論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Automatic English Vocabulary Question Generation for Efficient Measurement of Learner Proficiency
著者(和文)	YuniSusanti
Author(English)	Yuni Susanti
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10995号, 授与年月日:2018年9月20日, 学位の種別:課程博士, 審査員:徳永 健伸,岡崎 直観,宮崎 純,村田 剛志,藤井 敦,宇佐美 慧
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10995号, Conferred date:2018/9/20, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Automatic English Vocabulary Question Generation for Efficient Measurement of Learner Proficiency

Yuni Susanti
Department of Computer Science
Tokyo Institute of Technology
susanti.y.aa@m.titech.ac.jp

Doctor's Thesis
June, 2018

Abstract

Conducting a language test is indispensable to evaluate the proficiency of the language learners. However, manual construction of questions is a difficult task that requires a high level of skill from experts. Hence, an automatic question generation system can be a breakthrough by assisting the experts in making questions; thus, it makes the question construction easier. This thesis, consisting of seven chapters, presents a study on automatic generation of multiple-choice English vocabulary questions for efficient measurement of language learner proficiency. It consists of four topics: (1) automatic question generation (AQG), (2) distractor improvement, (3) question difficulty control and (4) integration of AQG into the computerised adaptive test (CAT). We proposed a new method for each topic and conducted evaluations involving learners, teachers and experts.

In the first topic, we proposed a novel method for automatically generating English vocabulary questions, modelling the generated questions after the TOEFL vocabulary questions. In this type of question, determining the word sense of the target word in a reading passage is crucial to creating the question options (the correct answer and distractors). We could use word sense disambiguation (WSD) techniques to identify the word sense. However, the accuracy of the state-of-the-art WSD method remains about 70-80%, which is far from satisfactory to the question generation task. Thus, we proposed a method that avoids word sense disambiguation. Instead, we took an information retrieval approach where given a target word and one of its word sense, we search a passage that uses the target word with the given word sense. We conducted two kinds of evaluation for assessing the quality of the generated questions: 1) test taker-based evaluation and 2) expert-based evaluation. In the test taker-based evaluation, we administered the machine-generated questions together with human-made questions to the real students. Both evaluations showed that the machine-generated questions have a comparable ability to the human-made questions in measuring the student proficiency, and the English teachers were not able to distinguish more than half of the machine generated-questions form human-made questions.

Through the evaluation, we found that distractors are the primary source of low-

quality machine-generated questions. Thus, the second topic focuses on improving the distractors. We proposed a new method to generate distractors that aggregates both semantic similarity and word collocation information. The method finds distractors which are close to the target word but far from the correct answer in their meaning, and also collocate with the adjacent words of the target word in the given context (the reading passage). The evaluation showed that the proposed method removes the problematic distractor candidates better than the baseline, and the generated distractors have comparable quality to the original human-made distractors. A further error analysis showed that we could use the problematic distractors generated by the proposed method for a real test despite their low score by the human expert.

Toward an efficient measurement of the language learner proficiency, we proposed to integrate the AQG with CAT, which presents items tailored to the test taker proficiency, e.g. the item difficulty suits to the test taker proficiency. Therefore, CAT needs a big collection of items with their item difficulty known in advance. The conventional CAT estimates item difficulty from the test taker's responses in a pretesting phase. However, this process is expensive and poses a risk of exposing the item before the real test. To cope with this problem, we proposed to control the difficulty of the generated question with the three predetermined factors: 1) target word difficulty (TWD), 2) similarity between the correct answer and distractors (SIM) and 3) distractor word difficulty level (DWD). The analysis of test taker-based evaluation revealed that we could control the item difficulty with the predetermined factors, and the SIM factor contributes the most to the item difficulty.

We conducted simulation-based experiments on the AQG and CAT integration using two types of item difficulty, i.e. the item difficulty estimated from the test taker's responses and that controlled during the question generation process, in Chapter 6. We evaluated the performance of the simulations by looking at the mean squared error (MSE) between the true proficiency of the test takers and the proficiency estimated by each simulation. The result showed that all proposed CAT simulations with the controlled item difficulty yielded smaller MSEs than the baseline (a linear test) simulation. Moreover, their MSEs were close to the MSE of the gold standard, i.e. the CAT simulation with the estimated item difficulty by test taker responses. This is an encouraging result showing a possibility of integrating the CAT and AQG with controlling item difficulty, which can eliminate the pretesting.

Acknowledgments

To my professor and my parents,
for their help, support, and patience.

Contents

List of Figures

Chapter 1

Introduction

The research in computer-assisted language testing is an increasing field of study which has attracted much theoretical and empirical work in the last decades. According to Cotton (1988), classroom teachers spend anywhere from 35% to 50% of their instructional time conducting questioning, or testing sessions. Questioning in the form of a written test is a common method to evaluate learner's knowledge or ability on a specific field, including language proficiency. Multiple-choice and open-ended questions (why, what, how and others) are two of the most popular types of questions for language proficiency evaluation.

Regarding the language itself, one of the most widely learnt second-languages is English. As the demands of communication across diverse communities have been developing in recent years, the use of English as the primary international language has increased to enable this interaction between different societies both in business and academic settings. Owing to this, English proficiency tests such as TOEIC®and TOEFL®are imperative in measuring English communication skill of a non-native English speaker. However, manual construction of questions for language proficiency tests requires a high level of skill and is a laborious and time-consuming task as well. Recent research has investigated how natural language processing techniques can contribute to generating questions automatically and this kind of research has received immense attention lately.

Since the past questions of standardised English proficiency tests are not freely distributed, test takers can only rely on a limited number of test samples and preparation books to study. Providing test takers with a rich resource of English proficiency test questions is one of the main motivations of this research. However, generating an unlimited number of questions should not be the only objective of the automatic question generation (AQG) research. It is also important to guarantee the quality of generated questions; otherwise those questions cannot be used for its intended purposes. Therefore, we conducted

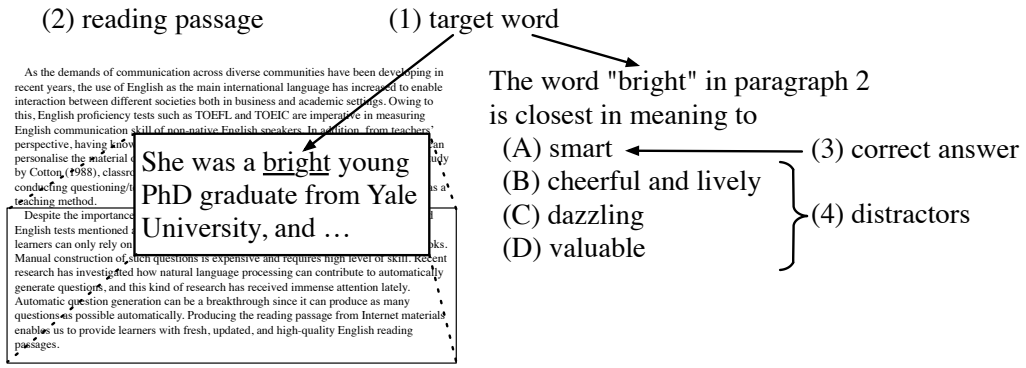

Figure 1.1: Four components of question in closest-in-meaning vocabulary question

a variety of evaluation involving English learners, teachers and professional item writers. The present study focuses on English, but the method and evaluation can be applied to any languages

In this thesis, we focus on generating multiple-choice vocabulary questions since it contributes to the majority of questions in the TOEFL®iBT[1] reading section (three to five questions out of a total of 12 to 14 in one reading passage) and it appears in other English proficiency tests such as TOEIC®as well.

TOEFL®vocabulary questions, which ask for the closest option in meaning to a given English word, is adopted as the model of vocabulary question in this work. As shown in Figure 1.1, this type of question is composed of four components: 1) a target word, 2) a reading passage in which the target word appears, 3) a correct answer and 4) distractors (incorrect options). To generate a question, we need to produce these four components.

One possible approach for generating such questions is using a manually-created lexical knowledge base such as WordNet (Fellbaum, 1998), which holds not only word glosses (definition), but also other information such as example sentences of the word, its synonyms, antonyms, hyponyms, hypernyms and so forth. Brown et al. (2005) generated multiple-choice questions by taking their components from WordNet. Lin et al. (2007) also adopted WordNet to produce English adjective questions from a given text. The candidates of options (correct answer and distractors) are taken from WordNet and filtered by Web searching. Unlike previous work, we propose a method for automatic question generation by utilising Web texts from the Internet in addition to information from WordNet. Producing the reading passage from Internet materials enables us to provide learners with fresh, updated and high-quality English reading passages.

Toward the efficient measurement of learner proficiency, we assessed the feasibility of

[1]TOEFL®iBT is an Internet-based test version of TOEFL®(www.ets.org/toefl)

integrating the AQG into a computerised adaptive test (CAT). The CAT aims at a precise and reliable measure of a test taker's proficiency or skill, by presenting items[2] that are appropriate to the test taker's proficiency (Linden and Glas, 2000). The CAT evaluates the test taker's proficiency after the response of each item and updates the estimated proficiency to select the next item to present the test taker. This can subside, or even eliminate the drawback of the conventional paper-and-pencil test (hereon called linear test) where the test takers answer the same set of items in the same order regardless of the difference of their proficiency. For instance, high-proficiency test takers might get bored of answering a whole test if it contains only items that they consider easy. On the contrary, low-proficiency test takers might get frustrated over difficult items and might give up on working on the test seriously. As the CAT presents items tailored to the test taker‘s ability, it reduces the frustration of the test takers to improve the reliability of the proficiency measurement.

However, successful implementation of CAT often relies on a large collection of previously administered items called the item bank. The item bank consists of items with their item parameters[3] estimated from the test takers‘ responses in a pretesting phase. Estimating the item parameters is called item calibration. Thus, CAT itself leads to a considerable cost in the item development, pretesting, and item calibration processes (Veldkamp and Matteucci, 2013). Also, conducting a pretesting poses a risk of exposing the items before they are used in a real test.

Integrating CAT with an AQG could possibly mitigate the problems of costly item development in CAT since the AQG can produce as many questions as needed. The present study focuses on integrating CAT with an AQG system without any item calibration process. In other words, the item parameters, e.g. the item difficulty, should be estimated in advance during the question generation process so that there is no need to administer the items in a pretesting. We propose a method that controls the item difficulty using three predetermined factors related to the question components: 1) target word difficulty (TWD), 2) semantic similarity between the correct answer and distractors (SIM) and 3) distractor word difficulty level (DWD). We generate items with various levels of difficulty by the combinations of these three factors. To evaluate a feasibility of the integration of CAT and AQG, we conduct a simulation-based experiment.

The contributions of this thesis are as follows.

1. We proposed a method to collect the reading passage and word sense pairs, which is essential to generate closest-in-meaning vocabulary questions without word sense

[2]From hereon, the term "item" is used interchangeably with "question"

[3]e.g. item difficulty, item discrimination, etc.

disambiguation technique. The proposed method combined with word sense disambiguation technique further improves the accuracy of the collection of the pairs (Chapter 3).

2. We proposed a method to rank the distractor candidates utilising word-embedding based semantic similarity and collocation, which is superior to the state-of-the-art method in producing less problematic distractors (Chapter 4).

3. We proposed a method to control the item difficulty using predetermined factors related to the question components. The ability to control the item difficulty is important for the integration of AQG and CAT (Chapter 5).

4. We proposed AQG and CAT integration with predetermined item difficulty and validated the feasibility through a simulation-based experiment using real data collected by administering the generated items to English learners (Chapter 6).

5. We proposed an evaluation method of generated questions from two perspectives, i.e. test taker's and examiner's (item writer, teachers) perspectives. Analysing the results froms different viewpoints and their interaction enable us to understand the characteristics of the automatically generated questions (Chapter 3-Chapter 5).

The remainder of this thesis is organised as follows. The next chapter presents the related work to this study, followed by thorough description on the proposed AQG method including the result and discussion of the evaluation in Chapter 3. Chapter 4 describes the method on improving the quality of distractors and its evaluation. Chapter 5 describes the method and discussion on controlling the item difficulty in AQG, followed by a discussion on the integration of AQG and CAT through simulation-based experiment in Chapter 6. Finally, we conclude the thesis in Chapter 7.

Chapter 2

Related work

This chapter surveys previous work on automatic question generation in language learning as well as the history of the field. We also provide a brief description on computerised adaptive test, followed by the related work on the integration of AQG and CAT. We further provide the description of Neural Test Theory, which is important for the evaluation in this thesis.

2.1 Automatic question generation

Studies in automatic question generation date back to the late of the 20th century. Wolfe (1976) introduced an experimental computer-based educational system called AUTOQUEST for assisting independent study of written text as one of the initial researches in this field. As he claims, students improve their reading comprehension ability by being periodically asked questions about what they read but a considerable human effort is needed to prepare the questions. Another study is by Coniam (1997), focusing on automatic generation of English cloze questions (fill-in-the-blank) using a large corpus of word frequency data. Since then, automatic question generation, particularly for the language learning purpose, is an emerging application because it has become possible with the availability of technologies in natural language processing such as WordNet as a machine-readable lexical dictionary.

Various types of question for assessing different skills in second-language acquisition has been proposed, with multiple-choice questions assessing vocabulary and grammar being the most popular. In the area of vocabulary question, many studies have been done, e.g. generation of cloze questions asking for completion of a sentence, generation of questions asking for a synonym and antonym of a certain word, and the like. Vocabulary questions also have been generated to test student's knowledge of the correct usage of

English verbs (Sakaguchi et al., 2013), prepositions (Lee and Seneff, 2007) and adjectives (Lin et al., 2007) in sentences. In what follows, we describe related studies and for each study we dissect its purpose, how the method works and how the evaluation was done.

Coniam (1997) investigated the extent to which it is possible to produce multiple-choice English vocabulary cloze questions from a text. The system allowed questions to be constructed in three different ways: 1) selecting every nth-word in the text to be a question, 2) selecting words with a user-specified frequency range and 3) specifying a certain word class (e.g., noun, verb, or adjective). Word class and word frequency of each question key are matched with similar word class and word frequency options to construct the question options. The evaluation was done by administering the generated questions to 60 students and determined the "acceptability" of the options using item analysis. The result showed that the questions produced by the nth-word-deletion mode was the least successful compared to the other two which are language-oriented modes (specifying a word class or a frequency).

Brown et al. (2005) generated multiple-choice questions by taking their components from WordNet. Based on the attributes and lexical relations in WordNet, six types of vocabulary questions are defined: questions asking for definition, synonym, antonym, hypernym, hyponym and cloze questions. A word is selected from a given text, and information of a word is extracted for all six question types. For example, the definition question requires a definition of the word, which is retrieved from the WordNet glosses. The evaluation was done by administering the questions to 21 native English speaking adults. The result suggested that the generated questions gave a measure of vocabulary skills that correlates well with human-developed questions and standardised vocabulary tests.

Sumita et al. (2005) described a method for automatic English cloze question generation by combining a corpus (for generating the questions), a thesaurus (to find the potential distractor candidates) and web filtering (to verify the distractors). An evaluation was done by comparing the non-native proficiency scores from the generated questions and their real TOEIC®scores, and they showed a high correlation . An English native speaker also did the test and scored 93.5%, which is higher than the highest-score by the non-native speaker, who scored 90.6%.

Lin et al. (2007) also adopted WordNet to produce English adjective questions from a given text. The candidates of options (correct answer and distractors) are taken from WordNet and filtered by Web searching. Human expert manually examined the validity of the generated questions. The result showed that the proposed correct answer determi-

nation and question filtering contributed to the high precision.

There are more studies for generating cloze questions. Lee and Seneff (2007) generated English cloze question for prepositions and focused on generating the distractors. They generated distractors using two methods: 1) using the collocation information of the preposition 2) employing information from the most frequent mistakes made by non-native speaker in a non-native corpus. For the evaluation, a native speaker took the generated questions and about 96% of the generated distractors were usable. They also measured the difficulty of the distractors by administering the questions on four non-native students. The result showed that the distractor generated by the methods worked well in distracting the students than the baseline.

Smith et al. (2010) attempted to produced draft question items for gap-fill exercises (cloze question). The system takes a correct answer as the input and generates distractors and the questions by utilising a corpus and a thesaurus. The evaluation involving two English native speakers showed that about 53% of the generated questions were acceptable as it is or with minor revision.

Sakaguchi et al. (2013) proposed discriminative methods to generate *semantically confusing* distractors of cloze question for English language learners using a large-scale language learners corpus. They focused on creating questions for verbs. The proposed methods aim at satisfying both reliability and validity of generated distractors; distractors should be exclusive against answers to avoid multiple answers in one question, and distractors should discriminate learner‘s proficiency. User evaluation with native speakers showed that 98.3% distractors are reliable and the accuracy of non-native speakers of English on generated questions showed a positive high correlation with the student’s TOEIC®scores.

Unlike the above work, we focus on generating a vocabulary question asking for closest-in-meaning of an English word, which is administered in the TOEFL®reading section. As ETS[1] claims, TOEFL®is the most widely respected English-language test in the world, recognised by more than 9,000 colleges, universities and agencies in more than 130 countries, including Australia, Canada, the U.K. and the United States. Considering the popularity of TOEFL®, modelling questions following TOEFL®could be beneficial.

The closest-in-meaning question has similar implication with a ‘synonym’ question (as had been generated by Brown et al. (2005)), but here we need to generate a reading passage that contains the target word with a selected word sense as well. Compared to a simple ‘synonym’ question, our target question is more advanced since it provides a

[1]Educational Testing Service, an organisation which administers tests such as the TOEFL®, TOEIC®, GRE®and Praxis Series®tests

reading passage for each question. A test taker who can not capture a specific word sense of the target word in the reading passage would not be able to correctly answer the question.

2.2 Computerised adaptive test

Emerged in the 70s, computerised adaptive test (CAT) is a test in which items were chosen to present to examinees based on their previous responses. Initially, such concept was called *tailored testing* by Lord et al. (1968). With computer technology that facilitated implementation of this concept, the name was changed into computerised adaptive testing. Unlike the conventional paper-and-pencil test, or linear test, CAT prepares different tests for different test takers.

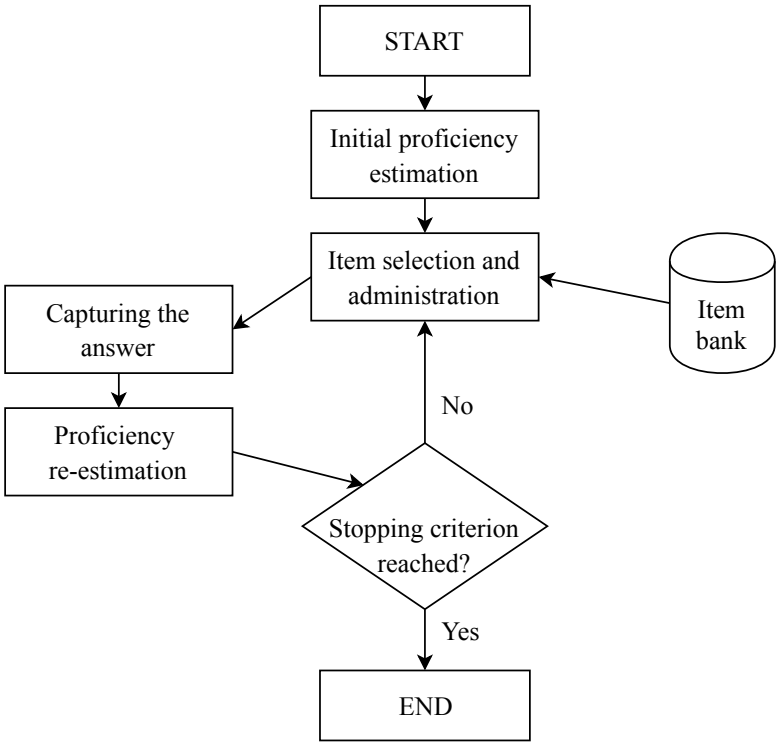

Figure 2.1: CAT procedures

The procedure of administering CAT is illustrated in Figure 2.1. The test starts with setting the initial proficiency of a test taker and the first item is selected according to the initial proficiency. The items are selected from the item bank, which is a collection of items. The proficiency of the test taker is then re-estimated with respect to their response

to the first item. This estimation is then used to determine the next item. The response to that item refines the proficiency estimation of the test taker and the cycle of the process continues until it reaches a certain stopping criterion. The following is detailed description of the four main steps of CAT.

1. Initial proficiency (θ_0) estimation. Ideally, the closer initial proficiency is to the true proficiency, the faster it converges to the test taker's true proficiency value. The initial proficiency may be set in various ways, including 1) a standard value for all test takers 2) a random value according to a probability distribution.

2. Item selection. An item is selected based on the current estimation of the test taker's proficiency. We listed several strategies in the following.

- Maximum information selection (Weiss, 1974) : it selects an item that maximises the information gain. This method guarantees faster decrease of standard error, but it can cause overexposure of items in the bank. In 1-parameter models, an item is most informative when its difficulty parameter is close to the test taker's proficiency (matched difficulty). This is the oldest and widely used item selection method.
- Stratified selection (Chang and Ying, 1999): the item selection begins by stratifying the item bank according to item discrimination. More informative (more discriminating) items are placed at the bottom stratum and less informative item are placed in the top. Selection is made from more discriminating stratum toward the middle of the test and changed into the selection from the most discriminating stratum by the end of the test. Within each stratum, items are selected by matched difficulty.
- Cluster selection (De Rizzo Meneghetti and Thomaz Aquino Junior, 2017): the item selection begins by clustering the items according to their parameter values and selects the items from the cluster that contains either the most informative (discriminating) item or item with the highest average information gain.

3. Proficiency (θ) re-estimation. The test taker's proficiency is re-estimated after the response to the administered item. This proficiency reflects the test taker's proficiency up to that item in a test. Common methods for the proficiency re-estimation include 1) Maximum-likelihood estimation 2) Bayesian estimation which uses prior knowledge of the distributions of the test taker's estimated proficiencies.

4. Stopping criterion. In CAT, a test ends when it reaches a predefined threshold of the standard error or when a fixed number of items is administered.

2.3 Integration of AQG and CAT

Attempts in the integration of CAT and AQG are scarce. One early attempt is by Bejar et al. (2002) which assessed the feasibility of an approach to adaptive testing based on item models. They selected several item models and used them to produce isomorphic items. They further calibrated the item models and applied the model calibration to all instances of the model. Another study is by Hoshino (2009) who developed item difficulty predictor using machine learning and applied the predictor to assign the difficulty to newly-generated items. In those related studies, the items still need to be calibrated by administering the items to test takers; either to obtain the model calibration or to train the difficulty predictor. Consequently, the cost of the calibration process could not be avoided. The present study focuses on integrating CAT with AQG without any item calibration process. In other words, the item parameters, e.g. item difficulty, should be estimated in advance during the generation process so that there is no need to administer the items in a pretesting.

The studies of item difficulty in language tests are directed more toward predicting the item difficulty than controlling it. These are fundamentally different tasks since the former concerns how difficult an item is, while the latter concerns with how to create items of various levels of difficulty. Earlier work on predicting item difficulty has been done on reading and listening comprehension questions using multiple regression combined with regression tree analysis (Rupp et al., 2001) and artificial neural networks (Perkins et al., 1995; Boldt and Freedle, 1996). More recently, Loukina et al. (2016) conducted a study to investigate the extent to which textual properties of a text affect the difficulty of listening questions in the English test. Trace et al. (2015) used item and passage characteristics to determine the item difficulty of cloze questions across the test taker's nationality and proficiency level. Other studies focused on vocabulary questions, as conducted by Hoshino and Nakagawa (2010) and Beinborn et al. (2014). Beinborn et al. (2014) worked on predicting the gap difficulty of the C-test[2] using a combination of factors such as phonetic difficulty and text complexity, whereas Hoshino and Nakagawa (2010) investigated factors affecting item difficulty of multiple-choice vocabulary questions.

Unlike the previous research, the aim of the present study is to control item difficulty in the automatic question generation task. Our long-term goal is to integrate AQG into CAT, where a question with a specific difficulty is created *on-the-fly* before they are presented to test takers. We conducted simulation-based experiments of the AQG and CAT integration by controlling the question difficulty intrinsically by the proposed method, and

[2] A test where some fraction of words have been removed from a text (gap).

showed that the integration is feasible.

2.4 Neural Test Theory

In this paper, we conducted a test taker-based evaluation by administering the machine-generated questions to the English learners. We further evaluated the quality of the questions by applying Neural Test Theory (NTT) (Shojima, 2007), which is a test theory for analysing test data. The NTT model evaluates academic achievements of the test takers in an ordinal scale. The motivation of this theory is that a test cannot distinguish test takers who have nearly equal abilities; the most that a test can do is to grade them into several ranks. The English proficiency of learners is commonly divided into groups of proficiency, e.g. 'high, intermediate, low', or 'good, fair, limited' as used by ETS in their explanation of TOEFL scores[3]. For that reason, we believe that using NTT to analyse the test data is suitable for our purpose in the present study since NTT divides the test takers into ordinal ranks.

NTT uses a self-organising map mechanism (SOM, Kohonen (1995)) to estimate the test taker's ranks and place them on the ordinal scale. Item category reference profile (ICRP) is a feature of NTT representing the probability that the test takers in a certain rank select a certain category (i.e. question options) in their responses to a certain question item. The ICRP is obtained by a statistical learning process as explained in Shojima (2007). We summarised it in the following.

Let us assume a latent rank scale with Q number of ranks, each rank is represented by node $R_q(q = 1, ..., Q)$. The ability of a test taker at node R_{q+1} is higher than that at $R_q(q = 1, ..., Q)$. Assuming the number of items is n, node R_q has an n-dimensional vector v_q called the reference vector. As an example, a latent rank scale for $(Q, n) = (7, 12)$ is shown in Figure 2.2. The black circles are the nodes representing the ranks and grey circles are the reference vector.

Let us further assume the test taker size is N and the response data of the test takers is $U = \{u_i\}(i = 1, \ldots, N)$, and that $v_q^{(t)}$ is the reference vector R_q at the t-th period, with the recommended initial value for v_q^1 is $q1/Q$. The learning procedure is as follows.

[3] https://www.ets.org/toefl/ibt/scores/understand/

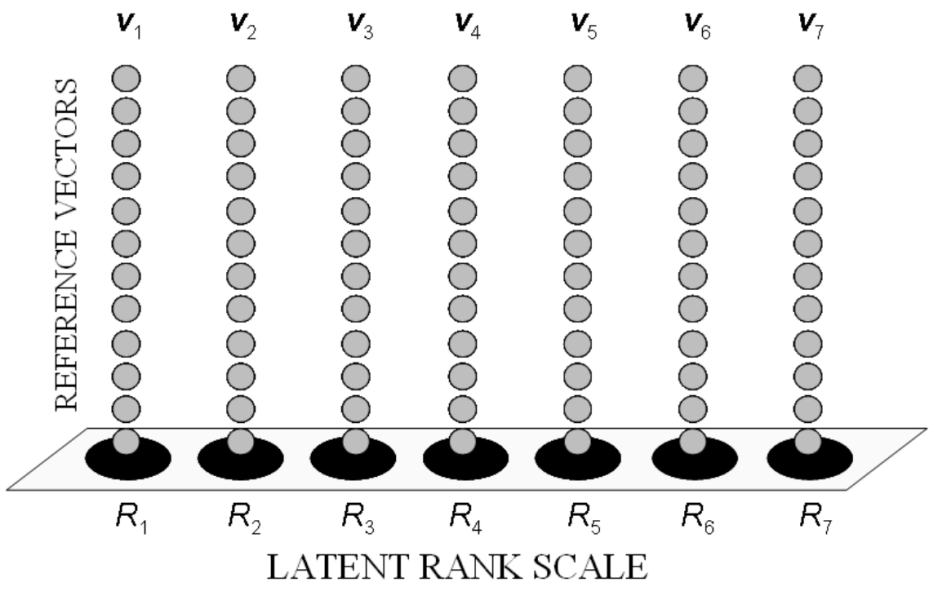

Figure 2.2: Latent rank scale and reference vectors of NTT

For $(t = 1; t \leq T; t = t + 1)$

- $U^{(t)} \Leftarrow$ Randomly sort the row vectors of U.
  - For $(h = 1; h \leq N; h = h + 1)$
    - Input $u_h^{(t)}$ the h-th row vector of $U^{(t)}$ and select the winner with the closest reference vectors in terms of discrepancy function d.
    - Obtain $V^{(t,h)}$ by updating the reference vectors of the winner and the neighbouring nodes.
- $V^{(t+1)} \Leftarrow V^{(t,N)}$

The fundamental part in the learning of NTT is reflected in the process of updating the reference vectors. First, on selecting the winner node, they recommend the Euclidian distance as the discrepancy function d as it is often used in SOM applications. When $u_h^{(t)}$ is the input, the winner node R_w by the square distance is determined as follows:

$$R_w : w = \arg\min_{q \in Q} \|v_q^{(t)} - u_h^{(t)}\|^2. \tag{2.1}$$

In updating the reference vectors, the reference vectors of the nodes that are closer to the winner should be designed to become numerically closer to the input data. For updating $v_{qh}^{(t)}$, the reference vector of Node R_q when $u_h^{(t)}$ is input at the t-th period, one of the valid candidates is as follows:

$$\begin{aligned}&\text{For } (q = 1; q \leq Q; q = q + 1)\\ &- v_{qh}^{(t)} = v_{qh-1}^{(t)} + h_{qw}(t)(u_h^{(t)} - v_{qh-1}^{(t)})\end{aligned}$$

where

$$\begin{aligned}h_{qw}(t|\alpha_t, \sigma_t^2) &= \alpha_t \text{exp}\{-\frac{(R_q - R_w)^2}{2\sigma_t^2}\},\\ \alpha_t &= \frac{T - t + 1}{T}\alpha_1,\\ \sigma_t &= \frac{(T - t)\sigma_1 + (t - 1)\sigma_0}{T - 1}\end{aligned}$$

The ICRPs are the itemwise reference vectors of the finally obtained $V^T, v_j^{(T)}(j = 1, \ldots, n)$. It can then be represented in a graph, as shown in the Figure 2.3. It shows an ICRP graph for a four-options item with three latent ranks of test takers. The x axis denotes the three latent ranks and the y axis denotes the probability of each rank to select each option. The lower the rank, the lower the proficiency as estimated by the NTT model.

ICRP shows how test takers in each rank behave against each option of the question, so it can be used to clarify the validity of the question options. For instance, it can be used to clarify if a distractor *correctly deceives* the low-proficiency test takers compared to the high-proficiency test takers. As an example, suppose the option 2 (yellow line) in the Figure 2.3 is the correct answer. From its ICRP graph we can confirm that the probability of the test takers to correctly select the correct answer is increasing from rank 1 (low-proficiency group) to rank 3 (high-proficiency group). This monotonically increasing line is the ideal one for the ICRP of the correct answer, while the opposite (a monotonically decreasing line) should be expected from the other three options, i.e. the distractors.

In this thesis, we used the Nominal Neural Test (NNT) model (Shojima et al., 2008), which is a variant of NTT for nominal-polytomous data that is suitable for the current vocabulary multiple-choice questions. The analysis of NTT is performed using Exametrika[4].

[4]Free software for NTT analysis, available at http://www.rd.dnc.ac.jp/ shojima/exmk/

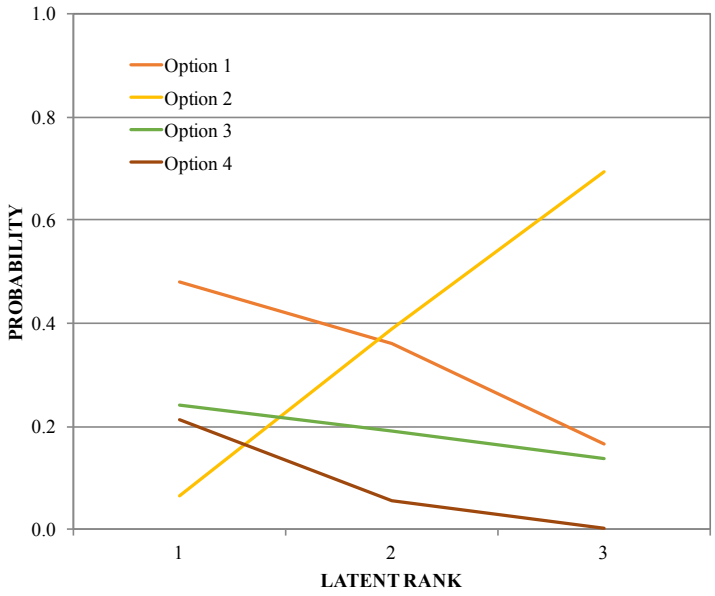

Figure 2.3: Example of ICRP graph for a four-options item with three ranks

Chapter 3

Automatic question generation

3.1 Method overview

Given a target word with its part-of-speech (noun, verb, adjective or adverb) and a word sense as the input, the task of generating a vocabulary question can be broken down into three: 1) reading passage generation, 2) correct answer generation and 3) distractor generation. The method is illustrated in Figure 3.1.

The reading passage generation retrieves a text that includes the target word used in the given word sense from the specified Web site. As we make a question asking for a word that is closest-in-meaning to the target word, the word sense of the target word plays a crucial role in generating the options, i.e. a correct answer and distractors, in the following steps.

Having obtained the reading passage, we generate the correct answer in a straightfor-

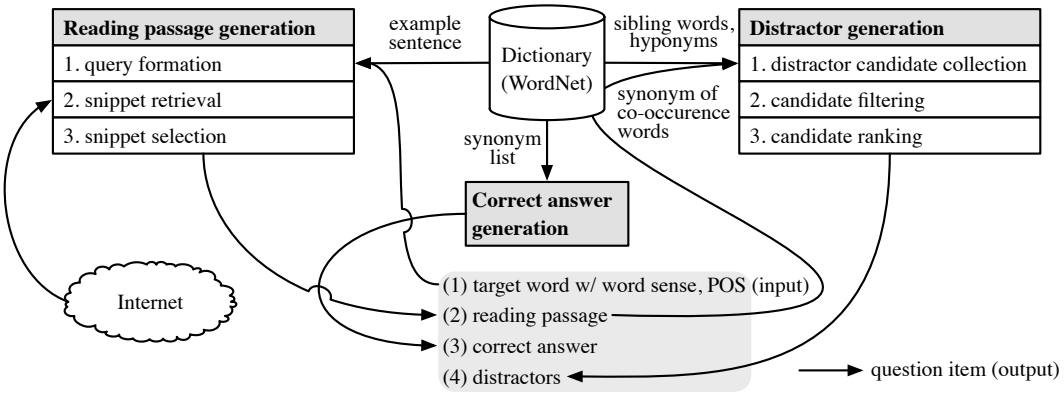

Figure 3.1: Automatic question generation method overview

ward way by referring to WordNet. We have two kinds of a correct answer: a single-word correct answer and a multiple-word correct answer. A single-word correct answer is generated by choosing a synonym of the target word with the given word sense. A multiple-word correct answer is generated by simplifying the gloss of the target word in WordNet.

Generating relevant distractors is crucial due to its great impact on the quality of the generated question. Too-easy or too-difficult distractors fail in distinguishing the test takers proficiency since the questions can either be answered by all test takers or nobody at all. The fundamental requirement is that distractors should be distracting while still keeping its distinctive meaning from the target word and the correct answer. To satisfy the requirement, we collect distractor candidates from the two different sources: the reading passage and the WordNet taxonomy. We rank the candidates and take the three highest-ranked candidates as the final distractors.

We describe each task in detail followed by the evaluation experiments in the Sections 3.2-3.4.

3.2 Reading passage generation

In English proficiency tests such as TOEFL®, the reading passage is taken from university-level academic texts with various subjects such as biology, sociology, and history. In this study we generate similar passages, but not limited to academic texts; we use the Internet as the source for retrieving the reading passages. We can control the text domain by choosing Web sites for retrieving the reading passage. Here the users, e.g. English teachers, can choose the sites depending on their purpose. For example, if the users prefer news articles on technology, they can choose sites such as 'www.nytimes.com' with specifying the 'Technology'. Retrieving a reading passage from the Internet, especially from news portals, gives a lot of benefits because such texts are expected to be new and up-to-date, in terms of both content and writing style. They also come from broad genres and topics, make them suitable for English language learning.

To obtain a reading passage in which the target word is used in the given word sense, we could take a straightforward approach to retrieve a text including the target word and check whether it is used in the specified word sense or not. Generally, a word in a dictionary has several meanings, whereas the word in a given text is used for representing one of those meanings. The task of identifying the correct word sense in context has been studied in natural language processing field under the name of 'word sense disambiguation (WSD)'.

3.2.1 Word sense disambiguation

Word sense disambiguation (WSD) is the task of identifying the meaning of a word in context in a computational manner (Navigli, 2009). Vocabulary questions in this research ask the test takers to select the option that is closest in meaning to the target word in context; to generate relevant options we need to identify the meaning of the target word in a reading passage in the first place. Therefore, WSD is crucial for generating vocabulary questions, especially to generate a correct answer and distractors. The state-of-the-art WSD methods as explained by McCarthy (2009) reach around .37 in accuracy with a knowledge-based approach, .88 with supervised and .82 with unsupervised machine learning approaches. Further explanation on WSD can be found in survey papers by Navigli (2009) and McCarthy (2009). In this research we use Lesk algorithm (Lesk, 1986) which chooses the sense that shares the highest number of words in its gloss (or example sentence) in a dictionary and the current context. For instance, given two word senses with their glosses for 'key' in a dictionary (WordNet):

1. *Metal device shaped in such a way that when it is inserted into the appropriate lock the lock's mechanism can be rotated.*

2. *Something crucial for explaining; "The key to development is economic integration."*

the word sense of 'key' in the context 'I inserted the **key** and locked the door.' should be identified as word sense 1, because its gloss has a three word overlap ('insert' and two 'lock's) with the context, while word sense 2 has no word overlap at all.

Since even with the state-of-the-art WSD method high accuracy is not always available, past attempts avoided the use of WSD for generating vocabulary questions by utilising the most frequent word sense with its example sentences as a context in WordNet (Brown et al., 2005). This is, however, obviously not enough to create decent questions because most frequently used senses in WordNet are based on a small corpus[1] and the length of reading passages is limited, at most a sentence.

To remedy insufficient performance of WSD, we propose combining WSD and our *context search* (CS) method described in the next section (3.2.2). Given a target word and its context, the task is to identify the correct word sense of the target word in that context. CS works in reverse; given a target word and one of its word senses, it searches for passages in which the target word is used with the given word sense. To combine both, WSD is applied to the target word in the retrieved passage by CS to confirm that

[1] mentioned in https://wordnet.princeton.edu/wordnet/documentation/

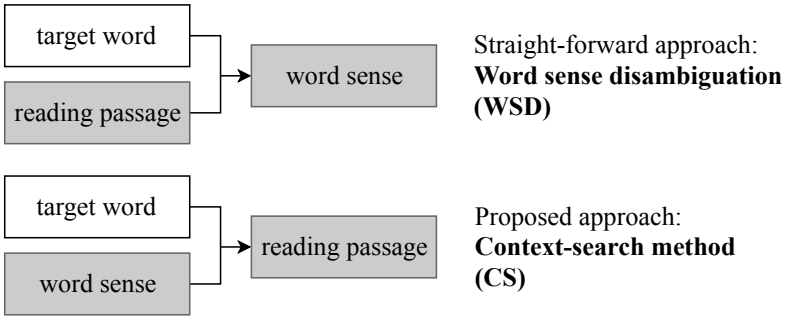

Figure 3.2: The straight-forward vs proposed method to obtain reading passage and word sense pairs for the AQG

the predicted word sense is the same as the given sense from CS. Figure 3.2 illustrates the difference between the two methods. In our experiment, we used target words from TOEFL®iBT sample questions. In real applications, however, the users can provide the target words with its word sense according to their purpose.

3.2.2 Context-search method (proposed)

Given a target word and one of its word senses, context search (CS) is a threefold process:

(1) query formation from the example sentence,

(2) retrieval of snippets with a search engine, and

(3) snippets scoring to choose one of them as an appropriate reading passage.

A query for the search engine is created from the example sentence of the specified word sense by taking the target word and its adjacent two words on both sides after removing stop words such as 'the', 'on', 'are' and so on. When the target word is located at the beginning or the end of the sentence, the two following or preceding words of the target word are taken for the query. The created query is submitted to the search engine to retrieve snippets containing the target word possibly with the given sense. The last step selects a snippet which is the most probable snippet containing the target word with the given sense. Plausibility that the word sense of the target word in the snippets is the same as the specified word sense is calculated based on the following three scores: 1) S_o: word overlap between the example sentence and the snippet, 2) S_a: the number of adjacent query words to the target words in the snippet after removing the stop words, 3) S_q: the number of query words appearing in the snippet.

The following is a detailed example of the score calculation. Assume our target word is 'bright' with *intelligent* sense, and given the example sentence 'She was a **bright** young graduate from my university', we have query words 'bright', 'young' and 'graduate'. Note that since 'she', 'was' and 'a' are stop words, the target word 'bright' is at the beginning of the sentence after stop word removal, thus 'young' and 'graduate' are used for the query.

Suppose that we retrieved the following two snippets where the query words are underlined and the target word is indicated in bold face.

S1. Mary is a **bright** young PhD graduate from Yale University. She was a sophomore in college when she found her true passion in research.

S2. Since she was a child, Mary has been a friendly girl. Mary always gives a **bright** smile to her friends in the university campus.

The first score S_o, the overlap word score, counts the word overlap between the example sentence and the snippets. The scores S_o for these snippets are $S_o(\mathrm{S1}) = 4$ since 'bright', "young', "graduate' and 'university' overlap, while $S_o(\mathrm{S2}) = 2$ for 'bright' and 'university'.

The second score S_a counts the number of adjacent query words to the target words in the snippet after removing the stop words. Thus, $S_a(\mathrm{S1}) = 1$ for 'young', and $S_a(\mathrm{S2}) = 0$.

The third score S_q counts the number of query words that appear in the snippet. Thus, $S_q(\mathrm{S1}) = 3$ for 'bright', 'young' and 'graduate', while $S_q(\mathrm{S2}) = 1$ for only 'bright'.

The three scores are combined to provide the final score for each snippet as given by

$$S = S_o + S_a + S_q. \tag{3.1}$$

The method then extracts three sentences: a sentence containing the target word, and the two sentences before and after it, as the reading passage for a question. However, if the target word is located in the first or last sentence of the retrieved text, the reading passage would be composed of two or three sentences: a sentence containing the target word, and one or two sentences before or after it.

3.2.3 Preliminary evaluation of reading passage generation

We conducted a preliminary experiment on two target word sets: 98 target words from TOEFL®iBT sample questions[2] and preparation books for TOEFL®iBT (ETS, 2007;

[2]Available at http://ets.org/toefl

Sharpe, 2006; Phillips, 2006), and randomly selected another 98 target words from Senseval-2 and Senseval-3 data which were prepared for Senseval WSD workshops[3]. These two target word sets share no common words. The Bing Search API[4] was used as the search engine, and we limited the target site to www.nytimes.com. In this experiment, we compare the results of the following three settings.

- **WSD:** We identify the word sense of the target word in a given context by using the Lesk algorithm.
- **CS:** For each target word in the test sets, context search is applied to find the context sentences in which the target word is used with a given sense.
- **CS+WSD:** WSD is applied after CS to confirm that CS has retrieved snippets containing the target word with a given word sense. The snippets with a word sense mismatch are discarded.

Evaluation was done manually to see if the identified word sense is correct for the WSD setting, and to see if the retrieved snippet with the highest score uses the target word with the given sense for the CS and CS+WSD settings. Thus, we evaluated to what extent we could correctly obtain pairs of word senses and their reading passages. Table 3.1 shows the accuracy of each setting. Note that the denominator for CS and CS+WSD methods are not 98 because there were cases where they did not retrieve the reading passage at all, e.g. because there is no article containing the target word in the specified target site (www.nytimes.com). The accuracy of CS reached .89 on the TOEFL®data. In addition, by combining with WSD the accuracy improved to .95. This means that the method successfully discarded the mismatch between the given and predicted word senses (discarded the article using the target word with incorrect word sense), resulted in the higher accuracy since the denominator also became smaller. We also evaluated in the Senseval data and it shows that the accuracy of CS is higher than the accuracy of WSD. The accuracy is also improved when we combined the two methods. Although it is still a preliminary evaluation, the proposed CS method combined with WSD shows promising results for continuing to the next step in generating vocabulary questions.

3.3 Correct answer generation

The correct answer in vocabulary questions is the option that has the closest meaning to the target word used in the reading passage. It does not ask for collocation; therefore

[3]http://senseval.org

[4]https://datamarket.azure.com/dataset/bing/search

Table 3.1: Accuracy of WSD and the proposed CS method in correctly obtain pairs of word senses and reading passages

setting\ data	TOEFL®	Senseval
WSD	.60 (58/98)	.30 (29/98)
CS	.89 (85/96)	.74 (73/98)
CS + WSD	.95 (80/84)	.78 (47/60)

the correct answer is not necessarily replaceable with the target word used in the reading passage.

In this work, we generate two kinds of correct answers: single-word and multiple-word correct answers. The following subsections (3.3.1 and 3.3.2) describe the generation of each kind of correct answer.

3.3.1 Single-word correct answer

A single-word correct answer is composed of one single word. Based on our observation of TOEFL®iBT official sample questions[5], the correct answer in a vocabulary question shows the following characteristics.

- It has the same part-of-speech as the target word.
- It shares a similar meaning to the target word.
- It does not share any substrings with the target word. Words with a similar meaning often share substrings in their spelling, for example, the word 'synchronisation' and 'synchronism'. Both words are nouns and have similar meaning, but these words should not be a target word and correct answer of each other because the test taker can easily estimate a correct option based on their similarity in spelling.

To realise the first and second characteristics, we take the candidates of a single-word correct answer from synonyms of the target word in the dictionary, WordNet[6] in our case. After filtering with respect to those three characteristics, we choose the first synonym as the single-word correct answer. For instance, given the target word 'bright' with word sense *bright.s.02*[7], all of its lemmas, 'brilliant' and 'vivid' are retrieved from WordNet, and we choose 'brilliant', its first synonym, as the correct answer. In case of a word with no synonym in the dictionary, we make a multiple-word correct answer by using its gloss.

[5] 38 vocabulary questions available at www.ets.org/toefl

[6] We use WordNet 3.0 available at https://wordnet.princeton.edu/wordnet/

[7] Roughly, the second word sense of adjective 'bright'.

3.3.2 Multiple-word correct answer

A multiple-word correct answer is the correct answer composed of more than one word as in option (b) in Figure 1.1. Note that past research on vocabulary question generation did not deal with multiple-word options which actually appear in the real English proficiency tests. Multiple-word options, both for correct answers and distractors, are generated from the gloss in a dictionary.

The multiple-word options in TOEFL®iBT sample questions are usually composed by no more than four words. However, sometimes the gloss can be longer than four. In such case, we simplify the long gloss. In the WordNet lexical dictionary, a long gloss tends to include disjunctive structures introduced by disjunctive markers like 'or'. In simplifying the gloss, we divide the gloss based on its disjunctive markers. We define the disjunctive markers depending on the dictionary. In the case of WordNet, we use 'or' and ';' for disjunctive markers.

In generating multiple-word correct answers, we directly use the gloss of the target word if it consists of no more than four words. If it is longer than four words, it is divided by disjunctive markers and the element which has the least number of words is adopted (elements which consist of only one word are excluded). When the numbers of words in the elements are the same, the left most element is selected. The following are some examples of gloss simplification. The target words and their glosses are shown with the result of simplification underlined, which is used for a multiple-word correct answer.

'accepting': consider or hold as true

'leaked': tell anonymously

'lived': inhabit or live in; be an inhabitant of

3.4 Distractor generation

Distractors are the incorrect options in a multiple-choice question. Many multiple-choice questions have four options; thus, we generate three distractors for a question.

There are two fundamental requirements for distractors. Firstly, they must be hard to distinguish from the correct answer and the target word, and secondly, they cannot be considered as a correct answer. These two requirements seem to be contradicting each other since the first requires distractors should be somehow similar to the target word, while the second requires the distractors to be different from the target word. Making a reasonable trade-off between these two requirements is important.

In this study, distractor generation is a three-fold process:

(1) collecting distractor candidates,

(2) filtering the candidates so that they fulfil necessary requirements, and

(3) ranking the candidates based on a scoring function.

3.4.1 Distractor candidate collection

Distractor candidates are collected from two sources: 1) the passage retrieved by CS and WSD for each target word and 2) the lexical hierarchy in the WordNet taxonomy. We use these two sources because each of them reflects a different aspect of 'similarity' relations to the target word. The first is the association relation that the words in a passage are somehow related to each other concerning the topic that the passage describes. Therefore, those co-occurring words with the target word are reasonable to be distractors. We only consider the co-occurring words with the same part-of-speech as the target word. However co-occurring words themselves are not appropriate for the distractors, since they actually appear in the passage. Therefore we collect their synonyms as distractor candidates for the target word.

The second source is the hierarchical relation in the lexical taxonomy that is defined in a dictionary. We focus on words being sibling and hyponym to the target word in the WordNet taxonomy. Words in the sibling and hyponym relations share the same ancestor (parent) with the target word; thus they are similar to the target word.

There are cases in which the number of distractor candidates from those two sources is not enough. When that happens, we take additional candidates from WordNet with the same part-of-speech and with close generality to the target word. All of the word senses with the same part-of-speech in WordNet are ordered in generality, from more general to less general word senses. For instance, the first word sense for noun is *entity.n.01* followed by *physical entity.n.01*, *abstraction.n.06*, *thing.n.12*, *object.n.01*, and so on. We select distractor candidates from words located around the order of the target word in this list.

3.4.2 Distractor candidate filtering

According to Heaton (1989), there are several requirements for options in multiple-choice questions. The following are Heaton's requirements followed by the descriptions of our implementation of the requirements. All examples mentioned below are taken from Heaton (1989).

(1) *Each option should belong to the same word class as the target word.*
We choose distractors with the same part-of-speech as the target word.

(2) *The options should be related to the same general topic or area.*
We collect distractor candidates from the synonyms of co-occurring words in a passage and sibling and hyponym words in the WordNet taxonomy (3.4.1) and further calculating the similarity between candidates and the target word (3.4.3).

(3) *Distractors and the correct answer should be at approximately the same level of difficulty.*
We use word difficulty level provided by JACET 8000 (Ishikawa et al., 2003), which is based on the British National Corpus[8] but supplemented with six million tokens of text targeted at the needs of Japanese students. Further explanation is described in 3.4.3.

(4) *All the options should be approximately the same length.*
Since we consider both single and multiple-word options, we do not take this requirement.

(5) *Avoid using pairs of synonyms as distractors. Such distractors can be ruled out easily by test takers.*
If there is a pair of synonyms in the candidates, we will remove one of them from the distractor candidates. For example, given a target word 'courteous' in the sentence '*The old woman was always courteous when anyone spoke to her.*', the options '(A) polite, (B) glad, (C) kind and (D) pleased' are not appropriate, since 'glad' and 'pleased' are synonyms. The test takers will be able to guess that the correct answer should be one of other two, 'polite' or 'kind'.

(6) *Avoid using antonyms of the correct answer as distractors. The test takers can also easily eliminate such distractors.*
We generate antonym list of the correct answer from WordNet and exclude them from distractor candidates. For example the options '(A) go up, (B) talk, (C) come down and (D) fetch' for the target word 'ascend' are not appropriate, since the antonym pair 'go up' and 'come down' immediately stand out, providing a clue for guessing the correct answer.

[8] http://www.natcorp.ox.ac.uk/

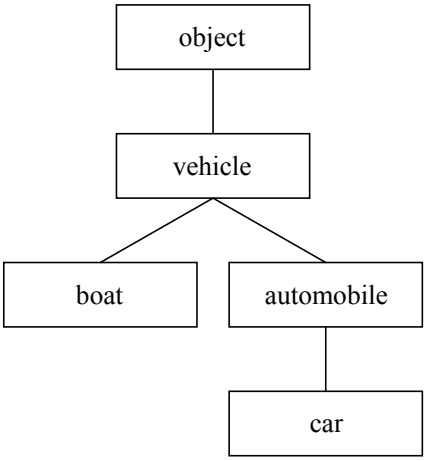

Figure 3.3: Illustration of the WordNet taxonomy

3.4.3 Candidate scoring and ranking

At this point, we already have distractor candidates filtered by the requirements mentioned in the previous section. Since we only need three distractors for a question, this step chooses the three most appropriate distractors from the candidates. As mentioned in Section **??**, a good distractor should be related to the correct answer and target word so that it will be hard to distinguish it from the correct answer.

We rank the distractor candidates regarding their 'closeness' to the correct answer by using a combination of the Path similarity and WU-Palmer similarity score calculated by using WordNet. The Path similarity (Pedersen et al., 2004) score is calculated from the shortest path length (number of nodes/ relation links) in the taxonomy, while Wu-Palmer similarity (Wu and Palmer, 1994) is calculated based on the depth of two nodes in the taxonomy and that of their LCS (Least Common Subsumer, the most specific ancestor node). Wu-Palmer similarity is defined in Equation 3.2.

$$sim_{ab} = \frac{2 * depth(LCS(a,b))}{depth(a) + depth(b)} \tag{3.2}$$

Taking the Figure 3.3 as an example, the shortest path for the 'boat' and 'car' is 'car-automobile-vehicle-boat' (path length = 4; thus similarity score is 1/4 = .25). As for the Wu-Palmer Similarity, the LCS for 'boat and 'car' is 'vehicle' (the depth is calculated from the root of the tree, hence the depth of the LCS is 1, depth of 'car' is 3, and depth of 'boat' is 2). Thus the Wu-Palmer similarity for 'boat' and 'car' is .4.

The resultant candidates are sorted in ascending order of the average of these two similarity scores. The top three ranked candidates with the closest difficulty level to the

correct answer are selected as the final distractors. Here we again make use of the JACET 8000-based word difficulty level. Given a difficulty level x of the correct answer (or target word in case the correct answer is multiple-word), we select the candidates following this list order: level $[x, x-1, x+1, x-2, x+2, .., 9]$[9]. For example, for a generated correct answer with level $x = 5$, we give preference to the distractor candidates that has difficulty level as close as possible to the correct answer's. In this particular example, we give preference following the order: level $[5, 4, 6, 3, 7, .., 9]$.

3.4.4 Single and multiple-word distractors

As the result of the distractor generation, we obtain three ranked single-word distractor candidates. As we generate both single and multiple-word correct answers, we also generate both types of distractors. Multiple-word distractors are created from the gloss of the single-word distractor candidates by the same gloss simplification method explained in **??**. However, when the length of a multiple-word distracter exceeds four words, we use its single-word version.

3.5 Evaluation of the AQG

The evaluation of automatically generated questions in language learning needs to consider at least the following two aspects. First, the questions can measure English learner's language proficiency precisely. This is important for both teachers and learners. Second, they have comparable quality to human-made questions. This aspect is particularly important from a teacher's perspective. We describe in detail the evaluation process of the machine-generated questions and provide thorough analyses of those two aspects.

3.5.1 Evaluation 3.1: measuring proficiency of English learners

The main purpose of this evaluation is to investigate if the machine-generated questions can measure English learners' proficiency precisely. We ask English learners[10] to complete sets of machine-generated and human-made questions and compare their scores on those two sets to see if there is a correlation between them. In addition, we compare their scores on the machine-generated question set with their commercial English test scores in-

[9] Note that we put level 9 ('level 0' in JACET 8000 means the words above level 8, non-English words, misspeling, etc) as the least priority.

[10] In this thesis, the term 'English learner' is used interchangeably with 'test taker'.

cluding TOEIC®, TOEFL®, and CASEC[11]. If we can observe strong correlation between these two scores, we could claim that machine-generated questions are well produced, at least they are comparable with human-made questions in measuring English proficiency.

By analysing the test taker responses, we also estimate the effectiveness of each question item using a statistical method called *item analysis* (Brown, 2012). There are two metrics used in the item analysis. One is the *difficulty index* which is the proportion of test takers who answered the question item correctly. The other is the *discrimination index* that indicates how well each question item can discriminate the test takers according to their proficiency. Effective question items would have a moderate value of the difficulty index and a high value of the discrimination index, i.e. the questions are not too easy but also not too difficult and can distinguish test takers' proficiency.

Experimental design

We used two types of question sets in this experiment: machine-generated questions (MQs) created by the automatic question generation and human-made questions (HQs) taken from the official sample question[12] of TOEFL iBT® and preparation books (ETS, 2007; Sharpe, 2006; Phillips, 2006; Gear and Gear, 2006). Fifty target words were compiled from the same sources as the HQs. We selected the target words considering the balance of their part-of-speech and word difficulty level. The source for reading passages of the MQs were NY Times[13], CNN[14], and Science Daily[15] websites.

Two question item sets of HQs and MQs were prepared; each consisted of 50 questions. While the target words of these two sets are the same, other components of the question item (a reading passage, a correct answer, and distractors) are different, as ones are created by a machine while the others are created by humans. We further mixed the HQ and MQ sets to create four question sets (QS_A1, QS_B1, QS_A2, and QS_B2) as shown in Table 3.2. For instance, QS_A1 includes human-made questions (HQs) for target word (TW) 01-13 and machine-generated question (MQs) for target word 14-25. The order of the target words in the question sets was randomised and was kept the same across the sets A and B.

We administered the created question sets to 79 Japanese university undergraduate students (46 first year, 20 third year and 13 fourth year students). The test takers were divided into two classes randomly, C_A (40 students) and C_B (39 students) with keeping

[11] http://casec.evidus.com/
[12] www.ets.org
[13] www.nytimes.com
[14] www.cnn.com
[15] www.sciencedaily.com

Table 3.2: Configuration of question sets (Evaluation 3.1)

question set	Contents		Test taker
	HQs	MQs	
QS_A1	TW#01–13	TW#14–25	C_A
QS_B1	TW#14–25	TW#01–13	C_B
QS_A2	TW#26–37	TW#38–50	C_A
QS_B2	TW#38–50	TW#26–37	C_B

close the distribution of student year across classes. The proportion of male and female students was roughly about 2:1. We assigned the QS_A1 and QS_A2 to class C_A, and QS_B1 and QS_B2 to class C_B, so that the test takers of different classes worked on different question items (HQs and MQs) for the same 50 target words in total. The time slot for one question set was about 20 minutes, with a one-week interval between conducting evaluation for QS_A1/QS_B1 and QS_A2/QS_B2.

Results and discussion

Comparison of the MQ score with the scores from other tests We compared test taker scores on MQs with their scores on HQs in the present experiment, and with their commercial English test scores: TOEFL®, TOEIC® and CASEC (total score and vocabulary section score). In the calculation of test scores, we merged two question sets QS_A1 and QS_A2 into set QS_A, and QS_B1 and QS_B2 into set QS_B. A test score of each test taker for MQs was calculated by dividing the number of correct responses by the total number of MQs in the question set, i.e. 50. Note that each test taker took either the question set Q_A or Q_B. The score for HQs was calculated in the same manner. In what follows, we provide the Pearson correlation coefficients[16] between the test scores of MQs and that of the others[17].

We first calculated the correlation between the MQ test scores with the HQ test scores on both sets. These resulted in correlation coefficients .63 (t = 5.039, df= 38, $p < .05$) for the set A and .71 (t = 6.08, df = 37, $p < .05$) for set B. As for the comparison with the commercial test scores, we used less data in calculating the correlation since we do not have the test scores for some test takers. Table 3.3 presents the result where n denotes the number of test takers.

[16]Pearson correlation is used since our data follows normal distribution.

[17]Calculation is done using the `cor()` function of R software (www.r-project.org).

Table 3.3: Pearson correlation coefficients of the test taker scores with their English test scores

commercial tests	MQs	HQs	n
TOEFL®	.71	.60	21
TOEIC®	.68	.60	21
CASEC (total)	.57	.59	73
CASEC (vocabulary)	.55	.68	73

All p is less than .05.

As we can see in Table 3.3, the MQ test scores maintain strong positive correlation with the commercial tests and their coefficients are comparable with that of HQs. The positive correlations indicate that the machine-generated questions are promising for measuring English proficiency of the test takers, achieving a comparable level to the human-made questions.

We also calculated the reliability of the MQs and HQs by calculating the internal consistency estimates using Cronbach alpha, and this yielded averaged alpha .69 for MQs and .74 for the HQs. Since the reliability of the item is not perfect ($\neq 1$), the correlation of the scores will be attenuated, i.e. smaller than the actual correlation of the true scores. We further calculated the corrected correlation coefficients between test taker scores on the MQs and HQs and these resulted in correlation coefficients .88 for the set A and 1.0 for set B. We also calculated the corrected correlation coefficients with the English test scores, and the result is presented in the Table 3.4.

Table 3.4: Pearson correlation coefficients of the test taker scores with their English test scores (corrected)

commercial tests	MQs	HQs	n
TOEFL®	.77	.70	21
TOEIC®	.74	.70	21
CASEC (total)	.66	.73	73
CASEC (vocabulary)	.66	.79	73

All p is less than .05.

Item analysis Item analysis is the process of collecting, summarising and using information from test taker responses to assess the effectiveness of question items. The diffi-

culty index, discrimination index and item-total correlation are some metrics which help to evaluate the standard of multiple-choice questions used in a test. The item analysis was performed on the 50 questions of both HQs and MQs, and the result is explained below.

Difficulty index The difficulty index is the proportion of test takers that correctly answered a question item. It ranges from 0 to 1; a lower value means a more difficult item. The difficulty index of the MQs ranged from .18 to .90 (mean = .51, SD = .20), while that of the HQs did from 0 to .92 (mean = .53, SD = .23). Figure 3.4 shows the distribution of the difficulty index of the MQs and HQs. The pale colour bars denote the HQ frequency and the dark colour bars denote the MQ frequency at each difficulty index. These values, which are quite close, indicate that both sets maintain similar difficulty index relative to the test taker's ability in the classes. In addition, both averaged difficulty indices indicate moderate values, which is an encouraging result since a moderate value of difficulty index means that the questions are not too easy nor too difficult.

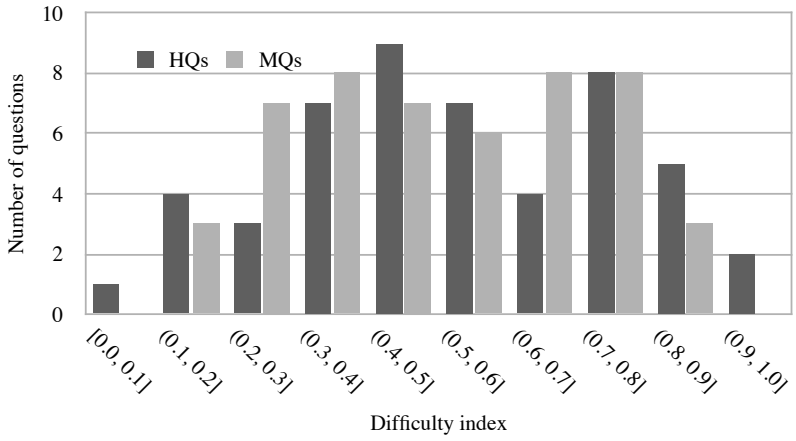

Figure 3.4: Difficulty index distribution of the MQs and HQs

Discrimination index The discrimination index indicates how well each item can discriminate test takers in terms of their ability. It ranges from -1 to 1 and the higher the value, the more discriminating the item is. For calculation of the discrimination index, we divided the test takers into three groups according to their total test scores. Given a ranking list of test takers based on their test scores, we define the top 27% of test takers as an 'high' group and the bottom 27% of test takers as a 'low' group. The rest is defined as a 'medium' group. We used the 27% boundary value for the high and low group determination following Kelley (1939). The discrimination index of a question item i (D_i) is

then calculated with Equation 3.3.

$$D_i = (H_i - L_i)/n \tag{3.3}$$

where H_i and L_i indicate the number of test takers in the high and low groups who correctly answered the question item i, and n is the total number of test takers in all groups (high, medium and low groups). An item is considered acceptable if its discrimination index is greater than or equal to .20 (Brown, 1983). Out of the 50 questions, 37 (74%) MQs have the discrimination index more than or equal to .20, and thus considered acceptable, while 40 (80%) HQs do. The small difference on those two values shows that the MQs achieve a comparable level to HQs in terms of discriminating high and low proficiency test takers.

Item-total correlation coefficients An item-total correlation analysis is performed to check if any item in a test is inconsistent with the averaged behaviour of the other items. Those kind of items are better to be discarded. It is simply the Pearson's product moment correlation coefficient of an individual items with the scale total calculated from the remaining items. As with the discrimination index, it ranges from -1 to 1 and the higher the value, the more correlated the item is with the test as a whole. A low item-correlation coefficient (less than .2) shows that the item is not measuring the same construct measured by the other items in a test, thus should be discarded (Streiner and Norman, 2003). Out of the 50 questions, 76-88% of the MQs has the correlation coefficient more than .2, whereas 80-100% of the HQs does. This result is expected from the HQ, however, it also shows that only a small portion of the MQs should be dropped.

Neural Test Theory analysis In NTT, we first need to decide how many ranks we want, and it usually lies within 1–10. As the same as in calculating the discrimination index, we grouped the test takers into 3 ranks: 'high', 'medium' and 'low'. We further separated the analyses for the set A and set B since they included different question items and were answered by different test takers. Table 3.5 shows the expected number of test takers in each latent rank of the question sets.

Categorisation of ICRP for the evaluation ICRP (item category reference profile) is a feature of NTT representing the probability that the test takers in a certain rank select a certain question option in their responses to a certain question item. The ICRP is obtained by a statistical learning process as summarised in the Section 2.4. Since it shows how test takers in each rank behave against each option of the question, it can be used to clarify

Table 3.5: Latent rank estimation for MQs and HQs (rank size)

question set	no. of test takers in ranks			
	low	medium	high	total
MQs(A)	12	15	13	40
MQs(B)	13	12	14	39
HQs(A)	12	12	16	40
HQs(B)	12	13	14	39

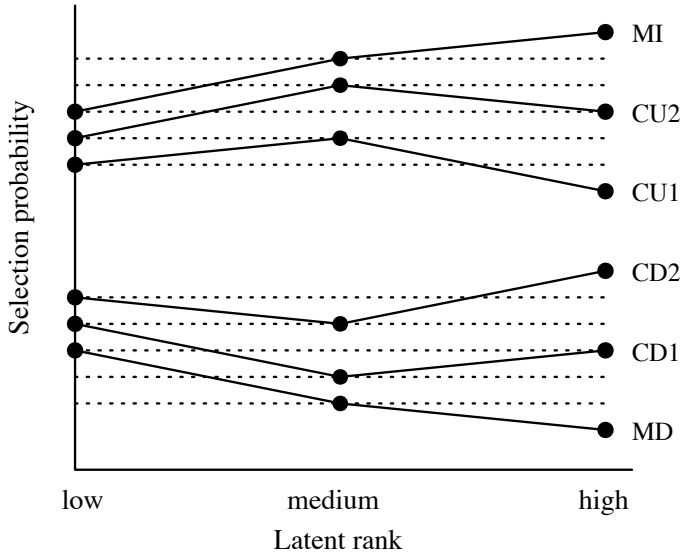

Figure 3.5: Six ICRP categories based on the magnitude relations between probabilities of test takers to select an option in a proficiency rank

the validity of the question options. For instance, it can be used to clarify if a distractor *correctly* deceives the low-proficiency test takers more compared to the high-proficiency test takers.

Since we have three latent ranks of the test takers in this evaluation, given an option, we have three independent magnitude relations between probabilities Ps that the test takers select the option in the corresponding rank, namely $P(\text{low}) \gtreqless P(\text{medium})$, $P(\text{medium}) \gtreqless P(\text{high})$ and $P(\text{low}) \gtreqless P(\text{high})$. According to their combination of the magnitude relations, we can classify the ICRP into six categories as shown in Figure 3.5: monotonically increasing (MI), monotonically decreasing (MD) and convex upward (CU1 and CU2) and convex downward (CD1 and CD2). The MI option has a trend spanning from the bottom-left to the top-right as shown in Figure 3.5. More strictly, its probability

scores should be $P(\text{low}) < P(\text{medium}) < P(\text{high})$, meaning that this type of option tends to be more selected by the high-rank test takers than the medium and low-rank test takers. The MD option has the opposite tendency and other four have mixed tendency of the MI and MD options.

As a correct answer, the MI options are favourable, since they tend to be more selected by the high-rank test takers than the medium- and low-rank test takers. They are expected to be able to discriminate test taker proficiency. On the other hand, the MD options are least favourable as the correct answer, since they discriminate the test takers in the wrong way; the higher ranked test takers have less probability of selecting this option correctly than the lower ranked test takers. The convex options show intermediate behaviour between the MI and MD options. Among three independent probability relations, the CU2 and CD2 options display two correct relations in terms of being a correct answer, for instance a CU2 option correctly represents the relations $P(\text{low}) < P(\text{medium})$ and $P(\text{low}) < P(\text{high})$ but fails for $P(\text{medium}) < P(\text{high})$. Likewise, a CU1 option correctly represents only the relation $P(\text{low}) < P(\text{medium})$. The same applies to the CD2 and CD1 options. Based on the number of correctly represented probability relations between ranks, we can say that the CU2 and CD2 options are better than the CU1 and CD1 options as a correct answer in measuring test taker proficiency.

As for the distractor, the MD options are most favourable since the role of distractors is deceiving the test takers into selecting them instead of the correct answer; the options that tend to be more selected by the lower ranked test takers are good distractors. Such options should show a decreasing curve similar to the MD options in Figure 3.5. Thus, distractors have the opposite order in goodness to the correct answer: the MD options are the best, followed by the CU1 and CD1 options, then the CU2 and CD2 options. The MI options are the worst options as being distractors.

Analysis of correct answers Table 3.6 shows the number of correct answers in each ICRP category. The table shows that the majority of correct answers in the MQ sets belongs to the MI category as similar to those in the HQ sets. This result is encouraging since the MI category is favourable for the correct answer.

Table 3.6 also indicates that there are in total six question items with the MD correct answer (the least favourable category for a correct answer) in our MQ sets. We calculated their difficulty index to see if those question items tend to be difficult (Figure 3.4) and found that these items with the MD correct answer are relatively more difficult than those with the MI correct answer; the average difficulty index of the former is .36 whereas that of the latter is .56.

Table 3.6: Distribution of correct answers across ICRP categories

question set	MI	CU2	CD2	CU1	CD1	MD	total
MQs(A)	13	2	4	1	2	3	25
MQs(B)	17	1	1	2	1	3	25
HQs(A)	19	3	1	0	0	2	25
HQs(B)	11	6	2	2	2	2	25

Exametrika also produces test reference profile (TRP) that is calculated by a weighted sum of ICRPs of correct answers (Shojima, 2007)[18]. The TRP summarises the overall tendency of a set of question items by representing an expected number of correctly answered items for each latent rank as shown in Figure 3.6. For example, medium-rank test takers are expected to correctly answer 15 question items in the set HQs(B). Notice that all four TRPs show the same tendency; TRP increases as the rank becomes higher. It implies that the test takers in the higher rank are expected to obtain a higher score than those in the lower rank. This result is encouraging since it means that the MQ set is comparable to the HQ set in appropriately discriminating test taker proficiencies.

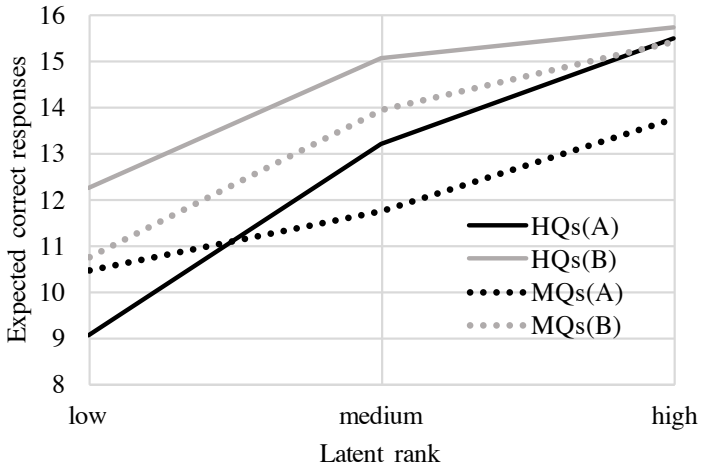

Figure 3.6: Test reference profile of the MQs and HQs

Analysis of distractors In contrast with the correct answers, the MD options are the most favourable for distractors and MI options are the least favourable. Table 3.7 shows

[18]We used uniform weighting in this study, i.e. the TRP was calculated by a sum of the ICRPs of correct answers.

the number of distractors in each ICRP category. We can see from the table that the majority of distractors in the MQ sets belongs to the MD category in contrast with the correct answers in Table 3.6. We found the same tendency in the HQ sets. This finding is promising because the numbers of the MD distractors are larger than those of other categories in all question sets.

Table 3.7: Distribution of distractors across ICRP categories

question set	MI	CU2	CD2	CU1	CD1	MD	total
MQs(A)	9	9	5	12	7	33	75
MQs(B)	14	3	2	10	4	40	73
HQs(A)	13	4	5	8	6	39	75
HQs(B)	15	2	5	6	7	34	69

Analysis of question with 'bad' question options To investigate the peculiar behaviour of the question options in the least favourable categories, i.e. the MD correct answer and the MI distractors, we further analyse the question items with those 'bad' options. According to Tables 3.6 and 3.7, there are six MD correct answers and 23 MI distractors. Since some of them are used in the same question items, we have in total 21 question items to be investigated. As a result, they are categorised into five groups based on their possible reasons.

(1) Multiple correct answers (MCA)

In this case, one or more distractors could be appropriate as the correct answer due to their closeness in meaning to the target word. Potential synonyms of the target word and the correct answer should have been ruled out from the distractor candidates when generating a question, but unfortunately, our dictionary (WordNet) happened to fail in having described that they were synonyms. In other words, this case could happen as a result of insufficient dictionary coverage.

One example is the distractor 'substantial' for the target word 'essential' in the following reading passage excerpt.

> ... It also allows for the book to lay flat, which is an essential feature of any cookbook. ...

The correct answer for this question item is 'basic and fundamental' with the distractors: 'substantial', 'of an obscure nature', and 'virtual'. In the evaluation result,

the correct answer 'basic and fundamental' belongs to the CU2 category; its ICRP increases from the low to medium latent ranks and decreases toward the high latent rank. On the other hand, the distractor 'substantial' belongs to the MI category which is the best as a correct answer but the worst as a distractor; its ICRP monotonically increases according to the latent ranks. It means that this particular distractor deceived the higher proficiency test takers more than the lower ones. One explanation is that 'substantial' and 'essential' share a common meaning which is why the higher proficiency test takers were deceived. Based on the Oxford Thesaurus of English[19], 'essential' is indeed one of the synonyms of 'substantial'.

There are also cases where the distractors are considered appropriate in the context of the reading passage although they are not necessarily a synonym of the target word. Here is one example. This question is asking for the closest meaning of 'proof' in the following reading passage excerpt among the choices: A. 'justification', B. 'symptom', C. 'establishment', and D. 'cogent evidence'.

> ... First real-life proof of principle that IVF is feasible and effective for developing countries ...

In this example, the distractor 'justification' belongs to the MI category, which means that the higher rank test takers tend to select this option. In the above reading passage excerpt, 'justification' could be the correct answer since it means 'an acceptable reason for doing something'[20], sharing a meaning with 'proof' in the above reading context.

Moreover, the probability of selecting the correct answer 'cogent evidence' decreases with the increase of the rank. The correct answer 'cogent evidence' is actually quite obvious, since 'evidence' definitely means 'proof'. Adding the modifier 'cogent' in front of 'evidence' might, however, have confused the test takers since they were most likely not aware of its meaning. According to the JACET 8000 word difficulty level, 'cogent' is considered as the most difficult word (difficulty level category *Others*[21]). One possible explanation is that the higher ranked test takers thought that the 'cogent evidence' option was a trap; the modifier 'cogent' might have varied the meaning of 'evidence' from its 'proof' meaning. Whereas

[19]http://www.oxforddictionaries.com/definition/english-thesaurus/substantial

[20]Merriam Webster dictionary, http://www.merriam-webster.com/

[21]Difficulty level category *Others* includes words over level 8, non-English words, and misspelling. We made sure that this word is neither non-English nor misspelling, so we treat this word as word over level 8 which is the most difficult level in JACET 8000.

the lower ranked test takers noticed that ‘evidence’ meant ‘proof’ and thus went with that option without much caring about its modifier.

(2) Unfamiliar word sense (UWS)
This case happens when the option is a word with an unfamiliar word sense to the test takers. This example asks for the closest meaning of ‘digit’ in the following reading passage excerpt among the choices: A. ‘trouble’, B. ‘skill’, C. ‘figure’, and D. ‘population’.

> . . . In each of today’s problems you will be given two sets of 6 two digit numbers. . . .

The correct answer for this item is ‘figure’, however, this option belongs to the MD category which means that the higher ranked test takers tend to not select this option compared to the other rank test takers. Moreover, the ICRP of the distractor ‘trouble’ increases with the increase of the latent rank (MI category). One possible explanation is that the correct answer ‘figure’ is less familiar when being used as the ‘digit’ meaning, whereas ‘trouble’, even though it has no relation with the target word, is related to the word ‘problem’ that appears in the reading passage.

(3) Collocationally odd word (COW)
This case happens when the correct answer is collocationally odd as the replacement of the target word in the reading passage. The vocabulary question here does not ask for the best replacement; instead, it asks for the closest in the meaning of the target word. However, the test takers often tend to find the correct answer by replacing the target word with all options and select the one which best replaces the target word. This example asks for the closest meaning of ‘spearheaded’ in the following reading passage excerpt among the choices: A. ‘educated’, B. ‘departed’, C. ‘were the leader of’ and D. ‘plowed’.

> . . . Jefferson County Mental Health has spearheaded the counseling effort, making sure victims receive the assistance they need. . . .

The correct answer is the option ‘were the leader of’. This is a multiple-word option generated from the definition of the target word. The ICRP of the correct answer ‘were the leader of’ monotonically decreases with the increase of the latent rank, whereas the ICRP of the distractor ‘departed’ monotonically increases with the increase of the rank. From a grammatical point of view, it is clear that the distractor ‘departed’ is better suited as the replacement for the target word than the correct answer ‘were the leader’.

(4) More reasonable word (MRW)
This is a case when one of the distractors looks better suited as the replacement of the target word in the reading passage, regardless of its meaning. This case might happen when the test takers do not know the meaning of the target word, but they do know the meaning of some or all the options. In other words, the test takers, similarly to the COW cases above, try to find the answer that best replaces the target word. One example is the distractor 'volatile' for the target word 'viable' and correct answer 'feasible' in the following reading passage excerpt.

> ... they described the bomb as a viable device capable of causing death or serious injury. ...

The distractor 'volatile' belongs to the MI category, meaning that its ICRP monotonically increases according to the latent ranks. This could happen because the word 'volatile' is highly reasonable in modifying the word 'device' in this context. Since the test takers probably did not know the meaning of the target word due to its high difficulty level, they selected the option related to 'device' which is suited to replace the target word, regardless of the meaning of the target word.

(5) Other
There are a few case which do not fit into the above groups; it is difficult to find consistent reasons for them. For example, this question item asks for the closest meaning of 'immeasurably' among the choices: A. 'firstly', B. 'plainly', C. 'beyond measurement' and D. 'to double the degree'.

> ... But Perez darted in and out of trouble long enough helped immeasurably when left fielder Endy Chavez shortcircuited a second-inning Cardinals rally by ...

The correct answer is the option 'beyond measurement', which should be obvious since it even shares substrings with the target word 'immeasurably'. However, the ICRP of the distractor 'plainly' is monotonically increasing (the MI category) as the increase of the latent rank. This might be because the distractor 'plainly' shares its suffix '-ly' with the target word.

Table 3.8 shows the breakdown of the types of investigated question items with at least one 'bad' option that showed the peculiar ICRP behaviour (MD for correct answers and MI for distractors). The COW and MRW question items make 38% of the total items. As explained above, in these types of question items, the test takers tend to select

a distractor that looks better suited as the replacement of the target word in the reading passage, regardless of its meaning. This means that even though the generated vocabulary question does not ask for the best replacement, the test takers in our experiments tend to look for answers which best replace the target word, especially if they do not know the meaning of the target word.

Table 3.8: Distribution of question with 'bad' option types

possible reason/ category	#questions
Multiple correct answers (MCA)	3
Unfamiliar word sense (UWS)	4
Collocationally odd word (COW)	2
More reasonable word (MRW)	6
other	6

3.5.2 Evaluation 3.2: similarity with human-made questions

In this evaluation, we mixed HQs and MQs and asked human experts to distinguish between two types of questions. This evaluation is similar to Turing test (Turing, 1950), evaluating to what extent the machine-generated questions are similar to those created by humans as the gold standard.

Experimental design

We used the same question items with Evaluation 3.1, but only half of them. By equally dividing QS_A2 and QS_B2 of Evaluation 3.1 into 5 sets, we created the question sets as shown in Table 3.9. The order of question items in a set was kept as the same as in experiment 1. In total, we had 25 HQs and 25 MQs to be evaluated by each evaluator in this experiment. We asked 8 English teachers (non-native English speakers: 4 Japanese and 4 Filipinos) to evaluate the question items by answering a questionnaire shown in Figure 3.7.

Results and discussion

We collected 400 responses in total, comprising 200 responses for the MQs and HQs. In what follows, we analyse the responses in relation to the questionnaire items.

Table 3.9: Configuration of question sets (Evaluation 3.2)

question set	contents	
	#HQ	#MQ
QS_1	4	6
QS_2	4	6
QS_3	6	4
QS_4	7	3
QS_5	4	6

(1) Is this question machine or human-generated?
(1: definitely created by machine; 5: definitely created by human)

(2) Which component made you decide that the question is either machine or human-generated (your answer for Q1)?

(3) Can this question be used in an actual test?
(1: definitely no; 5: definitely yes)

(4) What do you think is the difficulty level of this question is?
(1: very easy; 5: very difficult)

(5) Do you have any specific suggestions for this question? (optional)

Figure 3.7: A questionnaire for each question item in expert-based evaluation

Distinction between MQs and HQs In questionnaire item (1), we asked the evaluators to distinguish if the question is human-made or machine-generated using the 1–5 point scale. Scale 1 means that the question is definitely created by a machine, while 5 means it is definitely created by a human. We calculated the average scores given by the evaluators for each question item. The result is presented in Figure 3.8.

All human-made questions (dark colour bars) received an average score higher than or equal to 3, while 16 out of 25 of the machine-generated questions did. This suggests that at least those 16 machine-generated questions are hardly distinguishable from the human-made questions.

Rationale behind MQ-HQ judgement In the 200 responses to questionnaire item (2) for the MQs, there are 337 mentions to the reason of judgement. The breakdown is shown in Table 3.10 with the results of judgement. The column 'human-made' denotes the judgements of when the score greater than or equal to 3 in questionnaire item (1),

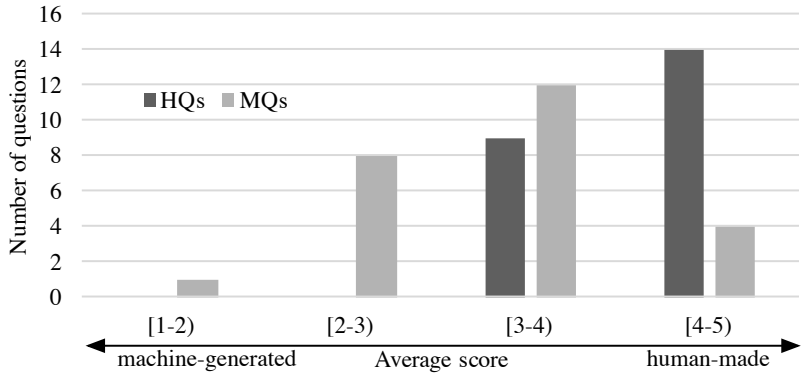

Figure 3.8: Result of distinguishing MQ and HQ by the experts

while the column 'machine-generated' denotes those with the score less than 3. Table 3.10 indicates that the reading passage and the correct answer tend to be more mentioned as the rationale for judging an item as human-made rather than as machine-generated. This suggests that these components are prominent in judging the question items as human-made.

Table 3.10: Rationale behind MQ-HQ judgement of MQs as judged by the experts

component	human-made	machine-generated	total
reading passage	82	53	135
correct answer	76	39	115
distractor	44	43	87

Usability of questions Questionnaire item (3) asked for the usability of the questions in a real test on a 5 point scale, with 5 meaning 'it can definitely be used in the actual test'. The result is presented in Figure 3.9.

Again, all human-made questions (dark colour bars) received an average score greater than or equal to 3, while 18 out of 25 machine-generated questions did. The figure clearly indicates that the human-made questions are better than the machine-generated questions in term of the usability in a real test. However, the result also suggests that more than half of the MQs were considered usable in a real test.

Item difficulty We asked for item difficulty on a 5 point scale with 5 being a very difficult question in questionnaire item (4). The results show that both MQs and HQs

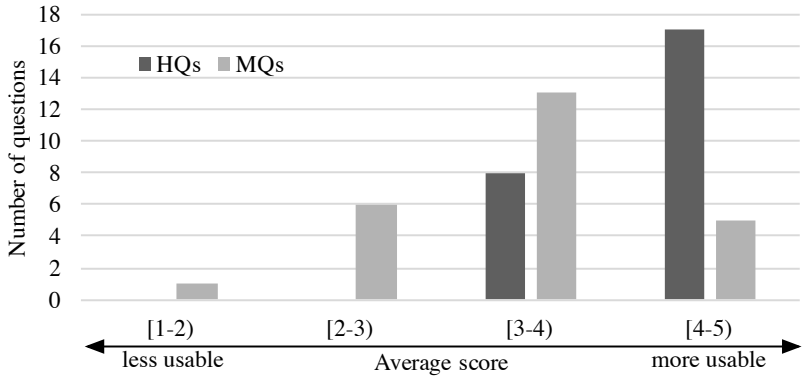

Figure 3.9: Usability of HQ and MQ for a real test, judged by the experts

have a medium difficulty level; the mean of the item difficulty for the MQs is 3.3 (SD = .70), while that for the HQs is 3.2 (SD = .77). The Pearson correlation coefficient was calculated between the item difficulty gained from questionnaire item (4) and that calculated from the difficulty index in Evaluation 3.1 (Figure 3.4)[22] to see to what extent both item difficulties from different perspectives correlated to each other. This resulted in positive correlation with .69 of the correlation coefficient (t = 4.56, df = 23, $p < .05$) for the HQs and .56 (t = 3.21, df = 23, $p < .05$) for the MQs. We can conclude that there is no big difference between the item difficulties from the test taker and the teacher perspectives.

General comments The evaluators provided various comments on the questions in response to questionnaire item (5). There are in total 75 comments for the HQs, and 85 comments for the MQs. We categorised these comments into one of the four categories: 1. positive (e.g. 'It has a well-written passage, excellent distractors and an appropriate answer choice.'), 2. negative (e.g. 'All of the distractors are not reasonable enough.'), 3. positive+negative (e.g. 'The passage is relevant to the word being identified, but I feel that the last sentence needs paraphrasing in order for it to be more comprehensible.'), and 4. neutral (e.g. 'Test takers can really answer this question if they would look for the context clues in the sentence.'). Table 3.11 shows the distribution of the comments for the HQs and MQs.

The following are the comments for each question component for MQs. Negative comments for the reading passage include: 'too long', 'too many clauses and run-on sentences', 'seems like it is retrieved from the web', and so on. Note that we did not tell

[22]The item difficulty was calculated by subtracting the difficulty index from one.

Table 3.11: Distribution of general comments from the expert

type	positive	negative	positive+negative	neutral	total
HQs	27	17	13	18	75
MQs	14	45	11	15	85

the evaluators that our passages had been retrieved from the Internet. On the positive side, the evaluators mentioned that the passage 'makes sense', 'well-written', 'gives enough context clues' as their motives to judge the MQ items as human-made.

Their negative comments on the correct answers include that the correct answer is 'too obvious thus makes the question too easy', 'could not find which one is the correct answer', 'it needs improvement', and so on. On the positive side, they mentioned that the correct answer is 'appropriate', 'advanced', 'well-made' and the like.

The distractors of the MQs also gained positive and negative comments. 'Too easy', 'out-of-context of the passage', 'neither reasonable nor challenging enough' are some of the negative comments mentioned. On the positive side, the distractors are said to be 'reasonable', 'serving their purpose well' and 'quite distracting'.

Discussion

In summary, based on the ratings on HQ-MQ distinction (Section 3.5.2) and usability in a real test (Section 3.5.2), it is clear that the HQs are better than the MQs. Dividing the question items into 'good' and 'bad' ones in the middle of the scale (3), we have only 16–18 out of 25 (64%–72%) good MQs, while all HQs are good.

We further analyse the bad and good-rated MQ items based on their ICRP categories that were introduced in 3.5.1. The good-rated items here are items with a rating greater than or equal to 3 on both HQ-MQ distinction and usability ratings, while the bad-rated items have a rating less than 3. Tables 3.12 and 3.13 show the distribution of the ICRP categories for the correct answer and distractors of the bad and good-rated items. Note that the total number of distractors does not always sum up to three times the number of questions since some distractors might not be selected at all by the test takers.

Table 3.12 indicates a tendency that the MI correct answers appear in the good-rated question items more than in the bad-rated items, while it indicates an opposite tendency for MD correct answers. Note that the MI options are favourable for the correct answers. This means that the result of the ICRP analysis based on the test taker responses (Evaluation 3.1) is consistent with the judgement of the human experts (Evaluation 3.2).

The similar tendency is found in the distribution of the ICRP categories for distractors,

Table 3.12: Distribution of the ICRP categories for correct answers in good and bad rated items

question items	MI	CU2	CD2	CU1	CD1	MD	total
good-rated	13	0	2	0	1	0	16
bad-rated	4	0	1	0	1	3	9

Table 3.13: Distribution of the ICRP categories for distractors in good and bad rated items

question items	MD	CU1	CD1	CU2	CD2	MI	total
good-rated	28	8	3	1	1	6	47
bad-rated	12	2	1	2	1	8	26

as shown in Table 3.13. Note that for the distractors, the most preferable ICRP category is MD and the least is MI, which is the opposite of the correct answer. However, the difference between the good and bad-rated items in terms of the proportion of the MD and MI categories is not so large compared with the correct answer (Table 3.12). A possible explanation is that when the evaluators gave ratings to the items, they would always consider the correct answer but might not always look at the distractors since they were more difficult to evaluate.

Chapter 4

Distractor improvement

The evaluation in the previous chapter revealed that problematic distractors were the main source of low-quality question items generated by a machine. This chapter focuses on reducing the number of problematic distractors to improve the quality of machine-generated question items.

Rodriguez (2005) stated that the quality of multiple-choice questions relies heavily on the quality of their options. His claim is supported by Hoshino (2013), who noted that test takers tended to employ a choice-oriented strategy when working on multiple-choice questions. Therefore, the quality of the question options, especially the distractors (incorrect options), affects the quality of the question, as inappropriate distractors enable the test takers to guess the answer easily (Moser et al., 2012) or cause them to unnecessary confusion.

Nevertheless, few studies on automatic question generation have focused on distractor generation. Haladyna (2004) pointed out that generating distractors was the most difficult part of multiple-choice question generation. As in the manual writing of questions, developing appropriate distractors remains a difficult task in automatic question generation. Some studies have generated distractors for fill-in-the-blank language questions using simple techniques such as random selection from words in the same document (Hoshino and Nakagawa, 2005), employing a thesaurus (Sumita et al., 2005) and collecting similar candidates of the target word in terms of their frequency and dictionary-based collocation (Liu et al., 2005).

Other studies have employed more advanced techniques and resources for distractor generation, mostly for the fill-in-the-blank English vocabulary questions. For instance, Pino and Eskenazi (2009) and Correia et al. (2010) used graphemic (morphological and orthographic) and phonetic variants of the target word as distractor candidates. Correia et al. (2010) employed lexical resources to filter distractor candidates considering the

target word's synonym, hyponym and hypernym. Sakaguchi et al. (2013) utilised common learner errors that were constructed from error-correction pairs on a language learning site, Lang-8[1]. In each pair of corrections, the error was a candidate distractor for the target word. Zesch and Melamud (2014) applied context-sensitive lexical inference rules to generate verb distractors that are not semantically similar to the target in the fill-in-the-blank context but might be similar in another context. More recently, Jiang and Lee (2017) proposed the use of a semantic similarity measure based on the word2vec model (Mikolov et al., 2013) for generating plausible distractors of Chinese fill-in-the-blank vocabulary questions.

In this study, we implemented a distractor generation method introduced by Jiang and Lee (2017) as a baseline because their work is the latest state-of-the-art method that targets the most similar task to the current study. Although their method generates Chinese fill-in-the-blank vocabulary questions, the method is independent of the language because it takes a corpus-based approach. We can hence adapt the method for English by replacing the corpus. Another difference is the question type to generate, i.e. fill-in-the-blank questions versus closest-in-meaning questions. These questions differ in whether the target word is present in the options as a correct answer (fill-in-the-blank) or present in the reading passage (closest-in-meaning). There is no difference in the characteristics of the distractors in both types of vocabulary questions.

4.1 Method: distractor generation

In the following, we describe the baseline in detail, followed by our proposed method. We then compare the two methods as summarised in Table 4.1. For all methods, distractor generation consists of three steps: 1) distractor candidate collection, 2) distractor candidate filtering and 3) distractor candidate ranking.

4.1.1 Baseline method

Distractor candidate collection and ranking To collect distractor candidates, Jiang and Lee (2017) extracted all the words in the Chinese Wiki corpus and ranked them on the basis of their various similarity criteria to the target word. The similarity criteria consist of the difficulty level (frequency-based) similarity, spelling similarity, PMI-based word co-occurrence with the target word and word2vec–based word similarity. They ranked the candidates according to each criterion and evaluated the results. Their evaluation

[1] http://lang-8.com

Table 4.1: Baseline vs proposed method for distractor generation

step	baseline (Jiang and Lee, 2017)	proposed
selection	all words in English Wiki Corpus with the same POS as the target word	(1) synonym of co-occurrence words in the passage (2) sibling words in the taxonomy and synonyms of synonyms (3) words in the JACET 8000 list with the close level to the correct answer
filtering	trigram filtering	criteria by Heaton (1989) and synonym filtering
ranking	word2vec–based semantic similarity between the target word and a distractor candidate	GloVe–based semantic similarity between the target word and a distractor candidate, and between the correct answer and a distractor candidate, and word collocation

showed that the word2vec–based criterion outperformed the others, thus in this study, we implemented this criterion for collecting the distractor candidates.

Jiang and Lee (2017) trained a word2vec model on the Chinese Wiki corpus[2]. Because we adapted their method for English vocabulary questions, we used a word2vec model pre-trained on English Wikipedia[3].

Distractor candidate filtering Jiang and Lee (2017) filtered the ranked distractor candidates to remove candidates that are also considered to be an acceptable answer. They examined whether the distractor candidates collocate with the words in the rest of the carrier sentence[4], by filtering based on the trigram and dependency relations.

- Trigram filtering: the trigram is formed from the distractor candidate and its two adjacent words (the previous and following words) in the carrier sentence. We implemented this filtering without modification in our implementation for English vocabulary questions.

- Dependency relation filtering: the implementation by Jiang and Lee (2017) considers all the dependency relations with the distractor as a head or child. We implemented this filtering with a small corpus[5], but the filtering did not remove any

[2]They used about 14 million sentences from Chinese Wikipedia.

[3]Available at https://github.com/idio/wiki2vec/, the model consists of 1,000 dimensions, 10 skipgrams and no stemming

[4]In their paper, they use the term "carrier sentence" because the text is usually only a sentence, not necessarily a reading passage as in the present study.

[5]Available as a package in the Natural Language Toolkit (NLTK).

candidates. Hence, we decided not to implement this filtering.

The three highest ranked candidates after the filtering were chosen as the final distractors for the question.

4.1.2 Proposed method

Distractor candidate collection We collected the distractor candidates from two main sources that reflect two different relations with the target word. The first source is synonyms of the words in the reading passage that have the same part-of-speech and tense as the target word, with an assumption that those words share the same topic of the reading passage. The second source is siblings of the target word in the WordNet taxonomy. Because siblings share the same hypernym, the siblings of the target word should share a similar meaning but also have a certain difference in meaning.

In addition to these two sources of distractor candidates, we utilise the JACET 8000 word list (Ishikawa et al., 2003) as the third source of distractor candidates. We consider JACET 8000 suitable for generating English vocabulary questions because it has been compiled for the purpose of English learning. Our observations tell us that most distractors in human-made vocabulary questions have the same or almost the same level of difficulty as the correct answer. Thus, as the distractor candidates, the present study utilises the words in the JACET 8000 word list for which the level differs at most by two levels from that of the correct answer. For example, if the correct answer is level 4, the distractor candidates are collected from the words of levels 2–6.

Furthermore, to top up insufficient distractor candidates from WordNet, we also add the synonyms of synonyms and words related to the target word according to the *Merriam-Webster Dictionary*.

Distractor candidate filtering The collected distractor candidates are further filtered following English vocabulary questions writing guidelines (Heaton, 1989), which are summarised below.

1) Question options should have the same part-of-speech as the target word.
2) Distractors should have a word difficulty level that is similar to that of the correct answer.
3) Question options should have approximately the same length.
4) A pair of synonyms in the question options should be avoided.
5) Antonyms of the correct answer should be avoided as distractors.

6) Distractors should be related to the correct answer, or come from the same general topic.

Vocabulary questions in the present study ask for the word closest-in-meaning to the target word. Thus, the distractors must not have the same or a very similar meaning to either the target word or the correct answer. To guarantee that the distractors are not synonyms of the target word, we filter out synonymous candidates using the synonym list from WordNet and the *Merriam-Webster Dictionary* in addition to the criteria specified by Heaton (1989).

Distractor candidate ranking Although the distractors must have a different meaning from both the target word and the correct answer, they must also be able to distract the test takers from the correct answer. Because the present study focuses on the closest-in-meaning vocabulary question, distracting distractors should be similar to the target word or correct answer in some respects. Unlike fill-in-the-blank questions, where the target word and correct answer are the same, in the closest-in-meaning questions, we can utilise both the target word and correct answer to generate distractors so that the distractors are semantically close to the target word but far from the correct answer. To rank distractor candidates, the baseline adopts a word embedding-based semantic similarity measure. In contrast, the present study introduces a new ranking metric $r(c)$ that aggregates word embedding-based semantic similarity and word collocation information for ranking the distractor candidates c with respect to the target word (tw), reading passage (rp) and correct answer (ca), which is given by

$$r(c) = \text{rank}(\text{sim}(c, tw)) + \text{rank}(\text{col}(c, rp)) - \text{rank}(\text{sim}(c, ca)) \tag{4.1}$$

where $\text{sim}(w_i, w_j)$ is the semantic similarity between words w_i and w_j; $\text{col}(w_i, context)$ is a collocation measure of word w_i and its adjacent two words in the given $context$, and $\text{rank}(f(\cdot))$ returns the rank of the value of $f(\cdot)$ in descending order. We use ranks instead of their raw scores because they are easier to integrate into a single score.

To calculate $\text{sim}(w_i, w_j)$, the present study employs the cosine similarity of the word vectors derived by the word embedding GloVe algorithm rather than word2vec because it is more efficient (Pennington et al., 2014). We used the pre-trained GloVe word vectors[6]. We calculate the collocation measure $\text{col}(w_i, context)$ on the basis of the frequencies of two bigrams: (w_{i-1}, w_i) and (w_i, w_{i+1}) in the $context$. The bigram statistics were

[6]The word vectors were trained on Wikipedia 2014+ Gigaword 5, which consists of 6B tokens 400K vocabulary uncased words and provides 100 dimensional vectors. This resource is available at https://nlp.stanford.edu/projects/glove/

generated using the module provided by the Natural Language Toolkit (NLTK) Python Package[7] and the English Text corpora in the same package[8].

The idea behind Equation (4.1) is that we want to obtain a distractor candidate c that is similar to target word tw (a large $\mathrm{sim}(c, tw)$, i.e. has a high $\mathrm{rank}(\mathrm{sim}(c, tw))$), and frequently collocates with the adjacent words in the reading passage rp (a large $\mathrm{col}(c, rp)$, i.e. has a high $\mathrm{rank}(\mathrm{col}(c, rp))$), but is not similar to the correct answer ca (a small $\mathrm{sim}(c, ca)$, i.e. has a low $\mathrm{rank}(\mathrm{sim}(c, ca))$). Thus, we prefer distractor candidates with a smaller value of $\mathrm{r}(c)$.

4.2 Evaluation design

4.2.1 Question data

We selected 45 target words (TW 1–45) from real closest-in-meaning vocabulary questions collected from the ETS official site[9] and preparation books of TOEFL®iBT, which are published by the official TOEFL®organisation[10]. The selection was made such that the part-of-speech categories of the target words were balanced.

For each question, we used the three human-made distractors in the original TOEFL® question as a reference distractor set. We then determined two additional sets of three distractors using the baseline and proposed methods. For each automatically generated method, the set of three distractors was made by selecting from the top three candidates in the ranked candidate list of that method. The original reading passage and correct answer were used to automatically generate the distractors. In total, we prepared 135 question items with 45 items each set. The order of the distractors was randomised in each question.

We conducted two evaluations, test taker-based and expert-based evaluations; they are explained in the following sections.

4.2.2 Evaluation 4.1: test taker-based evaluation

The aims of this evaluation is to evaluate the validity of the distractor candidates when they are used in a real test setting. We administered the question set described above to

[7] http://www.nltk.org/_modules/nltk/collocations.html

[8] Corpora: Brown, ABC, Genesis, Web Text, Inaugural, Gutenberg, Treebank and Movie Reviews, available at www.nltk.org/nltk_data/

[9] www.ets.org

[10] The Official Guide to the New TOEFL®iBT, 2007, published by McGraw-Hill, New York.

English learners and evaluated the quality of the distractors based on their responses. We used a Latin square design to design the question sets, as shown in Table 2. For instance, in question set QS.A, the distractor sets for target words (TWs) 1 to 15 are generated by the baseline method and TWs 16 to 30 by the proposed method and TWs 31 to 45 are the original TOEFL®distractors created by humans.

Participants A total of 80 Japanese university undergraduate students participated in the experiment. We divided them into three student groups, G1, G2 and G3 according to their school class and administered a different question set to each student group. Table 4.2 shows the assignment of the question sets to the student groups.

Table 4.2: Configuration of the question sets (Evaluation 4.1)

student group	#students	question set	baseline	proposed	TOEFL®
G1	19	QS_A	TW#01–15	TW#16–30	TW#31–45
G2	23	QS_B	TW#16–30	TW#31–45	TW#01–15
G3	38	QS_C	TW#31–45	TW#01–15	TW#16–30

Experimental procedure The experiment was conducted in the form of an online test. The participants completed the test using their own computer, but each group worked on the question set together in the same classroom. The experiment comprised three sessions. In each session, one of the three groups worked on their assigned question set. A session lasted roughly 30–40 min.

4.2.3 Evaluation 4.2: expert-based evaluation

The aim of this evaluation is to evaluate the quality of the automatically generated distractors using a human expert. Because of limited resources, we asked one human expert to evaluate the questions. However, we believe his judgement is reliable because he is an experienced professional writer of these questions.

We provided the expert with an evaluation guideline that includes the question writing guidelines presented in 4.1.2. Given a target word and its corresponding reading passage, the expert evaluated each of the three distractor sets used in the test taker-based evaluation by giving it a score of 1–5, where 1 indicates very low quality and 5 indicates very high quality. We also provided an optional "comment" field where he could write any possible reasons for giving a low score to a set of distractors, or explain why distractors were problematic, if any existed in the set.

4.3 Result and discussion

4.3.1 Evaluation 4.1: test taker-based evaluation

Correlation with test takers' proficiency scores We calculate the correlation between test takers' scores on the questions and their TOEIC®scores, which we treated as the ground truth proficiency scores. The idea is that if the test takers scores on the machine-made questions show a strong correlation with their TOEIC®scores, then the machine-made questions are able to measure the test taker's proficiency. Table 3 presents the Pearson correlation coefficients between the scores.

The scores on the questions generated by both automatic methods show a lower correlation with their TOEIC®scores than those of the original TOEFL®questions. However, all methods indicate a low correlation in absolute terms. This is because the TOEIC®score reflects various kinds of English proficiency of the test takers, whereas the generated questions concern only their vocabulary.

Focusing on the vocabulary ability, we also calculated the correlation coefficients between test takers' scores on the machine-generated questions and those of the original TOEFL®questions. This yielded positive correlations with coefficients of .425 (t = 5.039, df = 38, $p < .05$) for the proposed method and .302 (t = 6.08, df = 37, $p < .05$) for the baseline. Evans (1996) categorised a correlation coefficient of .425 as 'moderate', and .302 as 'weak' correlation. This result is encouraging because it indicates that questions using the proposed method are more successful than those created using the baseline at measuring the test taker's proficiency with respect to the original TOEFL®questions.

Table 4.3: Pearson correlation coefficients between test scores (averaged for all group)

test scores	baseline	proposed	TOEFL®
TOEIC®	.290	.290	.342
TOEFL®	.302	.425	1.000

All p is less than .05.

Neural Test Theory analysis We also applied the NTT model (Section 2.4) to the student responses data in this experiment. The ICRP of the NTT can be used to clarify the validity of a distractor since it shows how test takers in each rank behave against each option of the question. For instance, it can be used to clarify if a distractor correctly deceives the low-ranked test takers more compared to the high-ranked test takers.

As described in Section 3.5.1, we further categorised the ICRP into six categories based on the magnitude of the relations between the probability that the option is selected by test takers in the corresponding student rank, as shown in Figure 3.5. The MD options are most favourable for distractors because the role of a distractor is to deceive a test taker into selecting it instead of the correct answer. Hence, the options that tend to be more selected by the lower-ranked test takers are good distractors. Such options should show a decreasing curve similar to the MD options in Figure 3.5. The MD options are the best for distractors, followed by the CU1 and CD1 options, then the CU2 and CD2 options. The MI options are the worst options for distractors. We counted the number of distractors in each ICRP category for each method as shown in Table 4.4.

Table 4.4: ICRP of the distractors for each method

	baseline	proposed	TOEFL®
MI	40	26	24
CU2	6	8	5
CD2	13	6	9
CU1	10	13	12
CD1	9	15	18
MD	46	51	50

The three methods produced more or less a similar number of the favourable MD distractors. However, as shown in the first row of Table 4.4, the proposed method produces fewer MI distractors (least favourable category for a distractor) than the baseline. The original TOEFL®questions, as expected, produced the smallest number of MI distractors. This result is encouraging because it shows that the proposed method succeeded in removing the problematic distractor candidates better than the baseline.

We further analysed the MI distractors to find the reasons these distractors were categorised as MI distractors. The probability of choosing the MI distractors increases as the test taker's rank increases. This indicates that more high-proficiency test takers are deceived by this distractor than low-proficiency test takers. Knowing the reasons helps us to understand the behaviour of each method when producing those distractors. We found that MI distractors could be classified into the following four categories.

`SYN` The distractor is a synonym of the target word or correct answer, e.g. the distractor 'support' for the target word 'assistance', where the correct answer is 'help'. We looked up two dictionaries[11] and if the distractors are listed as a synonym in one of

[11] *Oxford Dictionary of English* (www.oxforddictionaries.com) and the *Merriam-Webster Dictionary*

the dictionaries, we classified them in this category. This type of distractors is not appropriate for use in tests.

CON This distractor can be replaced in the given context, e.g. the distractor 'move' for the target word 'cope' when the correct answer is 'adapt' in the following context '. . . dinosaurs were left too crippled to cope, especially if, as some scientists believe . . . '. In this example, the distractor 'move' is neither similar to the target word nor the correct answer, but it fits in the context even though it results in a different sentence meaning. We checked the collocation of these words by querying Google search with a distractor and the word it is adjacent to in the reading passage as the query. This kind of distractor is reasonable because the test takers sometimes try to select the option that best replaces the target word in the reading passage.

REL This distractor is defined as a word related to the target word or correct answer in a dictionary[12], e.g. the distractor 'storm' is defined as a related word of the target word 'bombard' in the *Merriam-Webster Dictionary*. This kind of distractor is also reasonable.

UNK This type of distractor has an MI curve (monotonically increasing) without any convincing explanation such as one of the above three categories. These distractors can be safely used as a distractor although they are not very distracting.

Table 4.5: Categorisation of MI (problematic) distractors based on the possible reasons

	baseline	proposed	TOEFL®
SYN	11	0	0
CON	11	13	23
REL	2	1	0
UNK	16	12	1

Table 4.5 presents the number of the MI distractors categorised according to the above reasons. The results in Table 4.5 suggest the following conclusions.

1. CON is the principal reason for the MI distractors across all methods.

(www.dictionaryapi.com).

[12]The *Merriam-Webster Dictionary*'s related-word feature.

2. On the basis of the above categorisation, the SYN candidates should be rejected as distractors because they are potentially dangerous. None of the MI distractors from the original TOEFL®questions and proposed method belong to this category, whereas 11 out of 40 MI distractors of the baseline do and should be rejected. These results indicate that the proposed method succeeded in filtering the problematic candidates in this SYN category.

3. The CON and REL distractors are considered to be reasonable distractors, even though they are MI distractors. According to Table 4.5, 23 out of 24 original TOEFL®distractors belong to this category. The proposed and baseline methods respectively made 14 out of 26 (54%) and 13 out of 40 (33%) reasonable distractors in the CON and REL categories. This result is encouraging because more than half of the MI distractors generated by the proposed method are distracting distractors.

4.3.2 Evaluation 4.2: expert-based evaluation

We calculate the average judgement scores for all 45 test questions and the result is as follows: 2.867 (SD: 1.471) for the baseline, 4.333 (SD: .977) for the proposed method and 4.444 (SD: .840) for the original TOEFL®distractors. These average scores indicate that the distractors generated by the proposed method have better quality than those generated by the baseline and comparable quality with respect to the original TOEFL®distractors.

The human expert also wrote in a total of 135 comments for all questions. As explained in Section 4.2.3, the comments were specifically given to low-scored distractors; in other words, the distractors with comments were the problematic distractors according to the human expert. In total, 71 distractors from the baseline had comments, followed by 39 distractors from the proposed method, and 25 distractors from the original TOEFL®distractors. This result is encouraging because the proposed method produced fewer of problematic distractors than the baseline. We grouped the comments into the seven categories presented in Table 4.6 along with the number of distractors in each category for each method. Note that a distractor can belong to more than one category, so the row 'total number of problematic distractors' is not necessarily the sum of the distractors in all categories. A description of the seven categories follows.

1) **Too similar to the correct answer or the target word**. Comments explicitly state that a distractor is too similar to the correct answer, e.g. 'the distractor "overcome" is too close to the correct answer or "refined" is too similar to the correct answer.'

Table 4.6: Categorisation of problematic distractors by the expert

	baseline	proposed	TOEFL®
Too similar with the correct answer or the target word	26	3	1
Different word class	12	10	10
No relation to the correct answer	0	2	0
Different word difficulty	34	22	11
Antonym of the correct answer	3	0	1
Synonym pair	10	0	0
Others	0	3	2
total problematic distractors	71	39	25

2) **Different word class**. Comments concern the difference in the word classes of the correct answer/target word and distractors, e.g. 'the distractor "arise" is an intransitive verb, "digging out" is a verb phrase while "extending" and "destroying" are not.'

3) **No relation to the correct answer**. Comments concern criterion 6 in Section 4.1.2, e.g. 'the distractor "battlefield" is not related to the correct answer.'

4) **Different word difficulty**. Comments concern criterion 2 in Section 4.1.2, e.g. 'all the distractors are much more difficult than the correct answer.'

5) **Antonym of the correct answer**. Comments concern criterion 5 in Section 4.1.2, e.g. 'the distractor "separate" is an antonym of the correct answer.'

6) **Synonym pair**. Comments concern criterion 4 in Section 4.1.2, e.g. 'the distractor "repel" and "repulse" are synonyms.'

7) **Others**. Comments that are not classified into the above categories, e.g. 'the distractor "financially rewarding" should be changed, because it involves the same word as the correct answer.'

Comment category 1 is the severest category because if a distractor is too similar to either the correct answer or the target word, it makes the question invalid because it has more than one correct answer. In this respect, our result is encouraging because the proposed method generated fewer invalid questions in this category. The other comment categories are considered not to be as severe because they do not affect the validity of the question.

We calculated the correlation of the expert scores of the distractor sets. The correlation coefficients are .313 (statistically significant at $p \leq .05$) for the proposed method and human pair and .012 (not statistically significant) for the baseline and human pair. This indicates that the expert tends to give similar scores to the proposed method's distractors and the original distractors. Hence, the distractors generated by the proposed method look more similar to the human-made distractors than those generated by the baseline method from the expert's point of view.

4.3.3 Comparison of the expert and test taker-based results

As previously stated, the distractors with comments from the human expert are potentially problematic. We further analysed the behaviour of the distractors in each comment category (Evaluation 4.1) when they were used in the real test, i.e. in the test taker-based evaluation (Evaluation 4.2). We analysed only the responses of the high-proficiency test takers because it is important to determine why high-proficiency test takers were deceived by the problematic distractors. We summarise the results in the following.

1) **Too similar to the correct answer or the target word**. In six out of 30 distractors, no test takers selected the distractor in this category, whereas an average 30% of the high-proficiency test takers selected the other 24 distractors. The distractors in this category must be verified by human experts before they are used in a real test because there is a chance that they are actually the correct answers. One example is the distractor 'notion' in a question with the target word 'concept' and the correct answer 'idea'. In this example, 'notion' is a synonym of both the target word and correct answer. Out of 19 test takers, eight test takers chose the distractor 'notion', whereas only two test takers chose the correct answer 'idea'. Those eight test takers did not necessarily choose the wrong answer because the distractor 'notion' was indeed correct. This supports the claim that a question should not have distractors with a meaning that is too similar to either the target word or the correct answer.

2) **Different word class**. Less than 23% of the high-proficiency test takers selected 29 out of 32 distractors in this category. Although these distractors are not necessarily problematic, they are not very distracting.

3) **No relation to the correct answer**. Less than 30% of the high-proficiency test takers selected these distractors. As above, although these distractors are not necessarily problematic, they are not very distracting.

4) **Different word difficulty**. The distractors that are easier or more difficult than the other options will stand out and might not be selected by the test takers because their difficulty looks salient. This is supported by the fact that 59 out of 67 distractors in this category were selected by less than 30% of the high-proficiency test takers. Hence, the distractors in this category are not distracting distractors.

5) **Antonym of the correct answer**. More than 50% of the high-proficiency test takers did not select three out of four distractors in this category. If the distractor and correct answer are antonym pair, this suggests that one of them is wrong. This kind of distractor is not distracting.

6) **Synonym pair**. No high-proficiency test takers selected the distractors in this category in four out of 10 questions. This is most likely because they found out that a synonym pair in the options could not be a correct answer because a question has only a single correct answer. The distractors in this category should be verified by a human expert before they are used in a real test. One example is the distractors 'life-sized' and 'lifelike' for a question with the target word 'miniature' and the correct answer 'small'. No test taker out of 23 chose neither 'life-sized' nor 'lifelike'. The test takers probably figured out that they were synonym pair, so both could not be the correct answer. This gives evidence that there should not be a synonym pair in the options because the test taker can easily rule them out as a correct answer.

We are also interested in how the problematic distractors from the test taker-based evaluation (the MI distractors) were evaluated by the human expert. The MI distractors are considered problematic because the probability of choosing this distractor increases as the proficiency of the test takers increases. This indicates that more high-proficiency test takers are deceived by this distractor than low-proficiency test takers. Table 4.7 shows the intersection of the MI distractors in the test taker-based evaluation and the commented distractors by the human expert, which is categorised according to Table 4.6. Again, because a distractor can belong to more than one category, the sum of distractors in all categories can be larger than the 'intersection' row.

Table 4.7 shows that 60% (24 out of 40) of the MI distractors in the baseline were also considered problematic by the human expert. This indicates that the baseline distractors that were judged as problematic by the human expert also behaved inappropriately in the real test. However, the same conclusion could not be drawn from the MI distractors generated by the proposed method and from the original TOEFL®questions. Only 35% (9 out of 26) and 21% (5 out of 24) distractors generated by the proposed method and those from the original TOEFL®questions, respectively, were judged as problematic by

Table 4.7: Categorisation of problematic distractors (intersection between test taker and expert-based evaluations)

problematic distractor	baseline	proposed	TOEFL®
expert-based	82	38	25
test taker-based	40	26	24
intersection	24	9	5
Too similar with the correct answer or the target word	9	1	0
Different word class	3	2	3
No relation to the correct answer	0	0	0
Different word difficulty	16	5	1
Antonym of the correct answer	1	0	0
Synonym pair	1	0	0
Others	0	1	1

the human expert. This is an encouraging result because those distractors, despite their low score given by the human expert, may still be used in a real test, i.e. the problem is not severe.

Chapter 5

Controlling item difficulty

Toward the efficient measurement of learner proficiency, this study also focuses on integrating AQG and CAT. However, the item parameters, e.g. the item difficulty, should be estimated in advance during the question generation process so that there is no need to administer the items in a pretesting. This chapter describes our method for controlling the item difficulty in the AQG system in this study and its evaluation.

In Classical Test Theory (CTT), item difficulty or the difficulty index is defined as the proportion of test takers who correctly answered a question item; thus item difficulty can be estimated only after administering the item to the test takers. The item difficulty can then be used to determine whether a certain item is appropriate (not too easy nor too difficult) for a group with a certain ability. Nevertheless, administering items in a pretest is costly, and there is a risk of exposing the items before they are used for a real test. Moreover, a considerable number of test taker's responses are required to obtain a reliable estimate of item difficulty (Loukina et al., 2016). Hence, we proposed method to control the item difficulty intrinsically, i.e. control the item difficulty through the characteristics of the question components (target word, reading passage, correct answer, and distractor) during the item generation process.

Hoshino (2013) analysed the relationship between different types of distractors and difficulty of question items in the multiple-choice vocabulary cloze test. Three sets of question items were prepared with various types of distractors based on their relation to the other components of the question: (a) distractors with a paradigmatic relationship with the correct answer, (b) distractors with a syntagmatic relationship with the context sentence, and (c) those with no relationship with either the correct answer or the context sentence. Their result showed that question items with type (a) and (b) distractors were more difficult than those with type (c) distractors. More recently, we also conducted an investigation of several potential factors affecting item difficulty in the vocabulary

question in our study (Susanti et al., 2016). All investigated factors are related to the components of the vocabulary questions. Our study revealed that: 1) the word difficulty level of the question components contributed up to 60% of the item difficulty, and 2) distractors had the greatest impact on item difficulty.

Both Hoshino (2013) and our investigation indicated that the distractor played a critical role in the difficulty of multiple-choice vocabulary questions. Thus, the following three factors are considered in the present study: 1) Target word difficulty (TWD), 2) Semantic similarity between correct answer and distractor (SIM), and 3) Distractor word difficulty level (DWD). For each of these factors, we define two levels, as shown in Table 5.1.

Table 5.1: Investigated factors affecting the item difficulty

ID	factor	level	
TWD	Target word difficulty	low	high
SIM	Semantic similarity between correct answer and distractor	low	high
DWD	Distractor word difficulty level	low	high

5.1 Method: investigated factors

The following is a detailed explanation of how we implement these three factors in the automatic question generation system.

Target word difficulty (TWD) The first factor investigated for its effect on item difficulty is the target word difficulty. It is natural to assume that item difficulty is, to a certain degree, related to the difficulty level of the target word. There are a number of studies on determining the difficulty level of an English word (or reading difficulty), and they are based on various word features such as the frequency of occurrence of the word in a certain corpus and word length (Heilman et al., 2008; Petersen and Ostendorf, 2009). JACET 8000 (Ishikawa et al., 2003) is a radically new word list designed for Japanese English learners. JACET 8000 ranks the word list based on the word frequency in many corpora. The 8,000 words are divided into 8 groups of 1,000 words with each based on their word difficulty level. Throughout this study, we use the JACET 8000 level system to assign a word difficulty level to words in a question item, as participants in our experiments are all Japanese university students. JACET 8000 uses the 1–8 levelling system in which level 1

is the easiest word. A special level *Other* or *O* is defined for words over level 8, which include non-English words, misspelling words, etc.

In this study, we set the JACET 8000 level ≤ 3 as low and ≥ 4 as high level, after considering the target word difficulty distribution in our list of target words.

Similarity between correct answer and distractors (SIM) The second factor is the semantic similarity between the correct answer and distractors. Distractors in a multiple-choice question act as a lure to distract the test takers from finding the correct answer. The vocabulary question used in the present study asks for the word with the closest meaning to that of the target word; thus, a distracting distractor would be one with a close but different meaning from the correct answer. Hence, the similarity between the correct answer and distractors is considered to be a factor affecting item difficulty.

To calculate the similarity between the correct answer and distractors, the present study employs the cosine similarity of the word vectors derived by the word embedding GloVe (Pennington et al., 2014). We used the GloVe word vectors pre-trained on Wikipedia articles[1].

The distractor candidates are collected from several sources, including the synonyms of the co-occurrence words with the same part-of-speech as the target word in the reading passage, and the sibling and hyponym words of the target word in a lexical taxonomy. These candidates are filtered and the cosine similarity is further used for calculating the word vector similarity between the correct answer and each distractor candidate. The three candidates with the lowest similarity are chosen for the low-level distractors, and the three candidates with the highest similarity are chosen as the high-level distractors.

Distractor word difficulty level (DWD) The last factor is the word difficulty level of distractors. Distractors in multiple-choice questions have been shown to affect item difficulty (Susanti et al., 2016; Hoshino, 2013). The JACET 8000 list of words (Ishikawa et al., 2003) is used to determine the word difficulty level in this work, as it is a list designed for Japanese English learners, who were the participants in the experiment.

When generating distractors, the system eliminates irrelevant distractor candidates following several requirements explained in 3.4.2. These filtered distractor candidates are then ranked based on the semantic similarity and further divided into two levels of distractor candidates: low and high, as explained in Section 5.1. Next, the word difficulty level is retrieved for each distractor candidate based on the JACET 8000. The three can-

[1]Wikipedia 2014+ Gigaword 5,6B tokens 400K vocabulary size, uncased, 100d. Available at https://nlp.stanford.edu/projects/glove/

didates with the lowest level are further adopted as the low (easy) level distractors, and the three highest level candidates are adopted as the high (difficult) level distractors. If a distractor is composed of multiple words, we calculate the average of the JACET 8000 word difficulty levels of each word after removing the stopwords[2].

5.2 Evaluation 5: experimental design

The main purpose of this evaluation is to investigate the impact of the three potential factors on the item difficulty of the vocabulary questions. We asked high school students to take a test composed of vocabulary questions that were generated by a machine using different combinations of factor levels (low and high).

Questions We created all eight possible combinations of the three factors affecting difficulty (Table 5.1) with two levels each, as shown in Table 5.2. Question items were generated conformed to each combination. We prepared 24 question items for each combination in Table 5.2, generating 192 question items in total. Note that we used 192 different target words for these question items. We divided the 192 items into six question sets, taking into account the balance of the combinations and parts-of-speech of the target words. One question set consisted of 32 question items, with four items for each combination in a set. The target words were selected from the Oxford3000 words[3] and GSL[4] word lists.

Participants The experiment was conducted as an online test. The participant took the test using a computer. The experiment comprised three sessions, in each of which one of the three groups worked on the assigned question set. A session was 30 minutes long. All participants in each group worked on the question set together in the same classroom.

Experimental procedure The experiment was conducted as an online test. The participant took the test using a computer. The experiment comprised three sessions, in each of which one of the three groups worked on the assigned question set. A session was 30 minutes long. All participants in each group worked on the question set together in the same classroom.

[2]Stopwords corpus from NLTK (http://www.nltk.org/nltk_data/).
[3]https://www.oxfordlearnersdictionaries.com/wordlist/english/oxford3000/
[4]http://www.eapfoundation.com/vocab/general/gsl/alphabetical/

Table 5.2: Combinations of three factors affecting the item difficulty

Combination	Factor		
	TWD	SIM	DWD
LLL	low	low	low
LLH	low	low	high
LHL	low	high	low
LHH	low	high	high
HLL	high	low	low
HLH	high	low	high
HHL	high	high	low
HHH	high	high	high

5.3 Results and discussion

In total, 22,272 responses for all question items (116 students worked on 192 question items) were collected in the experiment. We calculated the test takers' score on our experiment by dividing the number of their correct responses by the total number of questions in the question set, i.e. 32. Pearson correlation coefficients were further calculated between the student's test scores and their scores of the latest English term exam in each group. Table 5.3 presents the correlation coefficients for all groups.

Table 5.3: Correlation of test taker's scores in in the experiment with their latest term exam scores (Evaluation 5)

question set	class	correlation coefficient	#students
QS_A	C_A	.405	21
QS_B	C_B	-.190	20
QS_C	C_C	.289	18
QS_D	C_D	.521	19
QS_E	C_E	.579	19
QS_F	C_F	.301	19
	average	.301	116

The p is statistically significant ($< .05$) for class A, D, E.

As we can see in Table 5.3, in all classes, we do not get strong correlations between the test taker's score in the experiment and their term exam scores. The correlation is particularly bad in class B, in which the correlation coefficient is negative. The error

analysis in the class B data shows that there are several extreme cases where the test takers with the high exam score did not perform well in our experiment, and vice versa. The average of the correlation coefficient is .419 by excluding class B, which is considered as a moderate correlation (Evans, 1996).

This chapter addresses the three research questions:

1. Can the item difficulty be controlled by the investigated factors?

2. Which among the investigated factors contributes the most to the item difficulty?

3. How do these factors affect the item difficulty across test takers with different proficiencies?

Each question is dealt with in the subsequent sections.

5.3.1 Can the item difficulty be controlled by the investigated factors?

To answer the first research question, we first look at the average item difficulty for each combination. If the average item difficulty for each combination is different, it means that we can control the item difficulty by varying the factors. First, we estimate the item difficulty and analyse the variance of the average item difficulty for each combination.

Estimating the item difficulty There are several ways of estimating item difficulty from the test taker's responses. In test theory such as Classical Test Theory (CTT) and Item Response Theory (IRT), the difficulty is defined as the likelihood of correct responses, not as the perceived difficulty nor necessary amount of effort (DeMars, 2010).

According to CTT, item difficulty, more commonly referred to as the difficulty index (P), is the proportion of test takers who correctly answer the item. Suppose an item is correctly answered by eight out of ten test takers. Then the CTT difficulty index is .8 (8/10). Hence, the item difficulty ranges from 0 to 1, with a higher value suggesting an easier item. In IRT, as explained by DeMars (2010), the item difficulty (often denoted with parameter b) represents the proficiency of the test takers, half of which are expected to answer the item correctly. For instance, if b = .2, about 50% of the test takers with proficiency = .2 would answer the item correctly. In contrast to CTT, a higher b value indicates a more difficult item.

We calculated the estimated item difficulty of all question items using both CTT and

IRT (using R[5] software and the lazyIRT package[6]). We found that the item difficulties estimated by CTT (P) and IRT (b) are strongly correlated (average r = .825). Hence, for further analysis, we used only the CTT difficulty (P). The descriptive statistics for the estimated item difficulties from CTT are presented in Table 5.4.

Table 5.4: Descriptive statistics for the estimated item difficulty

	P(CTT)					
	QS_A	QS_B	QS_C	QS_D	QS_E	QS_F
n	32	32	32	32	32	32
$\bar{x}$	.507	.509	.460	.454	.507	.434
sd	.091	.095	.083	.098	.089	.085
max	.687	.687	.562	.625	.656	.594
min	.375	.312	.219	.281	.312	.312
r with b (IRT)	.789	.782	.830	.840	.891	.816

Analysis of variance on combinations The purpose of analysis of variance (ANOVA) is to see if the differences in the mean difficulty index between combinations are significant. If they are different, it means that the item difficulty can be controlled using the combination of the three factors as explained at the beginning of this chapter.

Figure 5.1 shows the boxplot of the average difficulty index P for each combination. The boxplot shows that the means (red circles) are different for each combination. However, the difference varies greatly depending on the combinations. Hence, these differences in means could have come about by chance. We performed a one-way ANOVA on the combinations to see if the differences between them are statistically significant. We subsequently looked at the p of the ANOVA results to determine to what extent the differences between the means are significant.

We performed the ANOVA on 1) the eight combinations shown in Table 5.2 and 2) four regrouped combinations, as explained below.

- Eight combinations. The one-way ANOVA was performed on the eight combinations, yielding in a p less than .01. This indicates that the mean differences between the eight combinations are statistically significant at a significance level of .01, suggesting that the three factors did affect the item difficulty.

[5]https://www.r-project.org

[6]http://www.ms.hum.titech.ac.jp/Rpackages.html

- Four regrouped combinations. We reduced the combinations into four groups based on the number of 'high' factors: 1) `LLL`, 2) `MID-H1 (LHL + LLH + HLL)`, 3) `MID-H2 (LHH + HHL + HLH)` and 4) `HHH`. The result of ANOVA shows that the difficulty differences between these four new groups are statistically significant ($p < .01$). This indicates that setting the factors to high or low influences the item difficulty; to be more concrete, the items with more 'high' factors are more difficult than those with fewer 'high' factors. Therefore, item difficulty can be controlled by varying the investigated factors, which answers the first research question. Figure 5.2 shows the box plot of the reduced combinations.

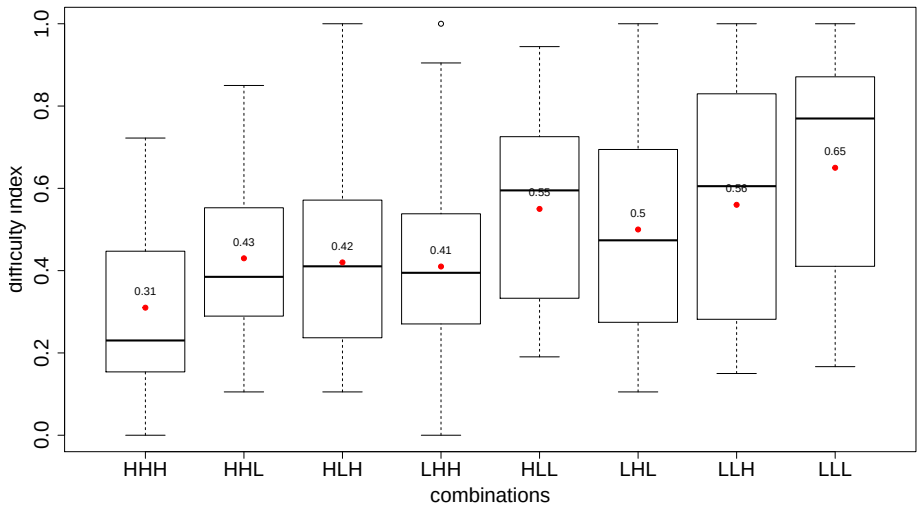

Figure 5.1: Box plot for the eight combinations of (item difficulty P)

5.3.2 Contribution of each factors

The following analysis presents the contribution of each factor towards explaining item difficulty. A three-way ANOVA was performed to determine if a three-way interaction effect exists among TWD, SIM, and DWD for explaining item difficulty, as well as to understand which factor contributes the most to item difficulty. The ANOVA result shows that there is no significant three-way interaction ($p \geq .05$), meaning that the investigated factors are independent of each other. The result also revealed that SIM has the biggest influence on item difficulty, followed by DWD and TWD which has about the same influence on the item difficulty (all $p \leq .05$). Looking at the proportion of variance (η^2), about

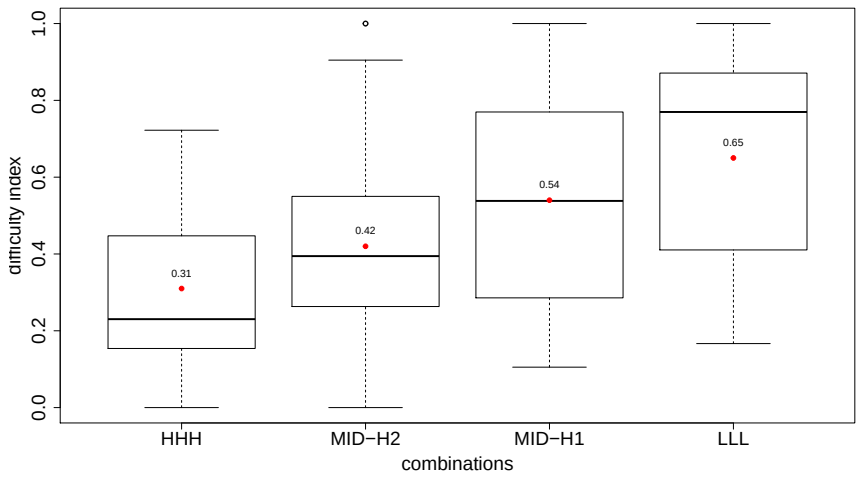

Figure 5.2: Box plot for regrouped four combinations (item difficulty P)

7% of the variability is explained by the SIM factor and 4.3% is explained by the other two factors. All the interaction factors only explained the variability by less than 0.1%.

Main effect In addition to the three-way interaction, there are three main effects that can be observed for each of the three factors. For example, the main effect of TWD is the difference between the means of the item difficulty for the two levels of TWD (TWD.high and TWD.low), ignoring the other two factors. The means for a three-factor experiment are often displayed in the form shown in Table 5.5.

Table 5.5: Item difficulty means values

		DWD. low	DWD. high	row mean
TWD. low	SIM. low	.652	.563	.608
	SIM. high	.502	.410	.456
	mean	.577	.487	.532
TWD. high	SIM. low	.548	.421	.484
	SIM. high	.425	.307	.366
	mean	.486	.364	.425
	column mean	.532	.425	

From the table, we can then calculate the main effect of each factor by subtracting the

means of the two levels, as presented in the following.

- TWD. $(\texttt{high} = .425 - \texttt{low} = .532) = -.106$
- SIM. $(\texttt{high} = (.456 + .366)/2 = .411) - (\texttt{low} = (.608 + .484)/2 = .546) = .135$
- DWD. $(\texttt{high} = .425 - \texttt{low} = .532) = -.107$

The large difference in the main effect above means that the factor affects on discerning between the high and low level, leading to a large impact on item difficulty. All three factors affect the item difficulty with a statistically significant p ($< .05$). In addition, the test takers were more affected by the factors related to the question options rather than the target word itself. It means that the test takers most likely only looked at the question options to determine the correct answer. This finding suggests that in the multiple-choice vocabulary questions, the question options are the most important factors for determining the question difficulty. This is supported by our another experiment with university students as the test takers. We conducted a similar experiment, using the reading passage difficulty instead of the target word difficulty (TWD) as a factor. The result showed that the distractors word difficulty affects the item difficulty the most. Furthermore, the result also suggested that the test takers hardly read the reading passage to answer the questions since the reading passage difficulty factor did not affect the item difficulty with a statistically significant p. This result has been published in Susanti et al. (2017).

5.3.3 Item difficulty in proficiency-based groups

The previous research question asked the impact of each factor on the overall item difficulty. The research question 3 further asks how the impact differs across different proficiency groups. We divided the test takers into three groups using the NTT. Table 5.6 shows the descriptive statistics for the three groups.

Table 5.6: Descriptive statistics of item difficulty *P* for proficiency-based groups

	low group	middle group	high group
n student	41	36	40
$\bar{x}$	.413	.518	.516
sd	.285	.320	.306
max	1	1	1
min	0	0	0

We carried out the three-way ANOVA test on each group to determine how the impact of the three factors on the item difficulty differs for each of the proficiency-based groups. The impact for all three groups are illustrated in Figure 5.3.

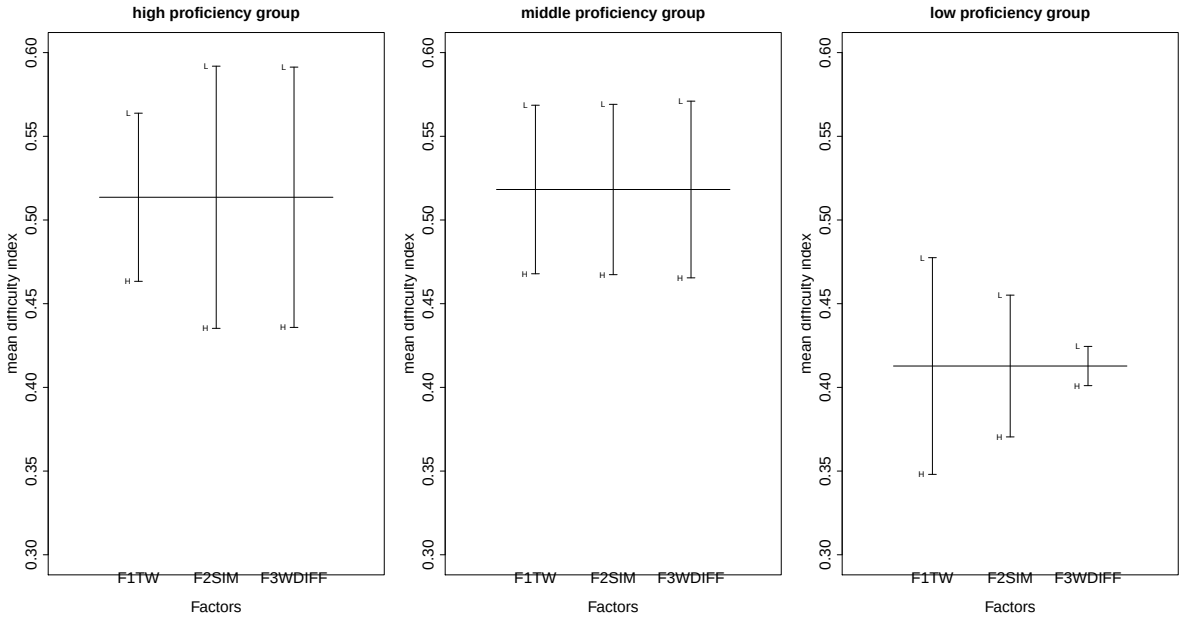

Figure 5.3: Mean differences across proficiency-based groups of test takers

High proficiency group As discussed in the previous sections, the general tendency of all test takers is that the SIM affects the estimated item difficulty the most. This tendency is retained in the high proficiency group of test takers, where SIM has the most significant influence on the item difficulty, followed by DWD and TWD (all $p \leq .05$), as shown in Figure 5.3. This finding suggests that the high proficiency students are more likely to get confused when the correct answer and distractors have a similar meaning to each other. This is reasonable because the test takers need to understand the meaning of every option distinctly to solve the item where the options are similar to each other.

Middle proficiency group For the middle proficiency group, DWD affects the item difficulty the most, followed by SIM and TWD (all $p \leq .05$). However, all three factors almost equally affect the item difficulty, as in Figure 5.3. This shows that the test takers in this group have a problem with the difficult target words. They also have problems even with familiar target words when the options are similar to each other and when the distractors are composed of difficult words.

Low proficiency group In contrast to the tendency of the middle and high proficiency groups, the test takers in the low proficiency group found the question items to be more difficult when TWD was high. The item difficulty is furthermore affected by SIM then followed by DWD (all p is significant under .05 except for DWD). This suggests that the low proficiency group could not solve the item when they have no idea about the target word at all, i.e. when the target word is difficult. Since they do not know the target word at all, the other two factors (SIM and DWD) do not really affect the item difficulty, as shown in Figure 5.3.

To sum up, the factors affecting item difficulty are different depending on the proficiency of the test takers. For instance, in our experiment, the TWD factor has the greatest impact on the item difficulty for the low proficiency group of test takers while the same factor has the least impact on the high proficiency group. Thus, to design the item difficulty, item writers must consider the proficiency of the test takers.

Chapter 6

Integration of AQG and CAT

This chapter describes the integration of AQG with CAT toward an efficient measurement of learner proficiency. As we stated in the Chapter 1, CAT aims at a precise and reliable measure of a test taker proficiency by presenting questions that have appropriate difficulty to measure their proficiency. It selects an item based on the test taker's proficiency evaluated after the response of each item. This is different to a linear test where all test takers take the same set of items in the same order. In the linear test, the high-proficiency test takers might get bored of answering a whole test if it contains only items that they consider easy, whereas the low-proficiency test takers might get frustrated by difficult items and give up working on the test seriously. Thus, CAT reduces the frustration of the test taker with items that are not suitable for his proficiency and it could improve the reliability of proficiency measurement of the test.

However, CAT leads to a considerable cost in the item development because it needs a large collection of previously administered items called the item bank. The item bank consists of items with their item parameters, e.g. item difficulty and item discrimination. Those item parameters are estimated from the test taker responses on the items in a test, and this process is called item calibration. This thesis focuses on the integration of the AQG with CAT without item calibration process, i.e. the item parameter is estimated during the question generation process so eliminating the need of administering the item beforehand to obtain item parameters.

Commonly, CAT makes use of Item Response Theory (IRT) model that represents the test takers and the items by a set of the model parameters. In the IRT models, a test taker's proficiency is represented as parameter θ, which usually follows the normal distribution. An item is represented by the following parameters: 1) item discrimination a, 2) item difficulty b, 3) guessing parameter c, and 4) upper asymptote d. In the present study, we use the 1-parameter logistic model where only the item difficulty b is considered. This

model is also called the Rasch model (Rasch, 1966). Thus, all the other parameters are set to their default values, which are 1 for the a and d and 0 for c.

6.1 Method: variation of item difficulty

To conduct an AQG-CAT integration simulation, we need three main elements: (1) items with their item parameters, (2) test taker's responses on every items, and (3) test taker's real proficiency to calculate the error of the proficiency measurement.

For the element (1) and (2), we re-used the result of the Evaluation 5 (Section 5.2). In Evaluation 5, first we generated six sets of questions where we control the difficulty of each question with three predefined factors related to the question components, as in Table 5.1. In total, we created 192 questions with eight combinations of the three factors. Next, we administered the machine-generated questions to 116 Japanese high school students as the test takers, who were divided into six classes in the experiment. As a result, we obtained test taker responses on 192 items from the Evaluation 5.

We use these items as element (1) and the test taker responses for the element (2) for the experiment in this chapter. In this thesis, we adapted a 1-parameter logistic model where only the item difficulty b is considered; thus, for element (1) we only need to define the item difficulty for every item. We use two types of item difficulty: 1) the estimated item difficulty from the test taker's responses as the gold standard and 2) the predefined item difficulty calculated from the component of the question, which is our proposed method. In the proposed method, the item difficulty of an item is calculated in advance without administering the item to the test takers beforehand. Accordingly, we prepared the following variations of item difficulty including the gold standard for the element (1).

a. EST item difficulty This is the item difficulty estimated from the test taker's responses in Evaluation 5, which is the gold standard for item difficulty. There are various ways to estimate the item difficulty from the test taker's responses, such as using CTT or IRT as described in Chapter 5. For a CAT, commonly the item difficulty is estimated from the test taker's responses; therefore, this first variation is the "gold standard" of a CAT simulation experiment.

b. REG item difficulty We calculate the predefined item difficulty by using linear regression in this variant. The question items used in this experiment were generated with the combination of three factors, i.e. TWD, SIM and DWD, as described in detail in Section 5.1. We need to estimate the regression coefficient to calculate the predefined item

difficulty to be used in this experiment. To apply the linear regression, we first calculate the numeric values for each factor, as following.

- TWD: we use the target word difficulty level of JACET 8000, with normalisation into range [0, 1].
- SIM: we use the average similarity score between the correct answer and distractors. The similarity score, range from 0–1, is calculated with cosine similarity on the GloVe word embedding.
- DWD: we use the average distractor word difficulty level of JACET 8000, with normalisation into range [0, 1].

Using the above numeric values for each factor and the CTT-based item difficulty calculated from the test takers' responses, we run the linear regression to get the regression coefficients. We further use the coefficients to calculate the predefined item difficulty for all items.

c. ORD item difficulty We represent the predefined item difficulty by an ordinal value from 1 to 4 in this variant, following the result of the analysis in the Chapter 5. In Chapter 5, each item is generated by one of the combinations of the three factors, as listed in Table 5.2. Further analysis shows that we can group the items into four difficulty levels based on the number of the 'high' factor in the combinations. The result also shows that the items with more high factors are more difficult than those with fewer high factors. The four groups are: 1) `LLL`: no high factor, 2) `MID-H1`: only one high factor, 3) `MID-H2`: two high factors and 4) `HHH`: all high factors. This group is illustrated in Figure 5.2.

We further represent the item difficulty of the items in each group with an ordinal value from 1 (`LLL`) to 4 (`HHH`). Hence, the items in the same group are assigned the same value representing the item difficulty.

d. AVG item difficulty We represent the item difficulty by averaging the item difficulty in the four groups of ORD. We estimated the item difficulty for all items and calculated the average of each group, as shown in the Figure 5.2. Since these values increases as the item becomes easier, we invert them for this variant. Thus, we have four average values, one for each group. Here, we represent the item difficulty of all items in each group with the item difficulty average value of the group. We further assign this average value to all items depending on the group the item belongs to. This item difficulty is an estimated item difficulty because it is calculated from test taker responses (averaged).

Finally, we use the latest term exam scores of the test takers in the Evaluation 5 for the element (3).

6.2 Experiment setting

In this thesis, we assess the feasibility of the AQG and CAT integration through a simulation-based experiment. We conducted the simulation using a CAT simulation package named `catsim` developed by De Rizzo Meneghetti and Thomaz Aquino Junior (2017) and adjusted it to our experiment setting. Using the test taker's responses in Evaluation 5, we conducted both CAT and linear test simulations.

We prepared the following simulation settings.

1. Linear test simulation (`LIN`). We use the EST item difficulty as described in Section 6.1. This setting serves as the baseline, and we use the same order of items in the test as in the Evaluation 5.

2. CAT test simulation. For the CAT test simulation, we initialise the proficiency of the test takers to a standard fixed value for all test takers ($\theta_0 = 0$). We use the maximum information selection strategy for item selection and the maximum-likelihood estimation for the proficiency re-estimation. The test stops when a total of 20 items is administered. The item bank size is 32, following the setting of Evaluation 5. The following shows the variations of the CAT simulation.

   - CAT with the EST item difficulty (gold standard) (`EST`).
   - CAT with the REG item difficulty (proposed) (`REG`).
   - CAT with the ORD item difficulty (proposed) (`ORD`).
   - CAT with the AVG item difficulty (`AVG`).

6.3 Result and discussion

In total, we have conducted five simulations including one linear test simulation as the baseline and four CAT simulations. We compared the result of all the simulations based on the mean squared error (MSE) calculated between the estimated proficiency at the end of each simulation and the proficiency of the students based on their latest term exam scores. The result is presented in Table 6.1. The smaller MSE indicates the better simulation because it means that the estimated item difficulty converges closer to the true

Table 6.1: Mean squared error (MSE) of the CAT simulations

group	#test takers	LIN	EST	REG	ORD	AVG
C_A	21	.142	.032	.054	.102	.049
C_C	18	.199	.072	.078	.146	.060
C_D	19	.156	.060	.069	.128	.044
C_E	19	.152	.024	.054	.087	.047
C_F	19	.328	.047	.090	.202	.106
	average	.195	.047	.069	.133	.061

proficiency of the test takers.

The LIN simulation, corresponding to a linear test, produced the biggest MSE (.195) compared to other four CAT simulations. This proves the effectiveness of an adaptive test to measure the test taker proficiency. We randomly sampled a single test taker in our experiment to show his progress during the test in each simulation. Figure 6.1 illustrates the test progress of a test taker in the LIN simulation. The x axis denotes the items and the y axis denotes the values for the item difficulty (orange line), the real proficiency of the test taker (black line) and the estimated proficiency of the test taker by the simulation (blue line). As shown in Figure 6.1, the LIN simulation presents the test taker with items in random order (following the order of items in the test) regardless of the test taker's proficiency.

Among the four CAT simulations, the EST simulation gave the smallest MSE (.047). This simulation is the gold standard of CAT because it uses the item difficulty calibrated from the test taker's responses. REG simulation used the predefined item difficulty estimated by the linear regression. Therefore, each item has a different value of item difficulty depending on the values of the factors it comprised. It means that this simulation adopts fine-grained item difficulty values. In the item selection step of CAT, it tries to present the test takers with item with a closest item difficulty value to the test takers' current proficiency. Thus, if the item difficulty values varies in the item bank, CAT could find a more appropriate item to present to the test taker. That being the case, the setting of the REG simulation is quite close to the gold standard, i.e. the EST simulation. The MSE of the REG simulation is bigger than that of the EST simulation (the gold standard), but fairly smaller compared to the LIN simulation (the baseline). This result is encouraging because the REG simulation that uses the proposed predefined item difficulty shows the

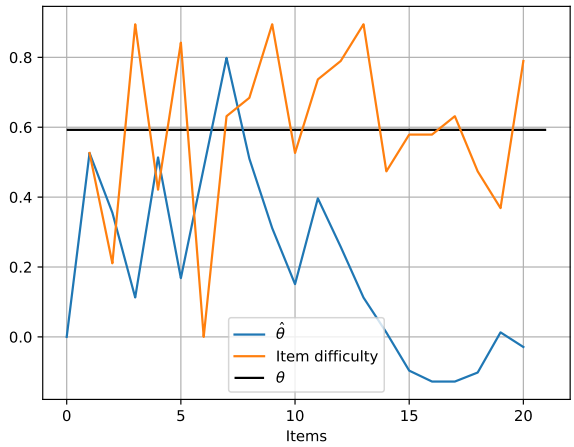

Figure 6.1: Test progress of a test taker (LIN simulation)

smaller MSE compared to the baseline.

Figure 6.2 illustrates the test progress of the same test taker of the LIN simulation progress (Figure 6.1) in the EST and the REG simulations. Unlike the LIN simulation, they present the test taker with items with item difficulty (orange line) that is close to the current estimation of the proficiency (blue line) during the test.

We also calculated the correlation between the estimated item difficulty from the test taker's responses (used in the gold standard, the EST simulation and the baseline, the LIN simulation) and the predefined item difficulty (used in the REG simulation). This yielded correlation coefficient r = .37 (statistically significant with $p < .01$), which is considered as a low correlation. However, this result is encouraging because it shows that even when the predefined item difficulty does not strongly correlate with the estimated item difficulty, it still performs good when incorporated into a CAT. This is proven by a smaller MSE of the REG simulation compared to the baseline LIN simulation, as shown in Table 6.1. A scatter plot between the two item difficulties are shown in Figure 6.3.

The ORD and AVG simulations use only four values of item difficulty. Figure 6.4 illustrates the test progress of the same test taker in the ORD and AVG simulation. The ORD simulation produced the biggest MSE (.133) among all CAT simulations. However, it still performed better compared to the baseline, the LIN simulation (MSE = .195). The AVG simulation, surprisingly, performed the best among the all proposed CAT simulations. Its MSE is even slightly smaller (.061) than that of the REG simulation that uses more fine-grained item difficulty values. This result is encouraging for incorporating a predefined item difficulty into CAT since it indicates that even only with four levels of the predefined item difficulty, it performed relatively better than the linear test.

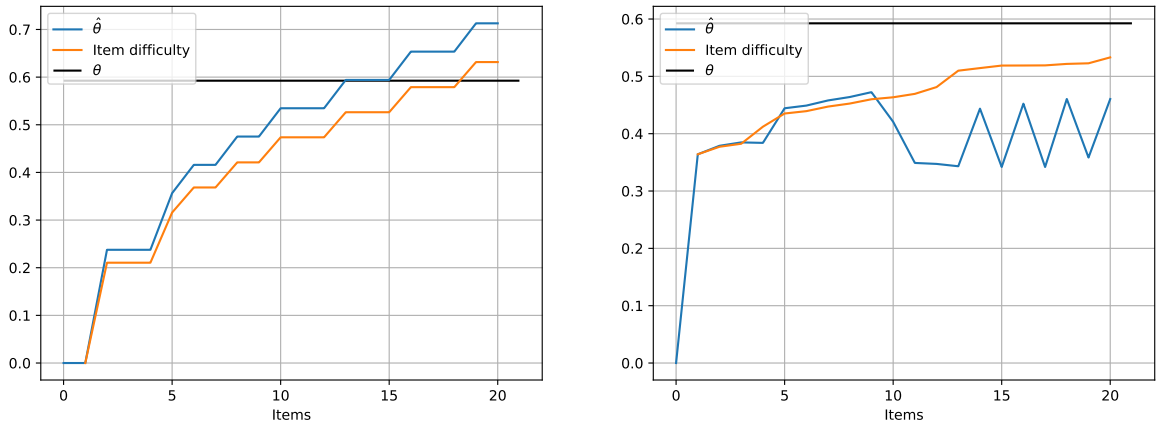

Figure 6.2: Test progress of a test taker (left: CAT EST simulation, the gold standard; right: CAT REG simulation)

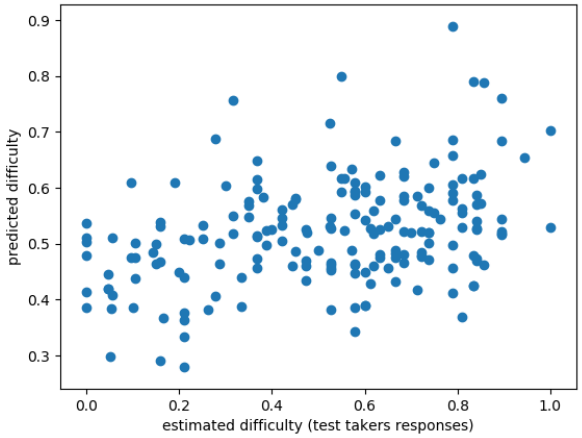

Figure 6.3: Scatter plot between the estimated and predicted item difficulty

In this chapter, we have conducted simulation-based experiments to assess the feasibility of integrating AQG into CAT without any item calibration. We calculated predefined item difficulty in various ways as described in Section 6.1. The result shows that all proposed CAT simulations using the predefined item difficulty (REG, ORD and AVG) produced smaller MSEs than the baseline LIN simulation. Moreover, their MSEs were also not further separated with the MSE of the gold standard, the EST simulation, which is a CAT simulation with the estimated item difficulty. Thus, we conclude that the integration of AQG and CAT is feasible from the experimental results.

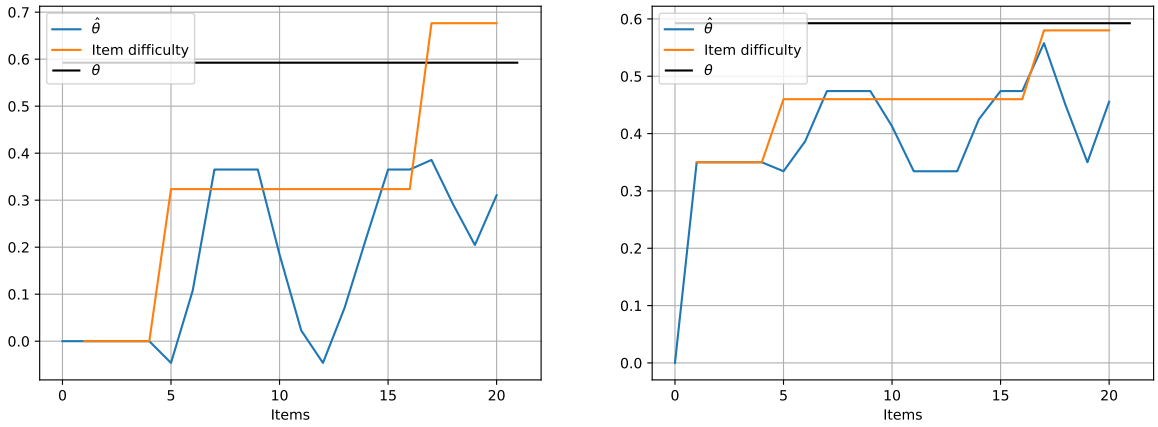

Figure 6.4: Test progress of a test taker (left: CAT `ORD` simulation; right: CAT `AVG` simulation)

Chapter 7

Conclusions

This thesis described a study on automatic generation of multiple-choice English vocabulary questions for efficient measurement of language learner proficiency. It consists of four topics: 1) automatic question generation (AQG), 2) distractor improvement, 3) question difficulty control, 4) integration of AQG into the computerised adaptive test (CAT).

In the first topic, we proposed a novel method for automatically generating English vocabulary questions, modelling the generated questions after the TOEFL®vocabulary questions. In this type of question, determining the word sense of the target word in a reading passage is crucial to creating the question options (the correct answer and distractors). We could use word sense disambiguation (WSD) techniques to identify the word sense. However, the accuracy of the state-of-the-art WSD method remains about 70-80%, which is far from satisfactory to the question generation task. Thus, we proposed a method that 'avoid' word sense disambiguation. Instead, we took an information retrieval approach where given a target word and one of its word sense, we search a passage that uses the target word with the given word sense.

We conducted two kinds of evaluation for assessing the quality of the generated questions: 1) test taker-based evaluation and 2) expert-based evaluation. In the test taker-based evaluation, we administered the machine-generated questions together with human-made questions to the real students. The analysis showed that the machine-generated questions were able to measure the proficiency of the students fairly comparable to the human-made questions.

In the evaluation of the question generation, the machine-generated question lacks a good quality mostly because of its distractors. Thus, the next topic focuses on improving the quality of the distractors. The proposed method extends the state-of-the-art method by introducing a new metric for ranking distractor candidates. The new metric aggregates both semantic similarity and word collocation information. The idea is to find distractors

which are close to the target word but far from the correct answer in their meaning, and also collocate with the adjacent words in the given context (the reading passage). We conducted test taker-based and expert-based evaluation for this topic, too. The result of the two evaluations showed that the proposed method succeeded in removing the problematic distractor candidates during their generation process compared to the baseline and generated distractors with comparable quality to the original human-made distractors. Further analysis showed that the problematic distractors from the proposed method can be used for a real test despite their low score from the human expert, which is an encouraging result.

We continued the direction of the AQG research to the integration with a computerised adaptive test (CAT), which is a type of test tailored according to the test taker ability. We proposed the integration of the AQG and CAT using predefined item difficulty, which can eliminate the need of item calibration.

We proposed to control the difficulty of the generated question with the three predetermined factors: 1) target word difficulty (TWD), 2) similarity between the correct answer and distractors (SIM) and 3) distractor word difficulty level (DWD). We again administered the generated questions with the various combination of the factors to the test taker. We analysed the collected data to answer three research questions: 1) whether the item difficulty can be controlled using the investigated factors, 2) which factor contributes the most to item difficulty and 3) how these factors affect the item difficulty across test takers with different proficiency. We performed ANOVA on the mean of estimated item difficulty to answer the first and second research questions and the result showed a statistically significant difference in the item difficulty. It means that the item difficulty can be controlled using the investigated factors. The analysis also revealed that the SIM factor contributes the most to the item difficulty. However, this tendency is not retained when the same analysis was conducted with different proficiency-based groups of test takers (the third research question). The factors affecting item difficulty are different depending on the proficiency of the test takers. For instance, in our experiment, the TWD factor has the least impact on the item difficulty for the high proficiency group of test takers while the same factor has the greatest impact on the low proficiency group. Thus, to design the item difficulty, item writers must consider the proficiency of the test takers.

We conducted simulation-based experiments on the AQG and CAT integration using two types of item difficulty i.e. the estimated item difficulty from the test taker's responses and the predefined item difficulty. The predefined item difficulty means that the item difficulty is calculated in advance, in the process of generating the question rather than from the test taker's responses at the pretesting. This way, there is no need to administer

a question to the test takers before using it for a CAT, enables the feasibility of integration of AQG with CAT with predefined item difficulty.

We evaluated the performance of the simulations by looking at the mean squared error (MSE) between the true proficiency of the test takers and the proficiency estimated by each simulation. The result showed that all proposed CAT simulations with the predefined item difficulty (`REG`, `ORD` and `AVG`) produced smaller MSEs than the baseline `LIN` simulation, which is a linear test simulation with the estimated item difficulty. Moreover, their MSEs were also not further separated with the MSE of the gold standard, i.e. the `EST` simulation, which is a CAT simulation with the estimated item difficulty. This is an encouraging result on the AQG and CAT integration with predefined item difficulty, which can eliminate the need of a pretesting.

This thesis concerns only a single type of questions, which is a closest-in-meaning vocabulary question. While the method in each chapter might be applied to another type of questions, some adjustment would be necessary. For instance, to control the item difficulty, this study determined three potential factors affecting item difficulty based on the author's observation and analysis, e.g. similarity between the correct answer and distractors. Heuristically, this factor could be applied for other type of questions such as open-ended questions (what, how, why, who questions) as well. However, unlike the current type of questions, open-ended questions commonly have sentences or phrases as their question options. Thus, an adjustment is necessary in the similarity calculation.

That being said, generalising the methods proposed in this study for wider type of questions is a future research direction. Evaluating the integration of AQG with CAT in a real setting is also another challenging task which is an important direction of the present study.

Bibliography

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. Predicting the difficulty of language proficiency tests. In Transactions of the Association for Computational Linguistics, volume 2, pages 517–529. Association for Computational Linguistics, 2014.

Isaac I. Bejar, Rene R. Lawless, Mary E. Morley, Michael E. Wagner, Randy E. Bennet, and Javier Revuelta. A feasibility study of on-the-fly item generation in adaptive testing. GRE Board Professional Report, 98(12-P), 2002.

R. F. Boldt and Roy Freedle. Using a neural net to predict item difficulty. ETS Research Report Series, 1996(2):i–19, 1996. ISSN 2330-8516. doi: 10.1002/j.2333-8504.1996.tb01709.x. URL http://dx.doi.org/10.1002/j.2333-8504.1996.tb01709.x.

FG Brown. Principles of educational and psychological testing 3rd Ed. New York: Holt, Rinehart and Winston, 1983.

James Dean Brown. Classical test theory. In Glenn Fulcher and Fred Davidson, editors, The Routledge Handbook of Language Testing, chapter 22, pages 323–335. Routledge, 2012.

Jonathan C. Brown, Gwen A. Frishkoff, and Maxine Eskenazi. 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, 2005.

Hua-Hua Chang and Zhiliang Ying. a-stratified multistage computerized adaptive testing. Applied Psychological Measurement, 23(3):211–222, 1999.

David Coniam. A preliminary inquiry into using corpus word frequency data in the automatic genera- tion of english language cloze tests. CALICO Journal, 14(2), 1997.

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

Kathleen Cotton. Classroom questioning. School Improvement Research Series, pages 1–10, 1988.

D. De Rizzo Meneghetti and P. Thomaz Aquino Junior. Computerized Adaptive Testing Simulation Through the Package catsim. ArXiv e-prints, July 2017.

Christine DeMars. Item Response Theory: Understanding Statistics Measurement. Oxford University Press, Inc., 2010.

ETS. The Official Guide to the New TOEFL iBT Internation edition. Mc Graw-Hill, 2007.

J. D. Evans. Straightforward statistics for the behavioral sciences. Pacific Grove, CA: Brooks/Cole Publishing, 1996.

Christiane Fellbaum. WordNet: A lexical database for English. A Bradford Book, 1998.

Jolene Gear and Robert Gear. Cambridge Preparation for the TOEFL Test 4th Edition. Cambridge University Press;, 2006.

Thomas M. Haladyna. Developing and Validating Multiple-Choice Test Items, 3rd edition. Lawrence Erlbaum Associates, 2004.

J. B. Heaton. Writing English Language Tests. Longman Pub Group, 1989.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. An analysis of statistical models and features for reading difficulty prediction. In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, EANL '08, pages 71–79, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. ISBN 978-1-932432-08-4.

Ayako Hoshino. Automatic Question Generation for Language Testing and its Evaluation Criteria. A Doctor Thesis, 2009.

Ayako Hoshino and Hiroshi Nakagawa. A real-time multiple-choice question generation for language testing -a preliminary study-. In Proceedings of the Second Workshop on Building Educational Applications Using NLP, pages 17–20. Association for Computational Linguistic, 2005.

Ayako Hoshino and Hiroshi Nakagawa. Predicting the difficulty of multiple-choice close questions for computer-adaptive testing. Special issue: Natural Language Processing and its Applications, page 279, 2010.

Yuko Hoshino. Relationship between types of distractor and difficulty of multiple-choice vocabulary tests in sentential context. Language Testing in Asia, 3(1):16, 2013. ISSN 2229-0443. doi: 10.1186/2229-0443-3-16.

S Ishikawa, T Uemura, M Kaneda, S Shimizu, N Sugimori, and Y Tono. JACET8000: JACET list of 8000 basic words. Tokyo JACET, 2003.

Shu Jiang and John Lee. Distractor generation for chinese fill-in-the-blank items. In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pages 143–148. Association for Computational Linguistics, September 2017.

T.L Kelley. The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology No. 30, pages 17–24, 1939.

Teuvo Kohonen. Self-organizing maps. Springer, 1995.

John Lee and Stephanie Seneff. Automatic generation of cloze items for prepositions. In Proceedings of Interspeech 2007, pages 2173–2176, 2007.

Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In Proceedings of the 5th Annual International Conference on Systems Documentation, pages 24–26, 1986.

Yi-Chien Lin, Li-Chun Sung, and Meng Chang Chen. An automatic multiple-choice question generation scheme for English adjective understanding. In Proceedings of Workshop on Modeling, Management and Generation of Problems/Questions in eLearning, the 15th International Conference on Computers in Education (ICCE 2007), pages 137–142, 2007.

Wim J. Van Der Linden and Gees A.W. Glas. Computerized Adaptive Testing: Theory and Practice. Springer, 2000.

Chao-Lin Liu, Chun-Hung Wang, Zhao-Ming Gao, and Shang-Ming Huang. Applications of lexical information for algorithmically composing multiple-choice cloze items. In Proceedings of the Second Workshop on Building Educational Applications Using NLP, EdAppsNLP 05, pages 1–8, Stroudsburg, PA, USA, 2005. Association for

Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1609829.1609830.

F.M Lord, M.R Novick, and Allan Birnbaum. Statistical theories of mental test scores. Oxford, England: Addison-Wesley, 1968.

Anastassia Loukina, Su-Youn Yoon, Jennifer Sakano, Youhua Wei, and Kathy Sheehan. Textual complexity as a predictor of difficulty of listening items in language proficiency tests. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3245–3253, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL aclweb.org/anthology/C16-1306.

Diana McCarthy. Word sense disambiguation: An overview. Language and Linguistics Compass, 3(2):537–558, 2009. ISSN 1749-818X.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, 2013. URL http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781.

Josef Robert Moser, Christian Gütl, and Wei Liu. Refined Distractor Generation with LSA and Stylometry for Automated Multiple Choice Question Generation, pages 95–106. Springer Berlin Heidelberg, 2012.

Roberto Navigli. Word sense disambiguation: A survey. ACM Computing Surveys, 41 (2):1–69, 2009.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. WordNet:: Similarity: measuring the relatedness of concepts. In Demonstration Papers at HLT-NAACL 2004, pages 38–41, 2004.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

Kyle Perkins, Lalit Gupta, and Ravi Tammana. Predicting item difficulty in a reading comprehension test with an artificial neural network. Language Testing, 12(1):34–53, 1995.

Sarah E. Petersen and Mari Ostendorf. A machine learning approach to reading level assessment. Comput. Speech Lang., 23(1):89–106, January 2009. ISSN 0885-2308.

Deborah Phillips. Longman Preparation Course for the TOEFL Test: iBT. Pearson Education Inc, 2006.

Juan Pino and Maxine Eskenazi. Semi-automatic generation of cloze question distractors effect of students' l1. In Proceedings of the SLaTE 2009 - 2009 ISCA Workshop on Speech and Language Technology in Education, Warwickshire, England, September 2009.

G. Rasch. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology,, 19(1):49–57, 1966. doi: 10.1111/j.2044-8317.1966.tb00354.x.

Michael C. Rodriguez. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educational Measurement: Issues and Practice, 24(2): 3–13, 2005. ISSN 1745-3992. doi: 10.1111/j.1745-3992.2005.00006.x.

Andre A. Rupp, Paula Garcia, and Joan Jamieson. Combining multiple regression and cart to understand difficulty in second language reading and listening comprehension test items. International Journal of Testing, 1(3-4):185–216, 2001.

Keisuke Sakaguchi, Yuki Arase, and Mamoru Komachi. Discriminative approach to fill-in-the-blank quiz generation for language learners. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistic, pages 238–242. Association for Computational Linguistic, August 2013.

Pamela J. Sharpe. Barron's TOEFL iBT Internet-Based Test 2006-2007 12th Edition with CD-ROM. Barron's Educational Series Inc, 2006.

Kojiro Shojima. Neural test theory. DNC Research Note, 07(02), 2007.

Kojiro Shojima, Tomoya Okubo, and Tomoichi Ishizuka. The nominal neural test model: A neural test model for nominal polytomous data. DNC Research Note, 07(21), 2008.

Simon Smith, PVS Avinesh, and Adam Kilgarriff. Gap-fill tests for language learners: Corpus-driven item generation. In Proceedings of ICON-2010: 8th International Conference on Natural Language Processing, pages 1–6, 2010.

David L. Streiner and Geoffrey R. Norman. Health Measurement Scales, 3rd edition. Oxford Medical Publications, Oxford, 2003.

Eiichiro Sumita, Fumiaki Sugaya, and Seiichi Yamamoto. Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions. In Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, pages 61–68, 2005.

Yuni Susanti, Hitoshi Nishikawa, Takenobu Tokunaga, and Obari Hiroyuki. Item difficulty analysis of english vocabulary questions. In Proceedings of the 8th International Conference on Computer Supported Education, pages 267–274. INSTICC, 2016.

Yuni Susanti, Takenobu Tokunaga, Hitoshi Nishikawa, and Hiroyuki Obari. Controlling item difficulty for automatic vocabulary question generation. Research and Practice in Technology Enhanced Learning, 12(25):16, Dec 2017.

Jonathan Trace, James Dean Brown, Gerriet Janssen, and Liudmila Kozhevnikova. Determining cloze item difficulty from item and passage characteristics across learner backgrounds. Language Testing, 1-24, 2015.

Alan M. Turing. Computing machinary and intelligence. Mind – A Quarterly Review of Psychology and Philosophy, LIX(236):433–460, 1950.

Bernard P. Veldkamp and Mariagiulia Matteucci. Bayesian computerized adaptive testing. Ensaio: Avaliao e Polticas Pblicas em Educao, 21(78), 2013.

D. J Weiss. Strategies of adaptive ability measurement (RR 74-5Z). Minneapolis: Psychometric Methods Program, Department of Psychology, University of Minnesota, 1974.

John H. Wolfe. Automatic question generation from text - an aid to independent study. In Proceedings of the ACM SIGCSE-SIGCUE technical symposium on Computer Science and Education, pages 104–112, 1976.

Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics (ACL), pages 133–138, 1994.

Torsten Zesch and Oren Melamud. Automatic generation of challenging distractors using context-sensitive inference rules. In Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, pages 143–148. Association for Computational Linguistics, June 2014.

Appendix A

Target word lists in the evaluation

Table A.1: Target word list (Evaluation 3.1)

immeasurably	disruption
retain	inaccessible
extracting	enhance
step	nourished
advocate	concept
obsession	assistance
digit	serve
maintain	essential
diverse	rare
despondent	inclination
fostered	gratify
inevitable	distort
assembled	proof
progressively	regulate
convey	corollary
spearheaded	seep
exhibit	supplant
happenstance	incriminated
ingredient	viable
alter	unsubstantiated
subtle	enactment
particular	perspective
interchange	relatively
staggering	cope
concur	primary

Table A.2: Target word list (Evaluation 4.1)

delicate	fluctuations
concept	excluded
relatively	despondent
rate	dispersal
sophisticated	consumed
ritual	illusory
readily	merge
project	fastidious
relevant	fostered
spheres	penetrate
progressively	accumulated
channel	distort
profound	disrupted
assistance	emit
sparsely	excavating
cope	bombard
miniature	diffused
inclination	synthesis
engaged	squander
prestigious	spearheaded
pose	
turbulent	
ingenuity	
enhance	
predominant	

Table A.3: Target word list (Evaluation 5)

picture	motion	system	employment	man	almost
rise	sensible	suspicious	freshly	associate	answer
nerve	usually	insert	connect	option	intention
commonly	fundamental	competitor	fortune	requirement	sum
world	ingredient	cycle	register	history	attempt
gain	program	store	test	mistake	final
terribly	cheap	develop	perhaps	exam	prospect
compose	standard	die	mind	error	routine
ripe	extend	unity	limit	expose	pretty
approval	enter	promptly	occupy	attract	impose
extra	consideration	pure	sharply	certainly	dump
organize	salary	retain	largely	arise	dull
permit	earn	withdraw	emphasis	essence	willingly
discuss	escape	divide	total	extremely	heavy
piece	conquer	still	deliver	foundation	branch
avoid	instance	death	direction	install	great
gentle	drop	thing	carry	totally	announce
realize	awful	discover	ban	always	huge
accommodation	substantial	assistance	formerly	primarily	demonstrate
ruin	pose	rude	prompt	significant	strictly
relieve	examine	eat	simple	national	allow
handle	swear	unite	strong	steady	generously
fix	really	former	precise	main	advance
actual	shorten	descent	upright	minor	depressed
ridiculous	approximately	draw	edition	large	peak
finish	encourage	compare	odd	engage	achieve
immediately	quickly	approximately	example	scare	calculate
mad	broad	ultimately	break	ensure	arrest
sincere	violent	wide	excess	grind	excessive
plan	previous	serious	call	begin	arise
newly	inquiry	reasonable	split	great	employ
resist	restrict	invest	confess	bear	thoroughly

Appendix B

English test scores distribution of the test takers

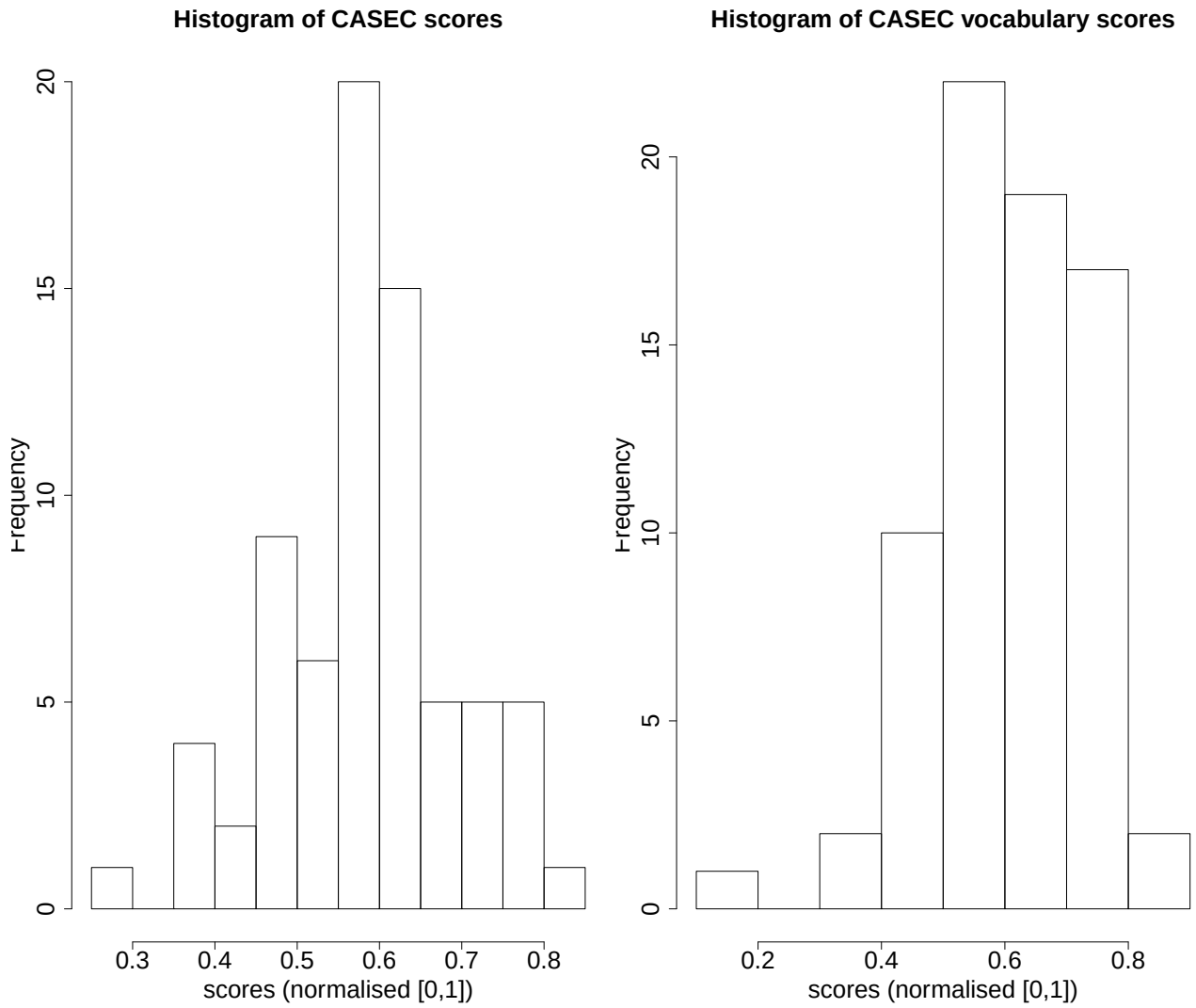

Figure B.1: English test scores (left: CASEC total scores; right: CASEC vocabulary section scores) of the test takers in Evaluation 3.1

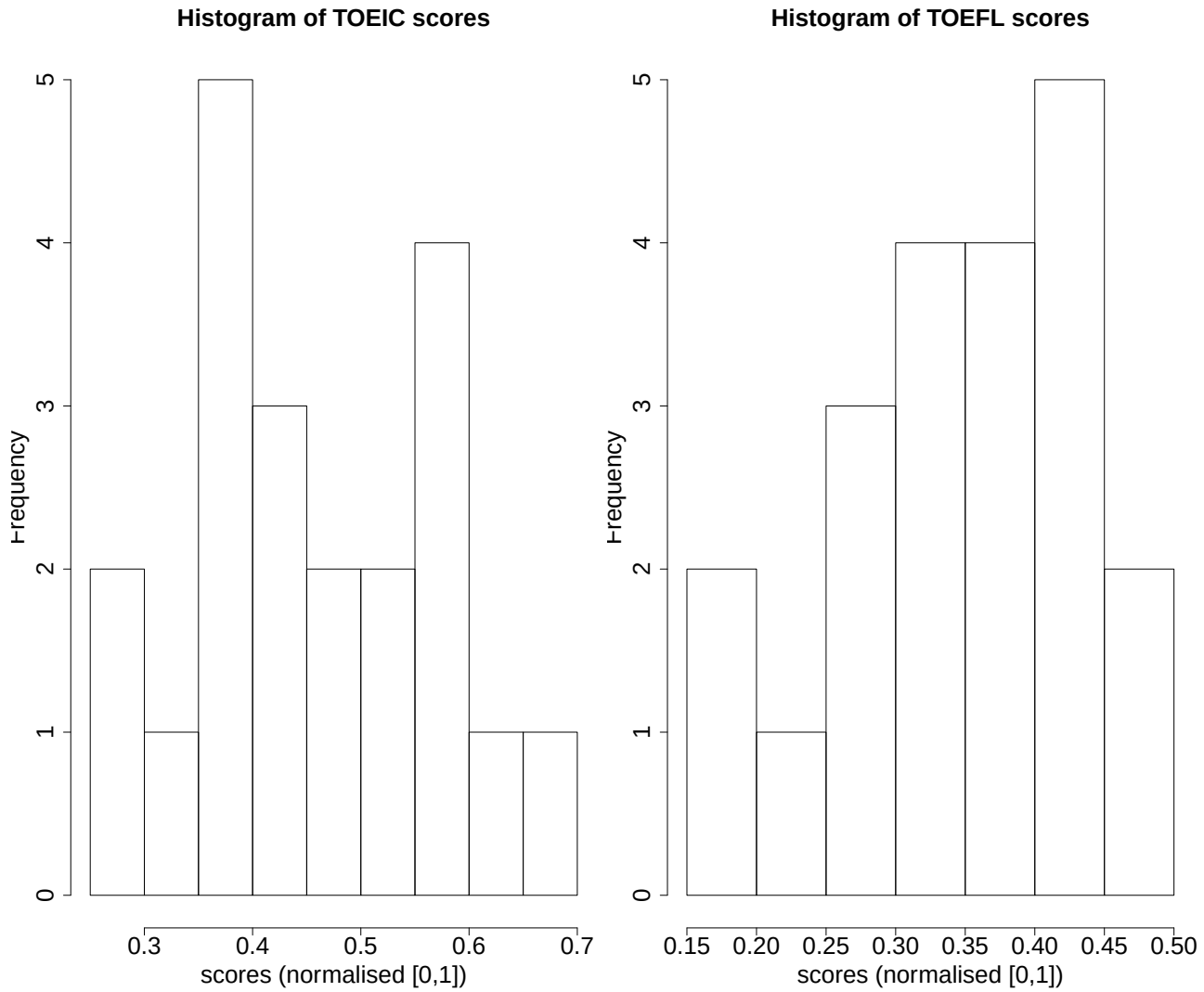

Figure B.2: English test scores (left: TOEIC scores; right: TOEFL scores) of the test takers in Evaluation 3.1

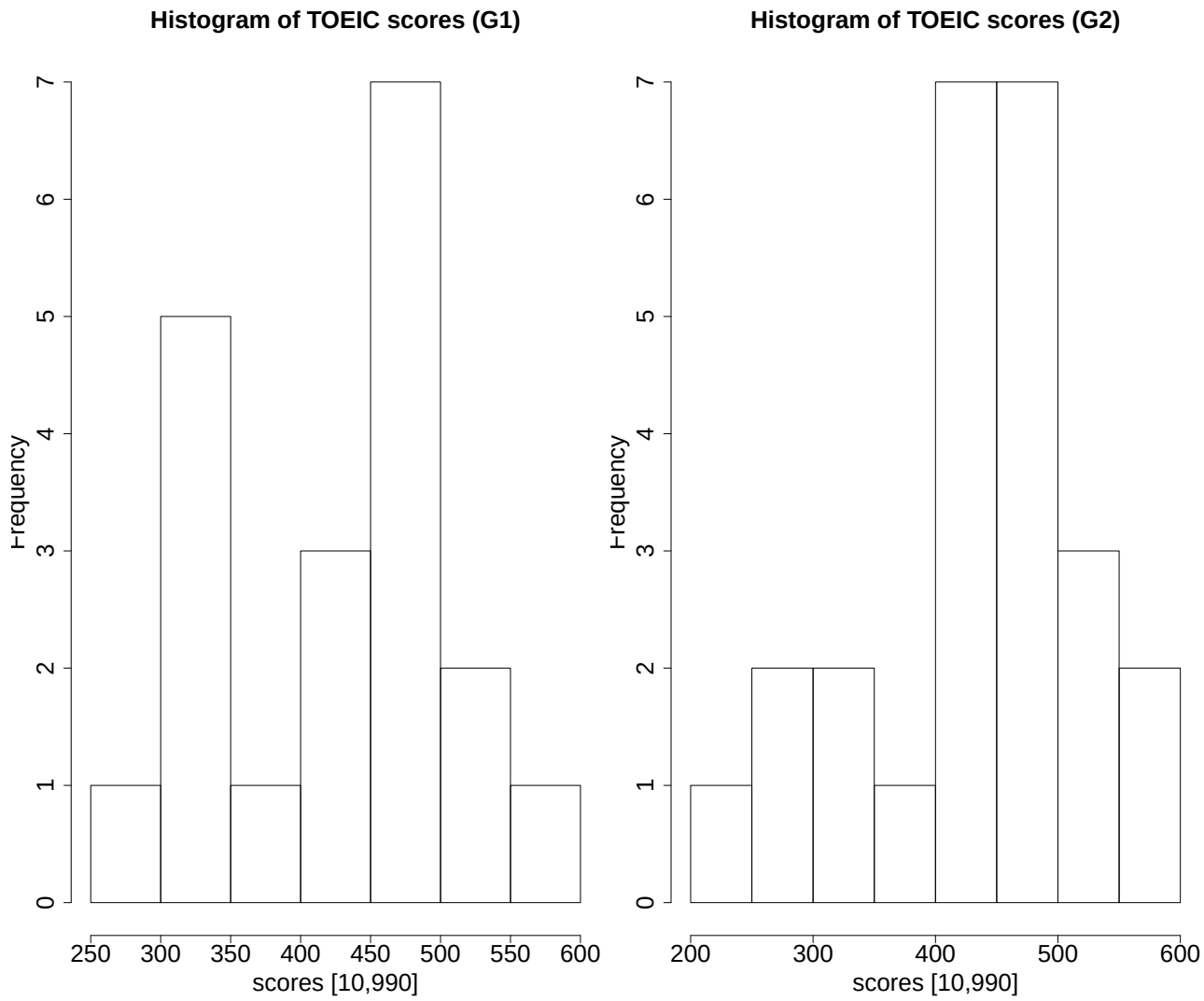

Figure B.3: TOEIC scores (left: G1; right: G2) of the test takers in Evaluation 4.1

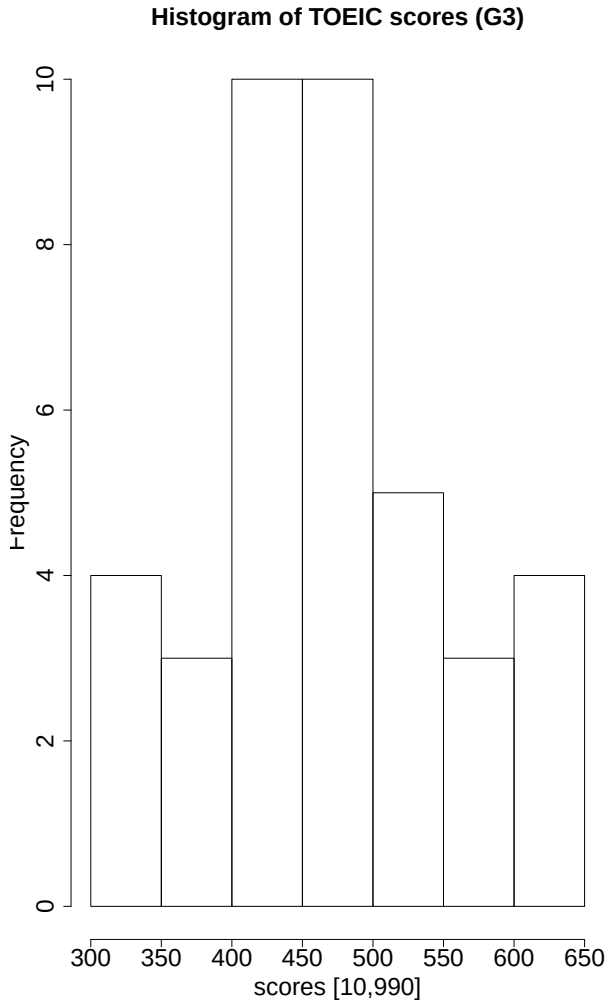

Figure B.4: TOEIC scores (G3) of the test takers in Evaluation 4.1

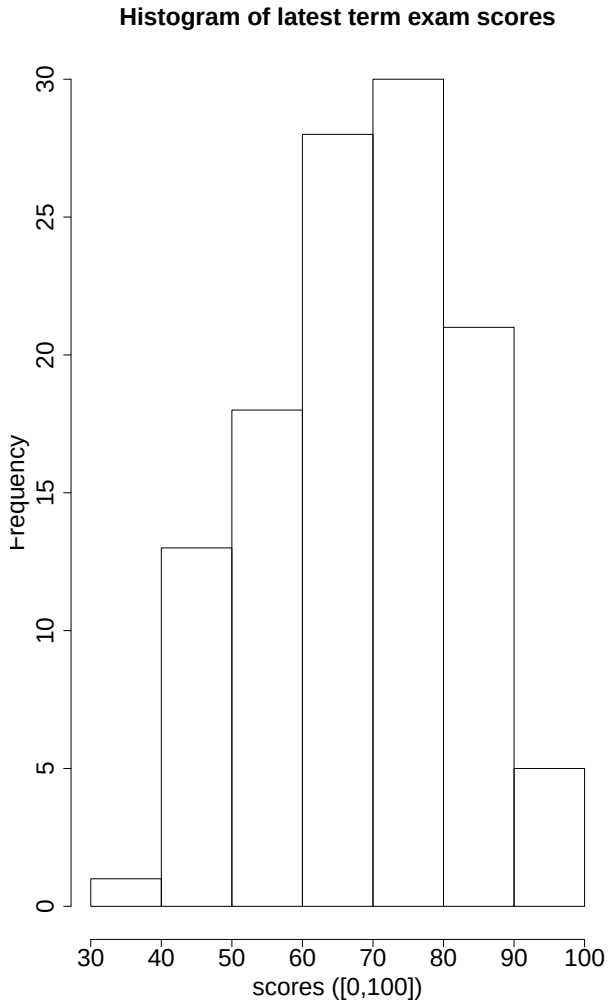

Figure B.5: English test scores of the test takers in Evaluation 5 and Evaluation in chapter 6