Incremental Methods to Select Test Sentences for Evaluating Translation Ability

Yasuhiro Akiba[†,‡], **Eiichiro Sumita**[†], **Hiromi Nakaiwa**[†],
Seiichi Yamamoto[†], and **Hiroshi G. Okuno**[‡]

† ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Keihana Science City, Kyoto 619-0288, Japan
‡ Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
{yasuhiro.akiba, eiichiro.sumita, hiromi.nakaiwa, seiichi.yamamoto}@atr.jp, okuno@i.kyoto-u.ac.jp

Abstract

This paper addresses the problem of selecting test sentences for automatically evaluating language learners' translation ability within a smaller error. In this paper, the ability to translate is measured as a TOEIC score. The existing selection methods only check whether an individual test sentence contributes to the estimation of the ability to translate or that of more general academic abilities, although combinations of test sentences may be used to contribute the estimation. This paper proposes two methods that solve the selection problem. The first method selects test sentences to minimize the estimation errors of learners' TOEIC scores. The second method selects test sentences to maximize the correlation coefficient between the number of correct translations and learners' estimated TOEIC scores. The optimization technique used in both of the proposed methods is the gradient technique in mathematical programming. The proposed methods proved to be more accurate than any of the existing methods we tested, and they estimated each TOEIC score within a permissible error of 69 points.

1 Introduction

This paper addresses the problem of selecting test sentences for automatically evaluating language learners' translation ability within a smaller error. This problem is regarded as an important issue in Test Theory (Wright and Stone, 1997) for precisely measuring learner ability. This paper attempts to solve the problem in the case of estimating a learner's TOEIC score based on other learners' scores. TOEIC (http://www.toeic.com) is the abbreviation of the "Test of English for International Communication" , which was created by ETS (Educational Testing Service) as TOEFL was.

The existing selection methods include Sugaya's selection method (Sugaya et al., 2002) and two selection methods in Classical Test Theory (Wright and Stone, 1997). These selection methods check whether an individual test sentence contributes to the estimation of the ability to translate or a more general human academic ability. However a combination of test sentences may contribute to the estimation.

This paper proposes two methods that solve the selection problem. In both methods, the selection problem is formalized as a 0-1 programming problem and is approximately solved by embedding it into a mathematical programming problem. The optimization technique used in both methods is the gradient technique in mathematical programming. This optimization technique enables the proposed selection methods to check whether a combination of test sentences contributes to the estimation.

To experimentally evaluate the proposed methods, the authors applied them to test sets of the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002) and the Travel Reservation Corpus (TRC) (Takezawa, 1999). The estimation errors by the proposed methods were smaller than those by each of the existing methods. Each TOEIC score was estimated within a permissible error of 69 points[1].

The next section outlines the TOEIC score estimation method used in this paper and the three existing selection methods. Section 3 proposes our two selection methods. Experimental results are shown and discussed in Section 4. Finally, our conclusions are presented in Section 5.

2 Related Works

This section outlines the methods to estimate TOEIC scores and three existing selection methods: Sugaya's selection method (Sugaya et al., 2002), the item-discrimination-power method, and the 50%-proportion-correct method. The first (Sugaya et al., 2002) is used in the discipline of machine translation evaluation (MTE) in order to select test sentences for evaluating speech-to-speech translation systems (Sugaya et al., 2000). The second and third (Wright and Stone, 1997) are used in the discipline of Classical Test Theory in order to select test sentences for evaluating a human academic ability.

2.1 Estimation of TOEIC scores

Sugaya et al. (2000) proposed the MTE method to estimate a Japanese-to-English translation system's TOEIC score based on human learners' TOEIC scores. In this MTE method, the translation result by the translation system and one by each human learner are manually compared for quality. This MTE method learns the regression line from the pairs of each human learner's TOEIC score and the percentage by which the system won against the learner. The trans-

[1] The original TOEIC scores are known to range in a band of ±69 points around the TOEIC scores with 95% confidence.

lation system's TOEIC score is estimated as the one that corresponds to the system's winning percentage of 0.5.

In this paper, the authors use the number of correctly translated sentences instead of the system's winning percentage to estimate TOEIC scores. This is because the calculation of the system's winning percentage needs a troublesome manual comparison and we would like to compare our selection method with the two selection methods in Classical Test Theory. The correctness of a translated sentence is simply judged by whether the translated sentence is identical to one of multiple reference translations, which are a set of correctly translated sentences in advance by professional translators. As proved later in Section 4.2, this simple test marking method is quite powerful.

2.2 Sugaya's selection method

In order to reduce the estimation error of the translation system's TOEIC score, Sugaya et al. (2002) proposed a test sentence selection method that selects test sentences that minimize the squared difference between the percentage that the translation system translated test sentences better than the i-th TOEIC examinee did, x_i, and their linear regression values predicted from the TOEIC score t_i of the i-th TOEIC examinee, $\beta_1 + \beta_2 * t_i$:

$$\begin{aligned} \sigma^2 &= \Sigma_{i=1}^{n}(x_i - (\beta_1 + \beta_2 * t_i))^2 \\ &= \beta_1^2 \, \Sigma_{i=1}^{n}(t_i - (1/\beta_2 * x_i - \beta_1/\beta_2)^2, \end{aligned}$$

where for each i (i=1,…,n)

$$x_i = \Sigma_{j=1}^{m} w_j * u_{i,j}.$$

$u_{i,j}$ takes one of three values, 1, 0, or 0.5, depending on whether the translation system translated the j-th test sentence better than, worse than, or even with a TOEIC examinee whose TOEIC score was t_i. Sugaya's selection method (Sugaya et al., 2002) solves the above minimization problem as a combinatorial optimization problem. Therefore Sugaya's selection method only checks whether an individual test sentence contributes to the estimation of the ability to translate.

2.3 Classical test theory

One of the major problems in Classical Test Theory (Wright and Stone, 1997) is to select test sentences for evaluating a human academic ability. In the discipline of Classical Test Theory, test sentences are called items. The strategy of the above selection is to maximize the correlation coefficient between the number of correct answers for items and a human academic ability. This section outlines two typical selection methods in Classical Test Theory: the Item-discrimination-power method and the 50%-proportion-correct method.

Item-discrimination-power method:

The Item-discrimination-power method only selects items that examinees with low academic ability cannot answer correctly and that examinees with high academic ability can answer correctly. Therefore, for the selection problem in this paper, the Item-discrimination-power method separately selects test sentences that low-scoring TOEIC examinees could not correctly translate and that high-scoring TOEIC examinees could correctly translate.

50%-proportion-correct method:

The 50% proportion correct method separately selects test sentences that have a 0.5 probability of being translated correctly. This approach aims to minimize the probability of answering items correctly by chance.

3 Proposed Methods

To solve the selection problem, this section proposes two methods. In both, the selection problem is formalized as a 0-1 programming problem as explained in Section 3.1 and is approximately solved by embedding it into a mathematical programming problem as explained in Section 3.2.

3.1 Two formalizations of the selection problem

Let us introduce the notations as follows:

1) let m denote the total number of test sentences,

2) let n denote the total number of TOEIC examinees used for the test sentence selection,

3) let t_i ($i = 1 \; to \; n$) denote the TOEIC score of the i-th examinee,

4) let $u_{i,j}$ ($i = 1 \; to \; n$, $j = 1 \; to \; m$) denote integer number 1 or 0 depending on whether or not the i-th examinee correctly translated the j-th test sentence, and

5) let w_j ($j = 1 \; to \; m$) denote a variable equal to integer number 1 or 0 depending on whether or not the j-th test sentence is selected.

Formalization 1.

Find numbers a, b, and w_j $(j = 1 \; to \; m)$ to minimize the squared difference between the TOEIC scores and their linear regression values:

$$L(a, b, w_1, \cdots, w_m) = \Sigma_{i=1}^{n}(t_i - y_i)^2,$$

where for each i (i=1,…,n)

$y_i = a * x_i + b, \; and \; x_i = \Sigma_{j=1}^{m} w_j * u_{i,j}.$ ∎

The minimization problem in Sugaya's selection method explained in Section 2.2 is equivalent to the problem that minimizes $L(\tilde{a}, \tilde{b}, w_1, \cdots, w_m)$, where $\tilde{a} = 1/\beta_2$ and $\tilde{b} = -\beta_1/\beta_2$.

Therefore, the differences between our Formalization 1 and Sugaya's one in the minimization problems is how the coefficients of the linear regression are treated. The coefficients in our Formalization 1 are adapted to the selected test sentences, while the coefficients in Sugaya's are fixed to constant real numbers and are not adapted to the selected test sentences.

Formalization 2.

Find numbers w_j $(j = 1 \; to \; m)$ to maximize the correlation coefficient between the TOEIC scores and the number of correctly translated sentences:

$$R(w_1, \cdots, w_m) \;=\; \frac{\Sigma_{i=1}^{n}(t_i - \bar{t_i})(x_i - \bar{x_i})}{\Sigma_{i=1}^{n}(t_i - \bar{t_i})^2 \Sigma_{i=1}^{n}(x_i - \bar{x_i})^2},$$

where for each i (i=1,. . . ,n)

$$x_i = \Sigma_{j=1}^{m} w_j * u_{i,j}.$$

$\bar{t_i}$ and $\bar{x_i}$ denote the averages of $\{t_i\}_{i=1,\cdots,n}$ and $\{x_i\}_{i=1,\cdots,n}$, respectively. ∎

3.2 Proposed methods

The first and second methods solve Formalizations 1 and 2, respectively. Each formalization is approximately solved by being embedded into a mathematical programming problem and by using the gradient technique.

Let $a^{(k)}$, $b^{(k)}$, and $w_j^{(k)}$ $(j \;=\; 1 \; to \; m)$, hereafter, denote the estimates of a, b, and w_j $(j \;=\; 1 \; to \; m)$, respectively, at the k-th iteration of the gradient technique.

3.2.1 Proposed method 1

The first method selects test sentences to minimize estimation errors of TOEIC scores in a similar way to that explained in Section 2.2. The first method principally differs from Sugaya's selection method in the following points. The first method uses the gradient technique for solving mathematical programming problems, while Sugaya's selection method uses a search technique for solving combinatorial optimization problem. The first method adapts the coefficients of the linear regression to the selected test sentences, while Sugaya's selection method does not.

The first method is described as follows:

(STEP 0)

If there are test sentences such that $u_{i,j} \;=\; 1$ for all i $(i \;=\; 1 \; to \; n)$ or such that $u_{i,j} = 0$ for all i $(i \;=\; 1 \; to \; n)$, remove such test sentences and let m denote the number of remaining test sentences.

(STEP 1)

Substitute 1 for $w_j^{(0)}$, substitute the coefficients of the linear regression that correspond to $w_j^{(0)}$ $(j \;=\; 1 \; to \; m)$ for $a^{(0)}$ and $b^{(0)}$, substitute 0 for k, and substitute 0.1 for λ. Here, λ is the parameter called "step width" for the gradient method.

(STEP 2)

$$w_j^{(k+1)} \;=\; w_j^{(k)} \\ -\lambda \times \frac{\frac{\partial L}{\partial w_j}(a^{(k)}, b^{(k)}, w_1^{(k)}, \ldots, w_m^{(k)})}{\sqrt{\sum_{j=1}^{m} \frac{\partial L}{\partial w_j}(a^{(k)}, b^{(k)}, w_1^{(k)}, \ldots, w_m^{(k)})^2}}.$$

Then, substitute for $a^{(k+1)}$ and $b^{(k+1)}$ the coefficients of the linear regression that corresponds to $w_j^{(k+1)}$ $(j \;=\; 1 \; to \; m)$ and add 1 to k.

(STEP 3)

(STEP 2) is repeated until $L(a^{(k+1)}, b^{(k+1)}, w_1^{(k+1)}, \cdots, w_m^{(k+1)})$ is not smaller than $L(a^{(k)}, b^{(k)}, w_1^{(k)}, \cdots, w_m^{(k)})$. Furthermore, let $\hat{k}$ denote the final iteration such that $L(a^{(\hat{k}+1)}, b^{(\hat{k}+1)}, w_1^{(\hat{k}+1)}, \cdots, w_m^{(\hat{k}+1)})$ is not smaller than $L(a^{(\hat{k})}, b^{(\hat{k})}, w_1^{(\hat{k})}, \cdots, w_m^{(\hat{k})})$.

(STEP 4)

For each j $(j \;=\; 1 \; to \; m)$, substitute 1 for w_j and select the j-th test sentence if $w_j^{(\hat{k})} \;\geq\; 0.5$. Substitute 0 for w_j and do not select the j-th test sentence if $w_j^{(\hat{k})} \;<\; 0.5$.

3.2.2 Proposed method 2

The second method selects test sentences to maximize the correlation coefficient between an independent and a dependent variable as in Classical Test Theory (Wright and Stone, 1997). The second method differs from the methods of Classical Test Theory in the maximization strategy.

The second method is described as follows:

(STEP 0)

The same as **(STEP 0)** of the first method.

(STEP 1)

Substitute 1 for $w_j^{(0)}$, substitute 0 for k, and substitute 0.1 for λ.

(STEP 2)

$$w_j^{(k+1)} \;=\; w_j^{(k)} \\ +\lambda \times \frac{\frac{\partial R}{\partial w_j}(w_1^{(k)}, \ldots, w_m^{(k)})}{\sqrt{\sum_{j=1}^{m} \frac{\partial R}{\partial w_j}(w_1^{(k)}, \ldots, w_m^{(k)})^2}}.$$

Then add 1 to k.

(STEP 3)

(STEP 2) is repeated until $R(w_1^{(k+1)}, \cdots, w_m^{(k+1)})$ is not larger than $R(w_1^{(k)}, \cdots, w_m^{(k)})$. Furthermore, let $\hat{k}$ denote the final iteration such that $R(w_1^{(\hat{k}+1)}, \cdots, w_m^{(\hat{k}+1)})$ is not larger than $R(w_1^{(\hat{k})}, \cdots, w_m^{(\hat{k})})$.

(STEP 4)

The same as **(STEP 4)** of the first method.

4 Experiment

The two proposed methods were evaluated in terms of how well they reduced the estimation errors. These methods were compared with the three existing methods, that is, Sugaya's selection method, the Item-discrimination-power method, and the 50%-proportion-correct method, explained in Section 2.

4.1 Experimental conditions

Test sets:

For the experiment, the authors applied each selection method to two test sets (Takezawa et al., 2002; Takezawa, 1999). One test set was randomly selected from the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002) and the other was selected from the Travel Reservation Corpus (TRC) (Takezawa, 1999). The first row in Table 1 shows the number of test sentences in each test set.

Reference translations:

To prepare the used multiple reference translations, the authors asked five native speakers of English who are familiar with Japanese to translate English sentences in the test

	[BTEC test set]				[TRC test set]			
	AVE.	SER.	MIN.	MAX.	AVE.	SER.	MIN.	MAX.
No selection	52.4	16.9	21.4	92.1	57.7	6.8	26.1	81.7
Proposed method 1	22.0	12.9	7.8	55.2	26.6	8.5	4.2	47.2
Proposed method 2	19.9	14.0	3.0	55.2	30.2	7.8	4.4	50.2
Existing method 1	34.8	15.9	13.0	70.7	33.8	7.8	13.9	66.1
Existing method 2	48.4	20.5	13.0	92.8	52.2	6.4	30.9	76.3
Existing method 3	58.2	26.2	0.5	109.4	95.8	27.0	6.9	214.9

Table 2: TOEIC estimation errors.

# of data	BTEC	TRC
Test sentences	510	330
Examinees	10	22
Candidates of leave-one-out examinees	4	7
Examinees for training the regression	9	21

Table 1: Numbers of data used

sets in three ways. Consequently, there were sixteen reference translations in English, including the English sentences existing originally in the test sets.

TOEIC examinees and their translation:

The TOEIC examinees were Japanese native-speakers who had scores between 255 and 745. The examinees were requested to present an official TOEIC score certificate showing that they had taken the test within the past six months. To measure the English capability of the Japanese native speakers, the TOEIC score was used. The second row in Table 1 shows the number of TOEIC examinees used.

The TOEIC examinees were asked to listen to Japanese text of the above test sets and provide an English translation on paper. The Japanese text was spoken twice within one minute, with a pause in-between.

The selection methods were evaluated in the leave-one-out cross validation. The leave-out TOEIC examinees for leave-one-out cross validation were restricted to the examinees who correctly translated more test sentences than the lower TOEIC-score examinees. This is because the examinees who correctly translated fewer test sentences than the lower TOEIC-score examinees may carelessly make a mistake or may not do his/her best. The last two rows in Table 1 show the number of leave-one-out TOEIC examinees and the number of the TOEIC examinees for training the regression, respectively.

4.2 Experimental results and Discussion

Table 2 shows the estimation errors by each method. The four columns for the test sets (labeled as AVE., SER., MIN., and MAX.) indicate the average estimation errors, the standard errors, the best-case estimation error, and the worst-case estimation error, respectively.

On average, the estimation errors by the proposed methods were smaller than those by each of the existing methods.

Even in the worst case (the columns labeled MAX.), the proposed methods estimated each TOEIC score within a permissible error of 69 points, which corresponds to a confidence interval at the 95% confidence level. The first existing method estimated a TOEIC score slightly outside the permissible range in the case of BTEC test set; however this is not permissible if we expect 95% confidence. The second and third existing methods estimated TOEIC scores extremely far outside the permissible error. The proposed methods proved to have a great potential for selecting the test sentences used to estimate the TOEIC score of language learners.

The correctness of a translated sentence in this paper is simply judged by whether the translated sentence is identical to one of multiple reference translations. The noteworthy point here is that the estimation errors by our proposed methods were reasonable small in spite of our adopting the simple test marking method.

This research is still at the very early stage. As a result, the prepared numbers of examinees and candidates of leave-one-out examinees were smaller than we would have like to prepare. To integrate e-learning systems, the authors plan to conduct a large-scaled experiment on estimating the TOEIC scores of language learners in the future.

5 Conclusions

This paper proposed two methods to automatically select test sentences for evaluating language learners' ability to translate. In automatically estimating learners' TOEIC scores, the proposed methods proved to be more accurate than any of the existing methods we tested. The proposed methods have the potential to select test sentences for estimating each TOEIC score within the permissible error.

Acknowledgment

This research was supported in part by the Telecommunications Advancement Organization of Japan.

References

Sugaya, F., Takezawa, T., Yokoo, A., and Yamamoto, S., 2000. Evaluation of ATR-MATRIX speech translation system with pair comparison method between the system and humans. In *Proc. of ICSLP2000*, volume 3, pages 1105–1108.

Sugaya, F., Yasuda, K., Takezawa, T., and Yamamoto, S., 2002. Quality-sensitive test set selection for a speech translation system. In *Proc. of ACL-02 Workshop on Speech-to-Speech Translation*, pages 109–116.

Takezawa, T., 1999. Building a bilingual travel conversation database for speech translation research. In *Proc. of the Oriental COCOSDA Workshop-99*, pages 17–20.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S., 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of LREC-2002*, pages 147–152.

Wright, B.D. and Stone, M.H., 1997. *Best Test Design*. Chicago: MESA.