Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions

Eiichiro SUMITA
Spoken Language Communication Research Laboratories
ATR
Kyoto 619-0288 Japan
eiichiro.sumita@atr.jp

Fumiaki SUGAYA
Text Information Processing Laboratory
KDDI R&D Laboratories Inc.
Saitama 356-8502 Japan
fsugaya@kddilabs.jp

Seiichi Yamamoto
Department of Information Systems Design
Doshisha University
Kyoto 610-0321 Japan
seyamamo@mail.doshisha.ac.jp
&
Spoken Language Communication Research Laboratories
ATR

Abstract

This paper proposes the *automatic* generation of *Fill-in-the-Blank Questions* (FBQs) together with testing based on *Item Response Theory* (IRT) to measure English proficiency. First, the proposal generates an FBQ from a given sentence in English. The position of a blank in the sentence is determined, and the word at that position is considered as the correct choice. The candidates for incorrect choices for the blank are hypothesized through a thesaurus. Then, each of the candidates is verified by using the Web. Finally, the blanked sentence, the correct choice and the incorrect choices surviving the verification are together laid out to form the FBQ. Second, the proficiency of non-native speakers who took the test consisting of such FBQs is estimated through IRT.

Our experimental results suggest that: (1) the generated questions plus IRT estimate the non-native speakers' English proficiency; (2) while on the other hand, the test can be completed almost perfectly by English native speakers; and (3) the number of questions can be reduced by using *item information* in IRT.

The proposed method provides teachers and testers with a tool that reduces time and expenditure for testing English proficiency.

1 Introduction

English has spread so widely that 1,500 million people, about a quarter of the world's population, speak it, though at most about 400 million speak it as their native language (Crystal, 2003). Thus, English education for non-native speakers both now and in the near future is of great importance.

The progress of computer technology is advancing an electronic tool for language learning called *Computer-Assisted Language Learning* (CALL) and for language testing called *Computer-Based Testing* (CBT) or *Computer-Adaptive Testing* (CAT). However, no computerized support for producing a test, a collection of questions for evaluating *language proficiency*, has emerged to date. [*]

Fill-in-the-Blank Questions (FBQs) are widely used from the classroom level to far larger scales to measure peoples' proficiency at English as a second language. Examples of such tests include TOEFL (Test Of English as a Foreign Language, http://www.ets.org/toefl/) and TOEIC (Test Of English for International Communication, http://www.ets.org/toeic/).

A test comprising FBQs has merits in that (1) it is easy for test-takers to input answers, (2) computers can mark them, thus marking is invariable and objective, and (3) they are suitable for the modern testing theory, *Item Response Theory* (IRT).

Because it is regarded that writing incorrect choices that distract only the non-proficient test-taker is a highly skilled business (Alderson, 1996), FBQs have been written by human experts. Thus, test construction is time-consuming and expensive. As a result, utilizing up-to-date texts for question writing is not practical, nor is tuning in to individual students.

[*] See the detailed discussion in Section 6.

Proceedings of the 2nd Workshop on Building Educational Applications Using NLP,
pages 61–68, Ann Arbor, June 2005. ©Association for Computational Linguistics, 2005

To solve the problems of time and expenditure, this paper proposes a method for generating FBQs using a corpus, a thesaurus, and the Web. Experiments have shown that the proficiency estimated through IRT with generated FBQs highly correlates with non-native speakers' real proficiency. This system not only provides us with a quick and inexpensive testing method, but it also features the following advantages:

(I) It provides "anyone" individually with up-to-date and interesting questions for self-teaching. We have implemented a program that downloads any Web page such as a news site and generates questions from it.

(II) It also enables on-demand testing at "anytime and anyplace." We have implemented a system that operates on a mobile phone. Questions are generated and pooled in the server, and upon a user's request, questions are downloaded. CAT (Wainer, 2000) is then conducted on the phone. The system for mobile phone is scheduled to be deployed in May of 2005 in Japan.

The remainder of this paper is organized as follows. Section 2 introduces a method for making FBQ, Section 3 explains how to estimate test-takers' proficiency, and Section 4 presents the experiments that demonstrate the effectiveness of the proposal. Section 5 provides some discussion, and Section 6 explains the differences between our proposal and related work, followed by concluding remarks.

2 Question Generation Method

We will review an FBQ, and then explain our method for producing it.

2.1 Fill-in-the-Blank Question (FBQ)

FBQs are the one of the most popular types of questions in testing. Figure 1 shows a typical sample consisting of a partially blanked English sentence and four choices for filling the blank. The tester ordinarily assumes that exactly one choice is correct (in this case, b)) and the other three choices are incorrect. The latter are often called *distracters*, because they fulfill a role to distract the less proficient test-takers.

Question 1 (FBQ)
I only have to _______ my head above water one more week.
a) reserve b) keep c) guarantee d) promise

N.B. the correct choice is b) keep.

Figure 1: A sample Fill-in-the-Blank Question (FBQ)

2.2 Flow of generation

Using question 1 above, the outline of generation is presented below (Figure 2).

A *seed* sentence (in this case, "I only have to keep my head above water one more week.") is input from the designated source, e.g., a corpus or a Web page such as well-known news site. [*]

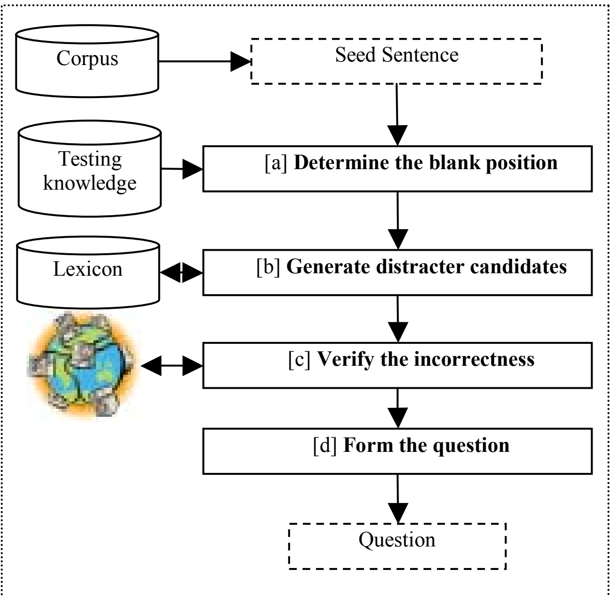

Corpus
Seed Sentence
Testing knowledge
[a] Determine the blank position
Lexicon
[b] Generate distracter candidates
[c] Verify the incorrectness
[d] Form the question
Question

Figure 2: Flow generating *Fill-In-The-Blank Question* (FBQ)

[a] The *seed* sentence is a correct English sentence that is decomposed into a sentence with a blank (*blanked sentence*) and the correct choice for the blank. After the seed

[*] Selection of the seed sentence (source text) is an important open problem because the *difficulty* of the seed (text) should influence the difficulty of the generated question. As for text difficulty, several measures such as Lexile by MetaMetrics (http://www.Lexile.com) have been proposed. They are known as *readability* and are usually defined as a function of sentence length and word frequency.

In this paper, we used corpora of business and travel conversations, because TOEIC itself is oriented toward business and daily conversation.

sentence is analyzed morphologically by a computer, according to the testing knowledge[*] the blank position of the sentence is determined. In this paper's experiment, the *verb* of the seed is selected, and we obtain the *blanked sentence* "I only have to ______ my head above water one more week." and the correct choice "keep."

[b] To be a good distracter, the candidates must maintain the grammatical characteristics of the correct choice, and these should be similar in *meaning*[†]. Using a *thesaurus*[‡], words similar to the correct choice are listed up as candidates, e.g., "clear," "guarantee," "promise," "reserve," and "share" for the above "keep."

[c] Verify (see Section 2.3 for details) *the incorrectness of the sentence* restored by each candidate, and if it is *not incorrect* (in this case, "clear" and "share"), the candidate is given up.

[d] If a sufficient number (in this paper, three) of candidates remain, form a question by randomizing the order of all the choices ("keep," "guarantee," "promise," and "reserve"); otherwise, another seed sentence is input and restart from step [a].

2.3 Incorrectness Verification

In FBQs, by definition, (1) the *blanked sentence* restored with the correct choice is *correct*, and (2) the *blanked sentence* restored with the distracter must be *incorrect*.

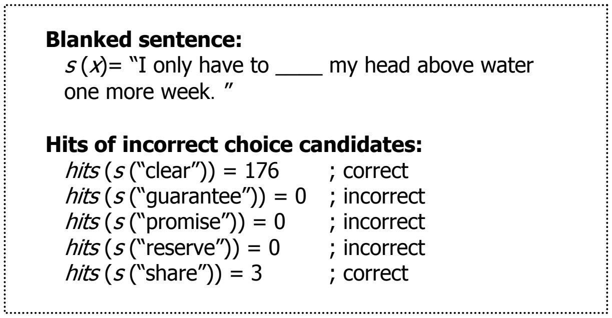
In order to generate an FBQ, the *incorrectness* of the sentence restored by each distracter candidate must be verified and if the combination is *not incorrect*, the candidate is rejected.

Zero-Hit Sentence

The Web includes all manners of language data in vast quantities, which are for everyone easy to access through a networked computer. Recently, exploitation of the Web for various natural language applications is rising (Grefenstette, 1999; Turney, 2001; Kilgarriff and Grefenstette, 2003; Tonoike et al., 2004).

We also propose a Web-based approach. We dare to assume that if there is a sentence on the Web, that sentence is considered *correct*; otherwise, the sentence is unlikely to be *correct* in that there is no sentence written on the Web despite the variety and quantity of data on it.

Figure 3 illustrates verification based on the retrieval from the Web. Here, *s* (*x*) is the blanked sentence, *s* (*w*) denotes the sentence restored by the word *w*, and *hits* (*y*) represents the number of documents retrieved from the Web for the key *y*.



Blanked sentence:
s (*x*)= "I only have to ____ my head above water one more week. "

Hits of incorrect choice candidates:
hits (*s* ("clear")) = 176 ; correct
hits (*s* ("guarantee")) = 0 ; incorrect
hits (*s* ("promise")) = 0 ; incorrect
hits (*s* ("reserve")) = 0 ; incorrect
hits (*s* ("share")) = 3 ; correct

Figure 3: Incorrectness and Hits on the Web

If *hits* (*s* (*w*)), is *small*, then the sentence restored with the word *w* is unlikely, thus the word *w* should be a good distracter. If *hits* (*s* (*w*)), is *large* then the sentence restored with the word *w* is likely, then the word *w* is *unlikely* to be a good distracter and is given up.

[*] Testing knowledge tells us what part of the seed sentence should be blanked. For example, we selected the *verb* of the seed because it is one of the basic types of blanked words in popular FBQs such as in TOEIC.

This can be a word of another *POS* (Part-Of-Speech). For this, we can use knowledge in the field of second-language education. Previous studies on errors in English usage by Japanese native speakers such as (Izumi and Isahara, 2004) unveiled patterns of errors specific to Japanese, e.g., (1) *article* selection error, which results from the fact there are no articles in Japanese; (2) *preposition* selection error, which results from the fact some Japanese counterparts have broader meaning; (3) *adjective* selection error, which results from mismatch of meaning between Japanese words and their counterpart. Such knowledge may generate questions harder for Japanese who study English.

[†] There are various aspects other than *meaning*, for example, *spelling*, *pronunciation*, and *translation* and so on. Depending on the aspect, lexical information sources other than a thesaurus should be consulted.

[‡] We used an in-house English thesaurus whose hierarchy is based on one of the off-the-shelf thesauruses for Japanese, called Ruigo-Shin-Jiten (Ohno and Hamanishi, 1984). In the above examples, the original word "keep" expresses two different concepts: (1) *possession-or-disposal*, which is shared by the words "clear" and "share," and (2) *promise*, which is shared by the words "guarantee," "promise," and "reserve." Since this depends on the thesaurus used, some may sense a slight discomfort at these concepts. If a different thesaurus is used, the distracter candidates may differ.

We used the **strongest** condition. If *hits* (*s* (*w*)) is ***zero***, then the sentence restored with the word *w* is unlikely, thus the word *w* should be a good distracter. If *hits* (*s* (*w*)), is ***not zero***, then the sentence restored with the word *w* is likely, thus the word *w* is *unlikely* to be a good distracter and is given up.

Retrieval NOT By Sentence

It is often the case that *retrieval by sentence* does not work. Instead of a sentence, a sequence of words around a blank position, beginning with a content word (or sentence head) and ending with a content word (or sentence tail) is passed to a search engine automatically. For the abovementioned sample, the sequence of words passed to the engine is "I only have to *clear* my head" and so on.

Web Search

We can use any search engine, though we have been using Google since February 2004. At that point in time, Google covered an enormous four billion pages.

The "correct" hits may come from non-native speakers' websites and contain invalid language usage. To increase reliability, we could restrict Google searches to Websites with URLs based in English-speaking countries, although we have not done so yet. There is another concern: even if sentence fragments cannot be located on the Web, it does not necessarily mean they are illegitimate. Thus, the proposed verification based on the Web is not perfect; the point, however, is that with such limitations, the generated questions are useful for estimating proficiency as demonstrated in a later section.

Setting aside the convenience provided by the off-the-shelf search engine, another search specialized for this application is possible, although the current implementation is fast enough to automate generation of FBQs, and the demand to accelerate the search is not strong. Rather, the problem of time needed for test construction has been reduced by our proposal.

The throughput depends on the text from which a seed sentence comes and the network traffic when the Web is accessed. Empirically, one FBQ is obtained in 20 seconds on average and the total number of FBQs in a day adds up to over 4,000 on a single computer.

3 Estimating Proficiency

3.1 Item Response Theory (IRT)

Item Response Theory (IRT) is the basis of modern language tests such as TOEIC, and enables *Computerized Adaptive Testing* (CAT). Here, we briefly introduce IRT. IRT, in which a question is called an *item*, calculates the test-takers' proficiency based on the answers for items of the given test (Embretson, 2000).

The basic idea is the *item response function*, which relates the probability of test-takers answering particular items correctly to their proficiency. The item response functions are modeled as *logistic curves* making an S-shape, which take the form (1) for item *i*.

$$P_i(\theta) = \frac{1}{1+\exp(-a_i(\theta - b_i))} \quad (1)$$

The *test-taker parameter*, θ, shows the proficiency of the test-taker, with higher values indicating higher performance.

Each of the *item parameters*, a_i and b_i, controls the shape of the item response function. The *a* parameter, called *discrimination*, indexes how steeply the item response function rises. The *b* parameter is called *difficulty*. Difficult items feature larger *b* values and the item response functions are shifted to the right. These item parameters are usually estimated by a maximal likelihood method. For computations including the estimation, there are many commercial programs such as BILOG (http://www.assess.com/) available.

3.2 Reducing test size by selection of effective items

It is important to estimate the proficiency of the test-taker by using as few items as possible. For this, we have proposed a method based on *item information*.

Expression (2) is the *item information* of item *i* at θ_j, the proficiency of the test-taker *j*, which indicates how much *measurement discrimination* an item provides.

The procedure is as follows.

1. Initialize *I* by the set of all generated FBQs.

2. According to Equation (3), we select the item whose contribution to *test information* is maximal.
3. We eliminate the selected item from *I* according to Equation (4).
4. If *I* is empty, we obtain the ordered list of effective items; otherwise, go back to step 2.

$$I_i(\theta_j) = a_i^{\,2} P_i(\theta_j)(1 - P_i(\theta_j)) \quad (2)$$

$$\hat{i} = \arg\max_i \left(\sum_j \sum_{i \in I} I_i(\theta_j) \right) \quad (3)$$

$$I = I - \hat{i} \quad (4)$$

4 Experiment

The FBQs for the experiment were generated in February of 2004. Seed sentences were obtained from ATR's corpus (Kikui *et al*., 2003) of the *business* and *travel* domains. The vocabulary of the corpus comprises about 30,000 words. Sentences are relatively short, with the average length being 6.47 words. For each domain 5,000 questions were generated automatically and each question consists of an English sentence with *one* blank and *four* choices.

4.1 Experiment with non-native speakers

We used the TOEIC score as the experiment's proficiency measure, and collected 100 Japanese subjects whose TOEIC scores were scattered from 400 to less than 900. The actual range for TOEIC scores is 10 to 990. Our subjects covered the dominant portion* of test-takers for TOEIC in Japan, excluding the highest and lowest extremes.†

We had the subjects answer 320 randomly selected questions from the 10,000 mentioned above. The raw marks were as follows: the average‡ mark was 235.2 (73.5%); the highest mark was 290 (90.6%); and the lowest was 158 (49.4%). This suggests that our FBQs are sensitive to test-takers' proficiency. In Figure 4, the y-axis represents estimated proficiency according to IRT (Section 3.1) and generated questions, while the x-axis is the real TOEIC score of each subject.

As the graph illustrates, the IRT-estimated proficiency (*θ*) and real TOEIC scores of subjects correlate highly with a co-efficiency of about 80%.

For comparison we refer to CASEC (http://casec.evidus.com/), an off-the-shelf test consisting of human-made questions and IRT. Its co-efficiency with real TOEIC scores is reported to be 86%.

This means the proposed automatically generated questions are promising for measuring English proficiency, achieving a nearly competitive level with human-made questions but with a few reservations: (1) whether the difference of 6% is large depends on the standpoint of possible users; (2) as for the number of questions to be answered, our proposal uses 320 questions in the experiments, while TOEIC uses 200 questions and CASEC uses only about 60 questions; (3) the proposed method uses FBQs only whereas CASEC and TOEIC use various types of questions.

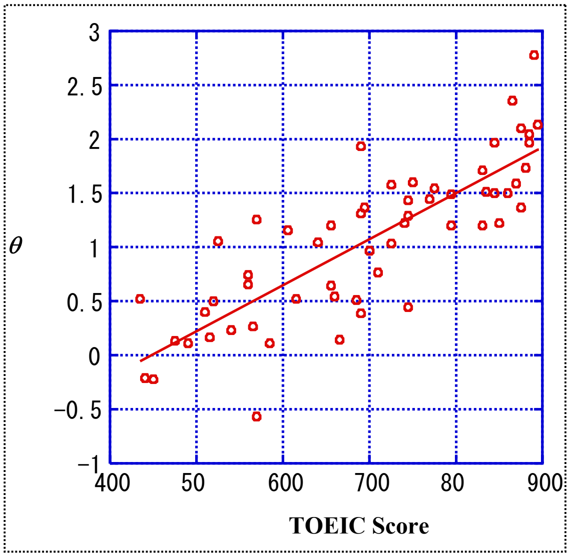


Figure 4: IRT-Estimated Proficiency (θ) vs. Real TOEIC Score

4.2 Experiment with a native speaker

To examine the quality of the generated questions, we asked a single subject§ who is a native speaker of English to answer 4,000 questions (Table 1).

The native speaker largely agreed with our generation, determining correct choices (type I). The

* Over 70% of all test-takers are covered (http://www.toeic.or.jp/toeic/data/data02.html).

† We have covered only the range of TOEIC scores from 400 to 900 due to expense of the experiment. In this restricted experiment, we do not claim that our proficiency estimation method covers the full range of TOEIC scores.

‡ The standard deviation was 29.8 (9.3%).

§ Please note that the analysis is based on a single native-speaker, thus we need further analysis by multiple subjects.

rate was 93.50%, better than 90.6%, the highest mark among the non-native speakers.

We present the problematic cases here.

- Type II is caused by the seed sentence being *incorrect* for the native speaker, and a distracter is bad because it is *correct*. Or like type III, it consists of ambiguous choices.
- Type III is caused by some generated distracters being *correct*; therefore, the choices are ambiguous.
- Type IV is caused by the seed sentence being *incorrect* and the generated distracters also being *incorrect*; therefore, the question cannot be answered.
- Type V is caused by the seed sentence being nonsense to the native speaker; the question, therefore, cannot be answered.

Table 1 Responses of a Native speaker

Type	***Explanation***		***Count***	***%***
I	**Single Selection**	**Match**	**3,740**	**93.50**
II		No match	55	1.38
III	No Selection	Ambiguous Choices	70	1.75
IV		No Correct Choice	45	1.13
V		Nonsense	90	2.25

Cases with bad seed sentences (portions of II, IV, and V) require cleaning of the corpus by a native speaker, and cases with bad distracters (portions of II and III) require refinement of the proposed generation algorithm.

Since the questions produced by this method can be flawed in ways which make them unanswerable even by native speakers (about 6.5% of the time) due to the above-mentioned reasons, it is difficult to use this method for high-stakes testing applications although it is useful for estimating proficiency as explained in the previous section.

4.3 *Proficiency θ* estimated with the reduced test and its relation to TOEIC Scores

Figure 5 shows the relationship between reduction of the test size according to the method explained in Section 3.2 and the estimated proficiency based on the reduced test. The x-axis represents the size of the reduced test in number of items, while the y-axis represents the correlation coefficient (R) between estimated proficiency and real TOEIC score.

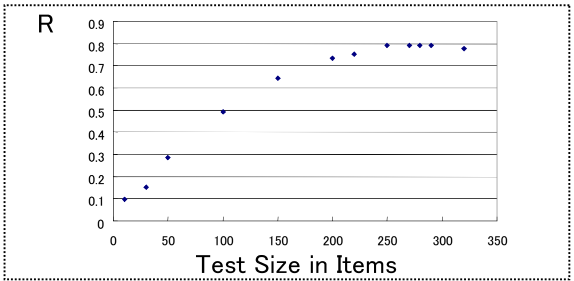


Figure 5 Correlation coefficient and Test size

5 Discussion

This section explains the on-demand generation of FBQs according to individual preference, an immediate extension and a limitation of our proposed method, and finally touches on free-format Q&A.

5.1 Effects of Automatic FBQ Construction

The method provides teachers and testers with a tool that reduces time and expenditure. Furthermore, the method can deal with any text. For example, up-to-date and interesting materials such as news articles of the day can be a source of seed sentences (Figure 6 is a sample generated from an article (http://www.japantimes.co.jp/) on an earthquake that occurred in Japan), which enables realization of a *personalized* learning environment.

Question 2 (FBQ)
The second quake ______ 10 km below the seabed some 130 km east of Cape Shiono.

a) put b) came c) originated d) opened

N.B. The correct answer is c) originated.

Figure 6: On-demand construction – a sample question from a Web news article in *The Japan Times* on “an earthquake”

We have generated questions from over 100 documents on various genres such as novels, speeches, academic papers and so on found in the enormous collection of e-Books provided by Project Gutenberg (http://www.gutenberg.org/).

5.2 A Variation of Fill-in-the-Blank Questions for Grammar Checking

In Section 2.2, we mentioned a constraint that a good distracter should maintain the grammatical characteristics of the correct choice originating in the seed sentence. The question checks not the grammaticality but the semantic/pragmatic correctness.

We can generate another type of FBQ by slightly modifying step [b] of the procedure in Section 2.2 to retain the stem of the original word *w* and vary the surface form of the word *w*. This modified procedure generates a question that checks the grammatical ability of the test takers. Figure 7 shows a sample of this kind of question taken from a TOEIC-test textbook (Educational Testing Service, 2002).

Question 3 (FBQ)

Because the equipment is very delicate, it must be handled with ______.

a) caring b) careful c) care d) carefully

N.B. The correct answer is c) care.

Figure 7: A variation on fill-in-the-blank questions

5.3 Limitation of the Addressed FBQs

The questions dealt with in this paper concern testing reading ability, but these questions are not suitable for testing listening ability because they are presented visually and cannot be pronounced. To test listening ability, like in TOIEC, other types of questions should be used, and automated generation of them is yet to be developed.

5.4 Free-Format Q&A

Besides measuring one's ability to receive information in a foreign language, which has been addressed so far in this paper, it is important to measure a person's ability to transmit information in a foreign language. For that purpose, tests for translating, writing, or speaking in a free format have been actively studied by many researchers (Shermis, 2003; Yasuda, 2004).

6 Related Work[*]

Here, we explain other studies on the generation of multiple-choice questions for language learning. There are a few previous studies on computer-based generation such as Mitkov (2003) and Wilson (1997).

6.1 Cloze Test

A computer can generate questions by deleting words or parts of words randomly or at every *N*-th word from text. Test-takers are requested to restore the word that has been deleted. This is called a "cloze test." The effectiveness of a "cloze test" or its derivatives is a matter of controversy among researchers of language testing such as Brown (1993) and Alderson (1996).

6.2 Tests on Facts

Mitkov (2003) proposed a computer-aided procedure for generating *multiple-choice questions* from textbooks. The differences from our proposal are that (1) Mitkov's method generates questions not about *language usage* but about *facts explicitly stated in a text*[†]; (2) Mitkov uses techniques such as term extraction, parsing, transformation of trees, which are different from our proposal; and (3) Mitkov does not use IRT while we use it.

7 Conclusion

This paper proposed the automatic construction of *Fill-in-the-Blank Questions* (FBQs). The proposed method generates FBQs using a corpus, a thesaurus, and the Web. The generated questions and *Item Response Theory* (IRT) then estimate second-language proficiency.

Experiments have shown that the proposed method is effective in that the estimated proficiency highly correlates with non-native speakers' real proficiency as represented by TOEIC scores; native-speakers can achieve higher scores than non-native speakers. It is possible to reduce the size of the test by removing non-discriminative questions with *item information* in IRT.

[*] There are many works on item generation theory (ITG) such as Irvine and Kyllonen (2002), although we do not go any further into the area. We focus only on multiple-choice questions for language learning in this paper.

[†] Based on a fact stated in a textbook like, "A prepositional phrase at the beginning of a sentence constitutes an *introductory modifier*," Mitkov generates a question such as, "*What* does a prepositional phrase at the beginning of a sentence constitute? i. *a modifier that accompanies a noun*; ii. *an associated modifier*; iii. *an introductory modifier*; iv. *a misplaced modifier*."

The method provides teachers, testers, and test takers with novel merits that enable low-cost testing of second-language proficiency and provides learners with up-to-date and interesting materials suitable for individuals.

Further research should be done on (1) large-scale evaluation of the proposal, (2) application to different languages such as Chinese and Korean, and (3) generation of different types of questions.

Acknowledgements

The authors' heartfelt thanks go to anonymous reviewers for providing valuable suggestions and *Kadokawa-Shoten* for providing the thesaurus named *Ruigo-Shin-Jiten*. The research reported here was supported in part by a contract with the NiCT entitled, "A study of speech dialogue translation technology based on a large corpus." It was also supported in part by the Grants-in-Aid for Scientific Research (KAKENHI), contract with MEXT numbered 16300048. The study was conducted in part as a cooperative research project by KDDI and ATR.

References

Alderson, Charles. 1996. *Do corpora have a role in language assessment?* Using Corpora for Language Research, eds. Thomas, J. and Short, M., Longman: 248—259.

Brown, J. D. 1993. *What are the characteristics of natural cloze tests?* Language Testing 10: 93—116.

Crystal, David. 2003. *English as a Global Language, (Second Edition).* Cambridge University Press: 212.

Educational Testing Service 2002. *TOEIC koushiki gaido & mondaishu.* IIBC: 249.

Embretson, Susan et al. 2000. *Item Response Theory for Psychologists.* LEA: 371.

Grefenstette, G. 1999. *The WWW as a resource for example-based MT tasks.* ASLIB "Translating and the Computer" conference.

Irvine, H. S., and Kyllonen, P. C. (2002). *Item generation for test development*. LEA: 412.

Izumi, E., and Isahara, H. (2004). *Investigation into language learners' acquisition order based on the error analysis of the learner corpus*. In Proceedings of Pacific-Asia Conference on Language, Information and Computation (PACLIC) 18 Satellite Workshop on E-Learning, Japan. (in printing)

Kikui, G., Sumita, E., Takaezawa, T. and Yamamoto, S., "Creating Corpora for Speech-to-Speech Translation," Special Session "Multilingual Speech-to-Speech Translation" of EuroSpeech, 2003.

Kilgarriff, A. and Grefenstette, G. 2003. *Special Issue on the WEB as Corpus.* Computational Linguistics 29 (3): 333—502.

Mitkov, Ruslan and Ha, Le An. 2003. *Computer-Aided Generation of Multiple-Choice Tests.* HLT-NAACL 2003 Workshop: Building Educational Applications Using Natural Language Processing: 17—22.

Ohno, S. and Hamanishi, M. 1984. *Ruigo-Shin-Jiten*, Kadokawa, Tokyo (in Japanese)

Shermis, M. D. and Burstein. J. C. 2003. *Automated Essay Scoring.* LEA: 238.

Tonoike, M., Sato, S., and Utsuro, T. 2004. *Answer Validation by Keyword Association.* IPSJ, SIGNL, 161: 53—60, (in Japanese).

Turney, P.D. 2001. *Mining the Web for synonyms: PMI-IR vs. LSA on TOEFL.* ECML 2001: 491—502.

Wainer, Howard et al. 2000. *Conputerized Adaptive Testing: A Primer, (Second Edition).* LEA: 335.

Wilson, E. 1997. *The Automatic Generation of CALL exercises from general corpora*, in eds. Wichmann, A., Fligelstone, S., McEnery, T., Knowles, G., Teaching and Language Corpora, Harlow: Longman:116-130.

Yasuda, K., Sugaya, F., Sumita, E., Takezawa, T., Kikui, G. and Yamamoto, S. 2004. *Automatic Measuring of English Language Proficiency using MT Evaluation Technology*, COLING 2004 eLearning for Computational Linguistics and Computational Linguistics for eLearning: 53-60.