta_{min}, \theta_{max}, t_{max}, \epsilon$)

```
Set $n = 0$
Set $A = \emptyset$
Set $T = []$
Set $U = []$
Set $\hat{\theta} = \theta_0$
while $n < n_{max} \wedge SE(\hat{\theta}) \geq s$ do
  Set $n = n + 1$
  Find item $x$ st. $x \in B \wedge x \notin A \wedge I_x(\hat{\theta}) = \max_{y \notin A} I_y(\hat{\theta})$
  Add$(A, x)$
  if test taker's answer to item $x$ is correct then
    Add$(U, 1)$
  else
    Add$(U, 0)$
  end if
  Set $t = 0$
  repeat
    Set $t = t + 1$
    Set $\theta_{tmp} = \hat{\theta}$
    Update $\hat{\theta}$ using Equation 4
    Set $\Delta = \left|\hat{\theta} - \theta_{tmp}\right|$
  until $\Delta < \epsilon \vee t > t_{max} \vee \theta_t \notin [\theta_{min}, \theta_{max}]$
  Set $\hat{\theta} = \max(\min(\hat{\theta}, \theta_{max}), \theta_{min})$
  Add$(T, \hat{\theta})$
  if $n > 15 \wedge (\min(\text{Last}(T, 5)) = \theta_{max} \vee \max(\text{Last}(T, 5)) = \theta_{min})$ then
    break while
  end if
end while
return $\hat{\theta}$
```

nunciation question performs best since it focuses only on the pinyin label, and its generation module is very simple. The generation of word selection questions is much more complicated. It involves appropriate selection of sentences, target words and distractors. Word collocation question can be considered as a simplified version of word selection question. We further analyze the feedback of the teachers, and find that the distractor selection works very well, indicating that our vocabulary knowledge base has a high quality. Meanwhile, the stem sentence selection and target word selection algorithms needs further improvement on both difficulty control and semantic analysis.

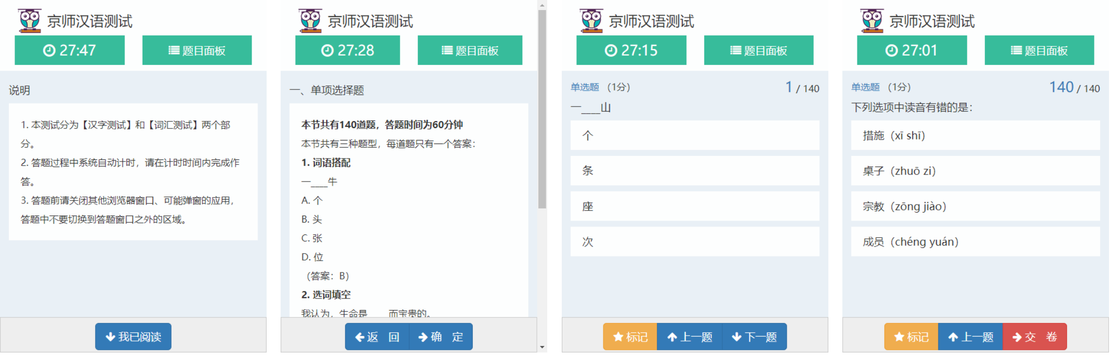
5.2 Online AIG Test

We build an online vocabulary test with accepted vocabulary questions of three types. Specifically, we select 140 questions from 14 word levels, i.e. 10 questions at each level. These questions are manually reviewed and adjusted to ensure they can be used in the student test. The score for each question is one point, thus, the test score equals the number of questions answered correctly. The vocabulary size of each student can be estimated with the method proposed by Beglar and Nation (2007). Since each level has 600 words, a student's test score will be multiplied by 60 to get their total receptive vocabulary size. The interfaces of the online testing system can be seen in Figure 3.

155 international students of different language proficiencies are organized to take a traditional written test of 90 minutes and the AIG online test of 30 minutes. The written test is a student placement test including listening, reading and writing questions constructed by professional CSL teachers. And the online test only includes vocabulary questions. These two tests are administrated on the same day to ensure the examinees' language

Table 1: Results of Expert Evaluation

Result	Word Selection	Word Pronunciation	Word Collocation	Average
OAR	19%	100%	40%	53%
AAR	65%	100%	80%	81.67%


京师汉语测试

27:47 题目面板

说明

1. 本测试分为【汉字测试】和【词汇测试】两个部分。

2. 答题过程中系统自动计时，请在计时时间内完成作答。

3. 答题前请关闭其他浏览器窗口、可能弹窗的应用，答题中不要切换到答题窗口之外的区域。

我已阅读

(a)

京师汉语测试

27:28 题目面板

一、单项选择题

本节共有140道题，答题时间为60分钟

本节共有三种题型，每道题只有一个答案：

1. 词语搭配

一___牛

A. 个

B. 头

C. 张

D. 位

（答案：B）

2. 选词填空

我认为，生命是____而宝贵的。

返 回 确 定

(b)

京师汉语测试

27:15 题目面板

单选题（1分） 1 / 140

一___山

个

条

座

次

标记 上一题 下一题

(c)

京师汉语测试

27:01 题目面板

单选题（1分） 140 / 140

下列选项中读音有错的是：

措施（xī shī）

桌子（zhuō zi）

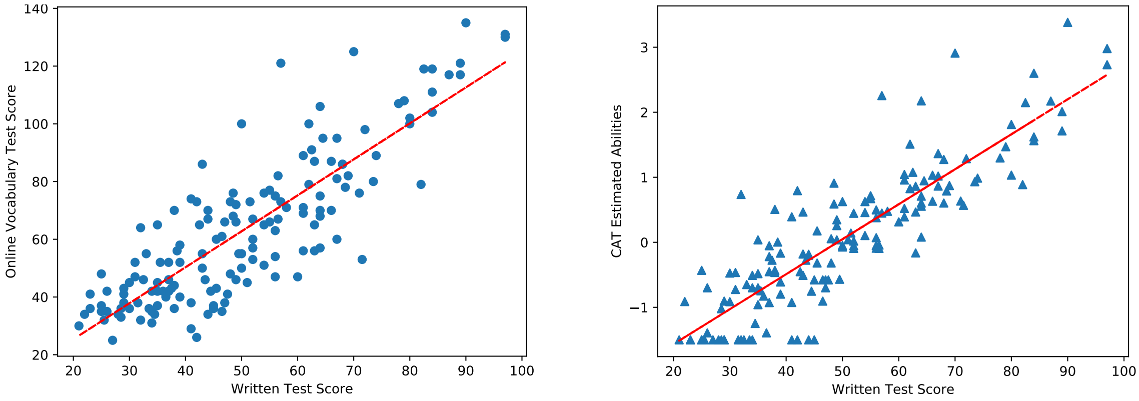
宗教（zōng jiào）

成员（chéng yuán）

标记 上一题 交 卷

(d)

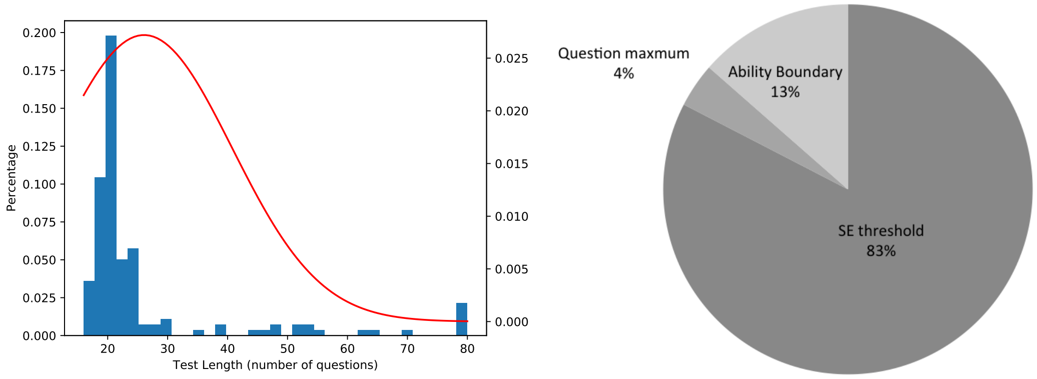
Figure 3: The online testing system on mobile devices (a) description of the test, (b) examples of test items, (c) the first item, (d) the last item.



(a)

(b)

Figure 4: Score Correlations. (a) Written test score and online AIG test score, $\rho = 0.8395$; (b) Written test score and CAT estimated ability, $\rho = 0.8715$.



(a)

(b)

Figure 5: CAT Simulation Results. (a) Test Length, the average length is 26; (b) Percentage of different stopping rules.

proficiencies are stable.

After the tests, we compute the correlation scores of them. As shown in Figure 4(a), the result is very inspiring that the scores are strongly correlated with Pearson correlation coefficient of 0.8395, given we only use AIG based vocabulary questions. Furthermore, the test time is greatly reduced from 90 minutes to 30 minutes. The online AIG test promisingly indicates that:

- Vocabulary knowledge is indeed a core part of second language proficiency, as stated in previous works (Nation, 2001; Cook, 2016).

- AIG is an effective tool for vocabulary assessment.

5.3 CAT Simulation

After the online test, we collect students' answer data, and estimate three parameters for each item, including difficulty parameter b, discrimination parameter a and guessing parameter c. The estimation is based on 3PL item response theory (IRT) and implemented with the R package ltm.

With this calibrated item pool, we implement the adaptive algorithm illustrated in Algorithm 1. The detailed parameter settings are as following: $\theta_0 = 0.2$, $s = 0.3$, $n_{max} = 80$, $[\theta_{min}, \theta_{max}] = [-1.5, 4.5]$, $t_{max} = 80$, $\epsilon = 0.0001$.

We simulate the CAT based vocabulary test with the 155 students' answers, and output estimated abilities when one of the stopping rules is triggered.

Figure 4(b) shows that the estimated abilities reaches an even higher correlation coefficient of $\rho = 0.8715$ than the fixed online AIG test. Meanwhile, the average test length is only 26, which decreases greatly by 81% compared to 140 of the AIG test.

Figure 5 further illustrates the CAT simulation result. Regarding the triggered stopping rules, 83% of the students end with the standard deviation threshold, which indicates that our CAT algorithm has a desirable estimate precision. However, there are still 13% of students end with the lower ability boundary, and 4% of students stop with maximum test length. These cases reflect that our item pool needs improvement by adding more very simple questions for low ability students and very hard questions for high ability students. It is an important feedback to the AIG algorithm, especially on the difficulty control and sampling method.

5.4 Vocabulary Size and Score Prediction

After estimating students' vocabulary abilities with CAT, we train a linear regression model to predict a student's vocabulary size and the written test score.

The vocabulary size vs is predicted with Equation 8.

$$vs = 60 \times (22.37\,\theta + 61.43), R^2 = 0.8505 \quad (8)$$

It has been implemented on our online testing system `http://test.aihanyu.org`. Users can quickly estimate their vocabulary sizes after taking a CAT test in a few minutes.

The written test score sc can be computed with Equation 9. The result could serve as an effective tool for student placement.

$$sc = 14.10\,\theta + 49.46, R^2 = 0.7594 \quad (9)$$

6 Conclusions and Future Work

In this paper, we propose a novel testing strategy by combining automatic item generation (AIG) and computerized adaptive testing (CAT) in vocabulary assessment. Experiments show that it is a promising and highly effective path to evaluate language proficiency. The advantages are obvious as below:

- AIG is an effective method to construct a balanced CAT item pool.

- CAT is also a good evaluation tool of AIG, since it can provide important feedback to AIG which is hard to be given by manual evaluation.

- The combination of AIG and CAT can reduce the test cost significantly.

We believe that this testing strategy can serve as a good basis for research of language testing, as well as various intelligent learning applications that need students' proficiencies for user modeling. In the future, we aim at enhancing the AIG algorithms and exploring the generation algorithms of more question types, as well as in more disciplines.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities, National Social Science Fund of China

(No. 18CYY029) and China Postdoctoral Science Foundation funded project (No. 2018M630095).

References

Itziar Aldabe and Montse Maritxalar. 2014. Semantic similarity measures for the generation of science tests in basque. *IEEE transactions on Learning Technologies*, 7(4):375–387.

Frank B Baker. 2001. *The basics of item response theory*. ERIC.

David Beglar and P Nation. 2007. A vocabulary size test. *The Language Teacher*, 31(7):9–13.

Jonathan C Brown, Gwen A Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 819–826. Association for Computational Linguistics.

Hua-Hua Chang. 2015. Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1):1–20.

Wanxiang Che, Zhenghua Li, and Ting Liu. 2010. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics.

Vivian Cook. 2016. *Second language learning and language teaching*. Routledge.

Rui Correia, Jorge Baptista, Nuno Mamede, Isabel Trancoso, and Maxine Eskenazi. 2010. Automatic generation of cloze question distractors. In *Proceedings of the Interspeech Satellite Workshop on Second Language Studies:Acquisition,Learning,Education and Technology*, Waseda University,Tokyo,Japan.

Xiang-Min Ding and Hong-Bin Gu. 2010. Automatic generation technology of chinese multiple-choice items based on ontology. *Computer Engineering and Design*, 31(6):1397–1400.

Mark J Gierl and Hollis Lai. 2018. Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied psychological measurement*, 42(1):42–57.

Renfen Hu. 2013. Research on phonetic symbols of phonograms in chinese mandarin. *Journal of Chinese Information Processing*, 27(3):41–47.

Renfen Hu, Jiayong Chen, and Kuang-hua Chen. 2016. The construction of a chinese collocational knowledge resource and its application for second language acquisition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3254–3263.

Ming Liu, Vasile Rus, and Li Liu. 2017. Automatic chinese factual question generation. *IEEE Transactions on Learning Technologies*, 10(2):194–204.

Ming Liu, Vasile Rus, and Li Liu. 2018. Automatic chinese multiple choice question generation using mixed similarity strategy. *IEEE Transactions on Learning Technologies*, 11(2):193–202.

Ruslan Mitkov and Le An Ha. 2003. Computer-aided generation of multiple-choice tests. volume 2, pages 17–22.

Ruslan Mitkov, Le An Ha, Andrea Varga, and Luz Rello. 2009. Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 49–56. Association for Computational Linguistics.

Ruslan Mitkov, Ha Le An, and Nikiforos Karamanis. 2006. A computer-aided environment for generating multiple-choice test items. *Natural language engineering*, 12(2):177–194.

Jack Mostow, Yi-Ting Huang, Hyeju Jang, Anders Weinstein, Joe Valeri, and Donna Gates. 2017. Developing, evaluating, and refining an automatic generator of diagnostic multiple choice cloze questions to assess children's comprehension while reading. *Natural Language Engineering*, 23(2):245–294.

Paul Nation. 2001. *Learning Vocabulary in Another Language*. Cambridge University Press.

Vasile Rus, Zhiqiang Cai, and Arthur C Graesser. 2007. Experiments on generating questions about facts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 444–455. Springer.

Arjun Singh Bhatia, Manas Kirti, and Sujan Kumar Saha. 2013. Automatic generation of multiple choice questions using wikipedia. In *Pattern Recognition and Machine Intelligence*, pages 733–738, Berlin, Heidelberg. Springer Berlin Heidelberg.

Goto Takuya, Kojiri Tomoko, Toyohide Watanabe, Iwata Tomoharu, and Yamada Takeshi. 2010. Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning : an International Journal*, 2.

Howard Wainer. 2000. *Computerized Adaptive Testing: A Primer*. Lawrence Erlbaum Associates.

David J Weiss and G Gage Kingsbury. 1984. Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4):361–375.

Jizhou Yang and Yongfen Jia. 2005. *The usage comparisons of 1700 pairs of synonyms*. Beijing Language and Culture University, Beijing, China.