Automatic Measuring of English Language Proficiency using MT Evaluation Technology

Keiji Yasuda
ATR Spoken Language Translation Research Laboratories
Department of SLR
2-2-2 Hikaridai,
"Keihanna Science City"
Kyoto 619-0288 Japan
keiji.yasuda@atr.jp

Fumiaki Sugaya
KDDI R&D Laboratories
2-1-15, Ohara, Kamifukuoka-city,
Saitama, 356-8502, Japan
fsugaya@kddilabs.jp

Eiichiro Sumita
ATR Spoken Language Translation Research Laboratories
Department of NLR
2-2-2 Hikaridai,
"Keihanna Science City"
Kyoto 619-0288 Japan
eiichiro.sumita@atr.jp

Toshiyuki Takezawa
ATR Spoken Language Translation Research Laboratories
Department of SLR
2-2-2 Hikaridai,
"Keihanna Science City"
Kyoto 619-0288 Japan
toshiyuki.takezawa@atr.jp

Genichiro Kikui
ATR Spoken Language Translation Research Laboratories
Department of SLR
2-2-2 Hikaridai,
"Keihanna Science City"
Kyoto 619-0288 Japan
genichiro.kikui@atr.jp

Seiichi Yamamoto
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai,
"Keihanna Science City"
Kyoto 619-0288 Japan
seiichi.yamamoto@atr.jp

Abstract

Assisting in foreign language learning is one of the major areas in which natural language processing technology can contribute. This paper proposes a computerized method of measuring communicative skill in English as a foreign language. The proposed method consists of two parts. The first part involves a test sentence selection part to achieve precise measurement with a small test set. The second part is the actual measurement, which has three steps. Step one asks proficiency-known human subjects to translate Japanese sentences into English. Step two gauges the match between the translations of the subjects and correct translations based on the n-gram overlap or the edit distance between translations. Step three learns the relationship between proficiency and match. By regression it finds a straight-line fitting for the scatter plot representing the proficiency and matches of the subjects. Then, it estimates proficiency of proficiency-unknown users by using the line and the match. Based on this approach, we conducted experiments on estimating the Test of English for International Communication (TOEIC) score. We collected two sets of data consisting of English sentences translated from Japanese. The first set consists of 330 sentences, each translated to English by 29 subjects with varied English proficiency. The second set consists of 510 sentences translated in a similar manner by a separate group of 18 subjects. We found that the estimated scores correlated with the actual scores.

1 Introduction

For effective second language learning, it is absolutely necessary to test proficiency in the second language. This testing can help in selecting educational materials before learning, checking learners' understanding after learning, and so on.

To make learning efficient, it is important to achieve testing with a short turnaround time. Computer-based testing is one solution for this,

and several kinds of tests have been developed, including CASEC (CASEC, 2004) and TOEFL-CBT (TOEFL, 2004). However, these tests are mainly based on cloze testing or multiple-choice questions. Consequently, they require labour costs for expert examination designers to make the questions and the alternative "detractor" answers.

In this paper, we propose a method for the automatic measurement of English language proficiency by applying automatic evaluation techniques. The proposed method selects adequate test sentences from an existing corpus. Then, it automatically evaluates the translations of test sentences done by users. The core technology of the proposed method, i.e., the automatic evaluation of translations, was developed in research aiming at the efficient development of Machine Translation (MT) technology (Su et al., 1992; Papineni et al., 2002; NIST, 2002). In the proposed method, we apply these MT evaluation technologies to the measurement of human English language proficiency. The proposed method focuses on measuring the communicative skill of structuring sentences, which is indispensable for writing and speaking. It does not measure elementary capabilities including vocabulary or grammar. This method also proposes a test sentence selection scheme to enable efficient testing.

Section 2 describes several automatic evaluation methods applied to the proposed method. Section 3 introduces the proposed evaluation scheme. Section 4 shows the evaluation results obtained by the proposed method. Section 5 concludes the paper.

2 MT Evaluation Technologies

In this section, we briefly describe automatic evaluation methods of translation. These methods were proposed to evaluate MT output, but they are applicable to translation by humans.

All of these methods are based on the same idea, that is, to compare the target translation for evaluation with high-quality reference translations that are usually done by skilled translators. Therefore, these methods require a corpus of high-quality human reference translations. We call these translations as "references".

2.1 DP-based Method

The DP score between a translation output and references can be calculated by DP matching (Su et al., 1992; Takezawa et al., 1999). First, we define the DP score between sentence (i.e., word array) W_a and sentence W_b by the following formula.

$$S_{DP}(W_a, W_b) = \frac{T - S - I - D}{T} \qquad (1)$$

where T is the total number of words in W_a, S is the number of substitution words for comparing W_a to W_b, I is the number of inserted words for comparing W_a to W_b, and D is the number of deleted words for comparing W_a to W_b.

Using Equation 1, $(S_i(j))$, that is, the test sentence unit DP-score of the translation of test sentence j done by subject i, can be calculated by the following formula.

$$S_{DP_i}(j) = \max_{k=1 \text{ to } N_{ref}} \left\{ S_{DP}(W_{ref(k)}(j), W_{sub(i)}(j)), 0 \right\} \qquad (2)$$

where N_{ref} is the number of references, $W_{ref(k)}(j)$ is the k-th reference of the test sentence j, and $W_{sub(i)}(j)$ is the translation of the test sentence j done by subject i.

Finally, S_{DP_i}, which is the test set unit DP-score of subject i, can be calculated by the following formula.

$$S_{DP_i} = \frac{1}{N_{sent}} \sum_{j=1}^{N_{sent}} S_{DP_i}(j) \qquad (3)$$

where N_{sent} is the number of test sentences.

2.2 N-gram-based Method

Papineni et al. (2002) proposed BLEU, which is an automatic method for evaluating MT quality using N-gram matching. The National Institute of Standards and Technology also proposed an automatic evaluation method called NIST (2002), which is a modified method of BLEU.

In this research we use two kinds of units to apply BLEU and NIST. One is a test sentence unit and the other is a test set unit. The unit of utterance corresponds to the unit of "segment" in the original BLEU and NIST studies (Papineni et al., 2002; NIST, 2002).

Equation 4 is the test sentence unit BLEU score formulation of the translation of test sentence j done by subject i.

$$S_{BLEU_i}(j) = \exp\left\{ \sum_{n=1}^{N} w_n \log(p_n) - \max\left(\frac{L^*_{ref}}{L_{sys}} - 1, 0 \right) \right\}$$

(4)

where

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in \{C\}} Count_{clip}\,(n-gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in \{C\}} Count(n-gram)}$$

$w_n = N^{-1}$
and

L^*_{ref} = the number of words in the reference translation that is closest in length to the translation being scored

L_{sys} = the number of words in the translation being scored

Equation 5 is the test sentence unit NIST score formulation of the translation of test sentence j done by subject i.

$$S_{NIST_i}(j) = \sum_{n=1}^{N} \left\{ \frac{\sum_{\text{all } w_1 \ldots w_n \text{ in sys output}} info(w_1 \ldots w_n)}{\sum_{\text{all } w_1 \ldots w_n \text{ in sys output}}(1)} \right\} \times \exp\left\{ \beta \log^2 \left[\min\left(\frac{L_{sys}}{\overline{L}_{ref}},\ 1 \right) \right] \right\} \quad (5)$$

where

$$info(w_1 \ldots w_n) = \log_2 \left(\frac{\text{the number of occurence of } w_1 \ldots w_{n-1}}{\text{the number of occurence of } w_1 \ldots w_n} \right)$$

$\overline{L}_{ref}$ = the average number of words in a reference translation, averaged over all reference translations

L_{sys} = the number of words in the translation being scored

and β is chosen to make the brevity penalty factor=0.5 when the number of words in the system translation is 2/3 of the average number of words in the reference translation. For Equations 4 and 7, N indicates the maximum n-gram length. In this research we set N to 4 for BLEU and to 5 for NIST.

We may consider the unit of the test set corresponding to the unit of "document" or "system" in BLEU and NIST. However, we use formulations for the test set unit scores that are different from those of the original BLEU and NIST.

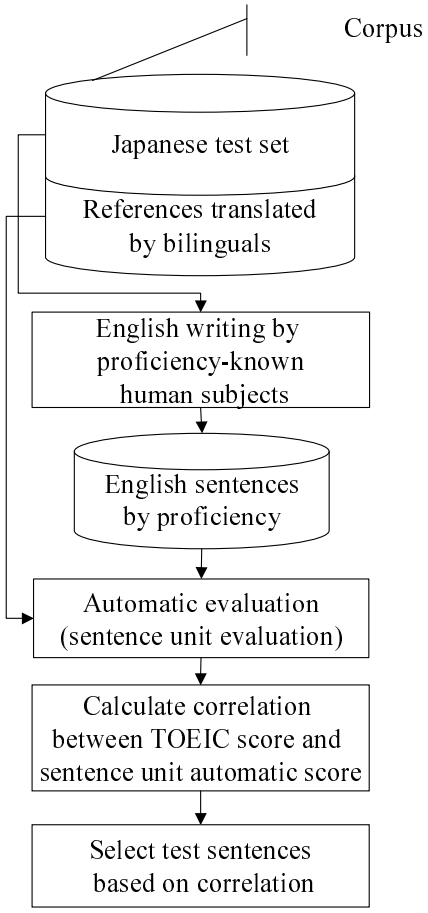

Figure 1: Flow of Test Set Selection

The test set unit scores of BLEU and NIST are calculated by Equations 6 and 7.

$$S_{BLEU_i} = \frac{1}{N_{sent}} \sum_{j=1}^{N_{sent}} S_{BLEU_i}(j) \quad (6)$$

$$S_{NIST_i} = \frac{1}{N_{sent}} \sum_{j=1}^{N_{sent}} S_{NIST_i}(j) \quad (7)$$

3 The Proposed Method

The proposed method described in this paper consists of two parts. One is the test set selection part and the other is the actual measurement part. The measurement part is divided into two phases: a parameter-estimation phase and a testing phase. Here, we use the term "subjects" to refer to the human subjects in the test set selection part and the parameter-estimation phase of the measurement part; we use "users" to refer to the humans in the testing phase of the measurement part.

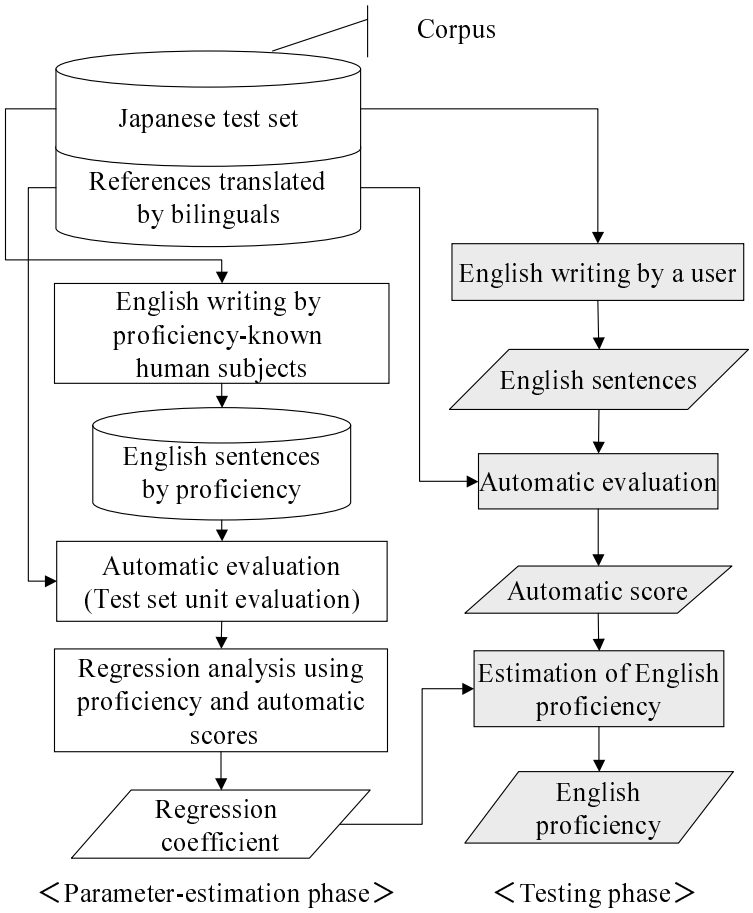

Figure 2: Flow of English Proficiency Measurment

We employ the Test of English for International Communication (TOEIC, 2004) as an objective measure of English proficiency.

3.1 Test Sentence Selection Method

Figure 1 shows the flow of the test sentence selection. We first calculate the test sentence unit automatic score by using Equation 2, 4 or 5 for each test sentence and subject. Second, for each test sentence, we calculate the correlation between the automatic scores and subjects' TOEIC scores. Finally, using the above results, we choose the test sentences that give high correlation.

3.2 Method of Measuring English Proficiency

Figure 2 shows the flow of measuring English proficiency. In the parameter-estimation phase, for each subject, we first calculate the test set unit automatic score by using Equation 3, 6 or 7. Next, we apply regression analysis using the automatic scores and subjects' TOEIC scores.

In the testing phase, we calculate a user's TOEIC score using the automatic score of the user and the regression line calculated in the parameter-estimation phase.

4 Experiments

4.1 Experimental Conditions

4.1.1 Test sets

For the experiments, we employ two different test sets. One is BTEC (Basic Travel Expression Corpus) (Takezawa et al., 2002) and the other is SLTA1 (Takezawa, 1999). Both BTEC and SLTA1 are parts of bilingual corpora that have been collected for research on speech translation systems. However, they have different features. A detailed analysis of these corpora was done by Kikui et al. (2003). Here, we briefly explain these test sets. In this study, we use the Japanese side as a test set and the English side as a reference for automatic evaluation.

BTEC

BTEC was designed to cover expressions for every potential subject in travel conversation. This test set was collected by investigating

"phrasebooks" that contain Japanese/English sentence pairs that experts consider useful for tourists traveling abroad. One sentence contains 8 words on average. The test set for this experiment consists of 510 sentences from the BTEC corpus.

The total number of examinees is 18, and the range of their TOEIC scores is between the 400s and 900s. Every hundred-point range has 3 examinees.

SLTA1

SLTA1 consists of 330 sentences in 23 conversations from the ATR bilingual travel conversation database (Takezawa, 1999). One sentence contains 13 words on average. This corpus was collected by simulated dialogues between Japanese and English speakers through a professional interpreter. The topics of the conversations are mainly hotel conversations, such as reservations, enquiries and so on.

The total number of examinees is 29, and the range of their TOEIC score is between the 300s and 800s. Excluding the 600s, every hundred-point range has 5 examinees.

4.1.2 Reference

For the automatic evaluation, we collected 16 references for each test sentence. One of them is from the English side of the test set, and the remaining 15 were translated by 5 bilinguals (3 references by 1 bilingual).

4.2 Experimental Results

4.2.1 Experimental Results of Test Set Selection

Figures 3 and 4 show the correlation between the test sentence unit automatic score and the subjects' TOEIC score. Here, the automatic score is calculated using Equation 2, 4 or 5. Figure 3 shows the results on BTEC, and Fig. 4 shows the results on SLTA1. In these figures, the ordinate represents the correlation. The filled circles indicate the results using the DP-based automatic evaluation method. The gray circles indicate the results using BLEU. The empty circles indicate the results using NIST. Looking at these figures, we find that the three automatic evaluation methods show a similar tendency. Comparing BTEC and SLTA1, BTEC contains more cumbersome test sentences. In BTEC, about 20% of the test sentences give a correlation of less than 0. Meanwhile, in the SLTA1, this percentage is about 10%.

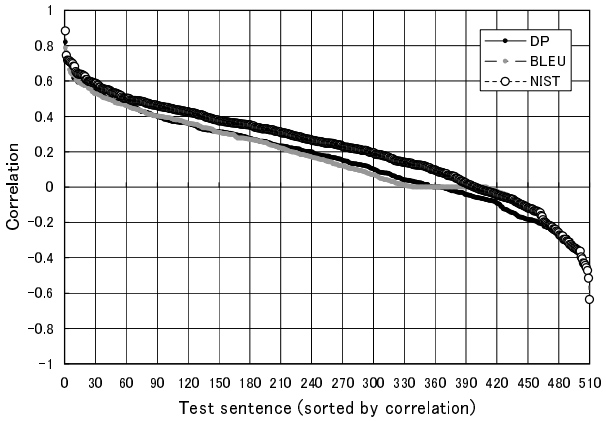

Figure 3: Correlation between test sentence unit automatic scores and subjects' TOEIC scores (BTEC)

Table 1 shows examples of low-correlated test sentences. As shown in the table, BTEC contains more short and frequently used expressions than does SLTA1. This kind of expression is thought to be too easy for testing, so this low-correlation phenomenon is thought to occur. SLTA1 still contains a few sentences of this kind ("Example 1" of SLTA1 in the table). Additionally, there is another contributing factor explaining the low correlation in SLTA1. Looking at "Example 2" of SLTA1 in the table, this expression is not very easy to translate. For this test sentence, several expressions can be produced as an English translation. Thus, automatic evaluation methods cannot evaluate correctly due to the insufficient variety of references. Considering these results, this method can remove inadequate test sentences due not only to the easiness of the test sentence but also to the difficulty of the automatic evaluation. Figures 5 and 6 show the relationship between the number of test sentences and correlation. This correlation is calculated between the test set unit automatic scores and the subjects' TOEIC scores. Here, the automatic score is calculated using Equation 3, 6 or 7. Figure 5 shows the results on BTEC, and Fig. 6 shows the results on SLTA1.

In these figures, the abscissa represents the number of test sentences, i.e., N_{sent} in Equations 3, 6 and 7, and the ordinate represents the correlation. Definitions of the circles are the same as those in the previous figure. Here, the test sentence selection is based on the correlation shown in Figs. 3 and 4.

Comparing Fig. 5 to Fig. 6, in the case of

Table 1: Example of low-correlated test sentences

		Japanese	English
BTEC	Example 1	おやすみなさい	Good night.
	Example 2	メニューを見せてください	Can I see a menu, please?
SLTA1	Example 1	はい、マスターカードでお願いします。	Yes, with my Mastercard please
	Example 2	それでお願いしたいんですが、予算の都合もありますんでとりあえずスイートの料金を教えてください。	I wish I could take that but we have a limited budget so how much will that cost?

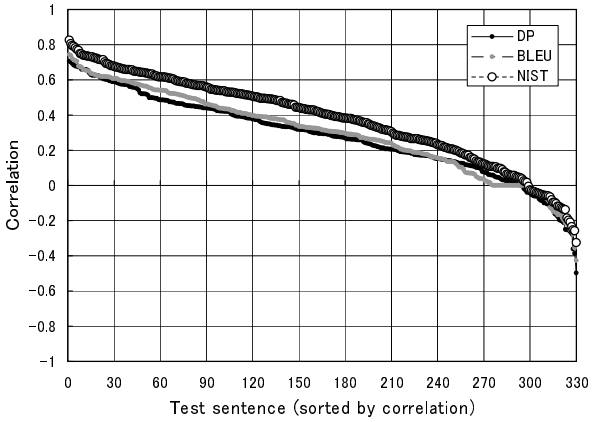

Figure 4: Correlation between test sentence unit automatic scores and subjects' TOEIC scores (SLTA1)

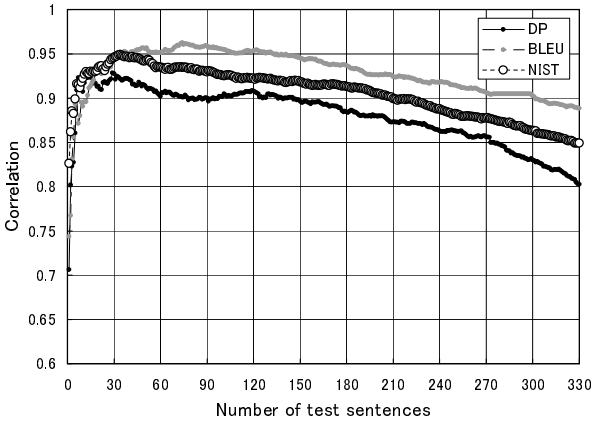

Figure 6: Correlation between test set unit automatic scores and subjects' TOEIC scores (SLTA1)

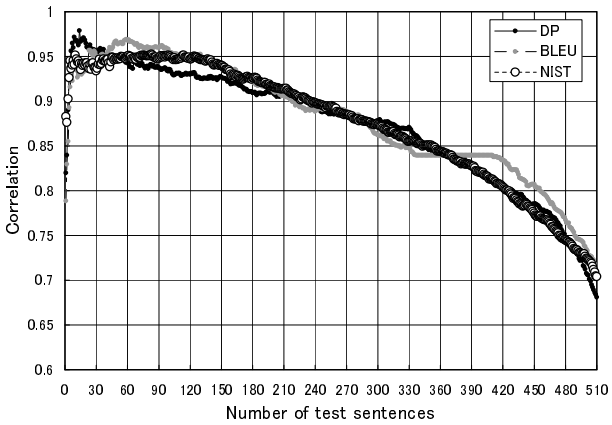

Figure 5: Correlation between test set unit automatic scores and subjects' TOEIC scores (BTEC)

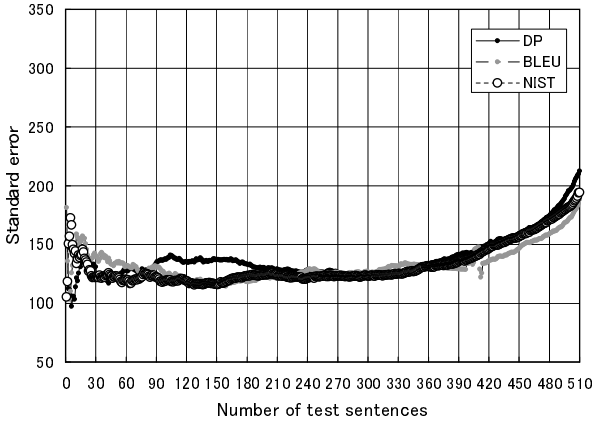

Figure 7: Standard error (BTEC)

using the full test set (510 test sentences for BTEC, 330 test sentences for SLTA1), the correlation of BTEC is lower than that of SLTA1. As we mentioned above, the ratio of the low-correlated test sentences in BTEC is higher than that of SLTA1 (See Figs. 3 and 4). This issue is thought to cause a decrease in the correlation shown in Fig. 5. However, by applying the selection based on sentence unit correlation, these obstructive test sentences can be removed. This permits the selection of high-correlated small-sized test sets. In these figures, the highest correlations are around 0.95.

4.2.2 Experimental Results of English Proficiency Measurement

For the experiments on English proficiency measurement, we carried out a leave-one-out cross validation test. The leave-one-out cross valida-

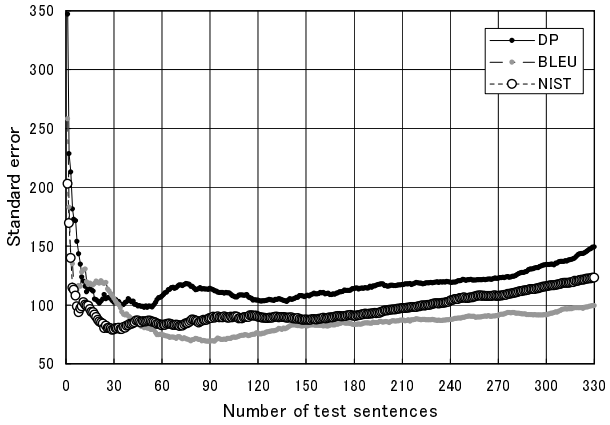

Figure 8: Standard error (SLTA1)

tion test is conducted not only for the measurement of the English proficiency but also for the test set selection.

To evaluate the proficiency measurement by the proposed method, we calculate the standard error of the results of a leave-one-out cross validation test. The following formula is the definition of the standard error.

$$\sigma_E = \sqrt{\frac{1}{N_{user}} \sum_{i=1}^{N_{user}} (T_i - A_i)^2} \qquad (8)$$

where N_{user} is the number of users, T_i is the actual TOEIC score of user i, and A_i is user i's estimated TOEIC score by using the proposed method.

Figures 7 and 8 show the relationship between the number of test sentences and the standard error.

In these figures, the abscissa represents the number of test sentences, and the ordinate represents the standard error. Definitions of the circles are the same as in the previous figure. Here, the test sentence selection is based on the correlation shown in Figs. 3 and 4.

Looking at Figs. 7 and 8, we can observe differences between the standard errors of BTEC and SLTA1. This is thought to be due to the difference of the number of subjects in the experiments (for the leave-one-out cross validation test, 17 subjects with BTEC and 28 subjects with SLTA1). Even though these were closed experiments, the results in Figs. 5 and 6 show an even higher correlation with BTEC than with SLTA1 at the highest point. Therefore, there is room for improvement by increasing the number of subjects with BTEC.

In the test using 30 to 60 test sentences in Figs. 7 and 8, the standard errors are much smaller than in the test using the full test set (510 test sentences for BTEC, 330 test sentences for SLTA1). These results imply that the test set selection works very well and that it enables precise testing using a smaller size test set.

5 Conclusion

We proposed an automatic measurement method for English language proficiency. The proposed method applies automatic MT evaluation to measure human English language proficiency. This method focuses on measuring the communicative skill of structuring sentences, which is indispensable in writing and speaking. However, it does not measure elementary capabilities such as vocabulary and grammar. The method also involves a new test sentence selection scheme to enable efficient testing.

In the experiments, we used TOEIC as an objective measure of English language proficiency. We then applied some currently available automatic evaluation methods: BLEU, NIST and a DP-based method. We carried out experiments on two test sets: BTEC and SLTA1. According to the experimental results, the proposed method gave a good measurement result on a small-sized test set. The standard error of measurement is around 120 points on the TOEIC score with BTEC and less than 100 TOEIC points score with SLTA1. In both cases, the optimum size of the test set is 30 to 60 test sentences.

The proposed method still needs human labour to make the references. To obtain higher portability, we will apply an automatic paraphrase scheme (Finch et al., 2002; Shimohata and Sumita, 2002) to make the references automatically.

6 Acknowledgements

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled "A study of speech dialogue translation technology based on a large corpus".

References

CASEC. 2004. Computer Assessment System for English Communication. `http://www.ets.org/toefl/`.

A. Finch, T. Watanabe, and E. Sumita. 2002. "Paraphrasing by Statistical Machine Translation". In *Proceedings of the 1st Forum on*

Information Technology (FIT2002), volume E-53, pages 187–188.

G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. 2003. "Creating Corpora for Speech-to-Speech Translation". In *Proceedings of EUROSPEECH*, pages 381–384.

NIST. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurence Statistics. `http://www.nist.gov/speech/tests/mt/mt2001/resource/`.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

M. Shimohata and E. Sumita. 2002. "Automatic Paraphrasing Based on Parallel Corpus for Normalization". In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 453–457.

K.-Y. Su, M.-W. Wu, and J.-S. Chang. 1992. A new quantitative quality measure for machine translation systems. In *Proceedings of the 14th International Conference on Computational Linguistics(COLING)*, pages 433–439.

T. Takezawa, F. Sugaya, A. Yokoo, and S. Yamamoto. 1999. A new evaluation method for speech translation systems and a case study on ATR-MATRIX from Japanese to English. In *Proceeding of Machine Translation Summit (MT Summit)*, pages 299–307.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. "Toward a Broad-Coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World". In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 147–152.

T. Takezawa. 1999. Building a bilingual travel conversation database for speech translation research. In *Proceedings of the 2nd International Workshop on East-Asian Language Resources and Evaluation – Oriental COCOSDA Workshop '99 –*, pages 17–20.

TOEFL. 2004. Test of English as a Foreign Language. `http://www.ets.org/toefl/`.

TOEIC. 2004. Test of English for International Communication. `http://www.ets.org/toeic/`.