Precise Measurement Method of a Speech Translation System's Capability with a Paired Comparison Method between the System and Humans

Fumiaki Sugaya[1], Keiji Yasuda[1,2], Toshiyuki Takezawa[1], Seiichi Yamamoto[1]

[1]ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
Email: sugaya@slt.atr.co.jp
[2]Graduate School of Engineering, Doshisha University
1-3, Tatara-miyakodani, Kyotanabe, Kyoto 610-0394, Japan

Abstract

The main goal of the present paper is to propose new schemes for the overall evaluation of a speech translation system. These schemes are expected to support and improve the design of the target application system, and precisely determine its performance. Experiments are conducted on the Japanese-to-English speech translation system ATR-MATRIX, which was developed at ATR Interpreting Telecommunications Research Laboratories. In the proposed schemes, the system's translations are compared with those of a native Japanese taking the Test of English for International Communication (TOEIC), which is used as a measure of one's speech translation capability. Subjective and automatic comparisons are made and the results are compared. A regression analysis on the subjective results shows that the speech translation capability of ATR-MATRIX matches a Japanese person scoring around 500 on the TOEIC. The automatic comparisons also show promising results.

Keywords

ATR-MATRIX, speech translation system, evaluation, paired comparison, TOEIC

1. Introduction

ATR Interpreting Telecommunications Research Laboratories earlier developed the ATR-MATRIX speech translation system (Takezawa et al., 1998), which translates both ways between English and Japanese. At ATR-SLT, we have been carrying out overall evaluations of this system through dialog tests and analyses (Sugaya et al., 1999). We have shown the effectiveness of the system in the basic hotel reservation task/domain.

Dialog tests are effective for evaluating the system. However, they do have demerits too, e.g., a lot of labor is required like test control, transcription, and tagging.

The Verbmobil project (Wahlster, 2000) conducted end-to-end evaluations to analyze the Verbmobil system. From our experiences, however, it is difficult to enlarge the evaluation target domains/tasks in the same way for ATR-MATRIX. Additional measures would be necessary to support the design of the system to meet performance expectations.

Machine translation systems have been evaluated with A, B, C, and D ranks (Sumita et al., 1999). This rank evaluation approach is useful for making relative system comparisons in time series among several schemes. However, one of its demerits is the lack of a direct relationship with the objective performance levels of the real target application systems. Tomita (Tomita et al., 1993) proposed a new scheme using the Test of English as a Foreign Language (TOEFL) to evaluate the quality of translated text as a whole. Evaluation results with this scheme can support the design of the target application systems and determine their performance. The scheme, however, cannot be applied to present speech translation systems, because its task/domain is limited.

We propose a new method (Sugaya et al., 2000) that is applicable to speech translation systems with a limited task/domain capability. In this method, both the system and humans with variable translation capabilities answer questions on the translations of test utterances taken from the target task/domain. The answers are compared by native evaluators. The comparison results show the existence of a matched point where both capabilities of the system and the humans match. Regression analysis clarifies the precise points.

In section 2, the new method is explained. In section 3, results for the language translation subsystem TDMT are presented. In section 4, results for speech recognition (SPREC) are shown. The effect of recognition errors on the language translation subsystem is also discussed.

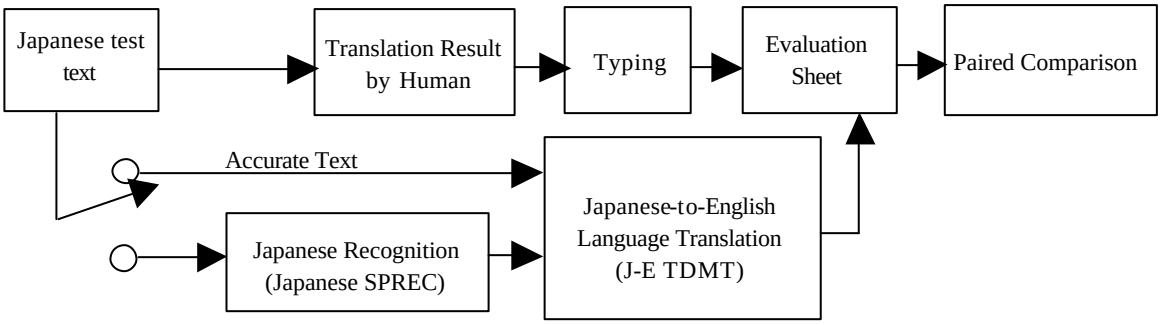

Figure 1: Diagram of translation paired comparison method

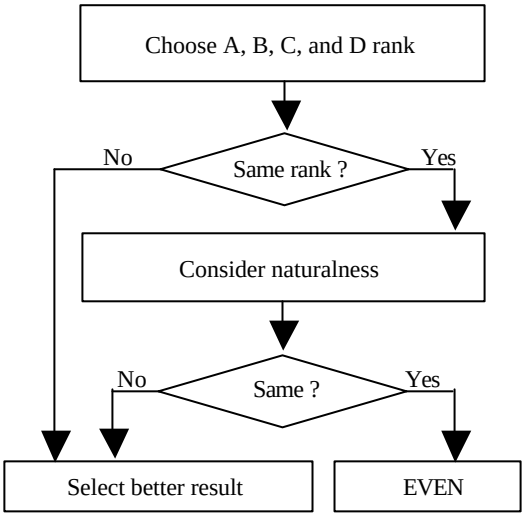

Figure 2: Procedure of comparative selection

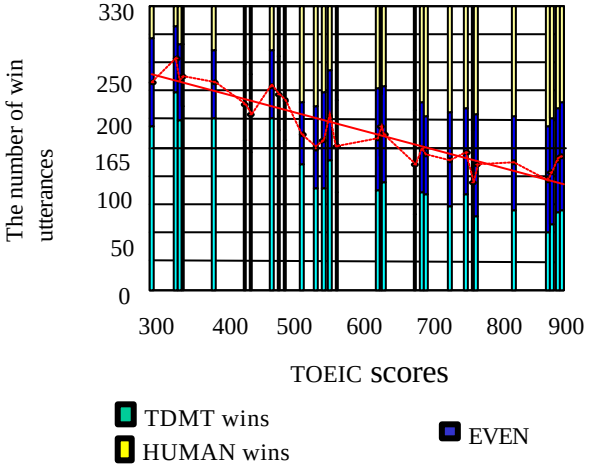

Figure3: Evaluation results for a language translation subsystem

An accuracy analysis of the proposed method is made in section 5. Section 6 shows effects of evaluators and summarizes evaluation results. In section 7, an automatic evaluation method is explained. A conclusion is given in section 8.

2. Translation Paired Comparison Method

Figure 1 shows a diagram of the proposed translation paired comparison method in the case of Japanese to English translation. The Japanese examinees are asked to listen to Japanese text and provide English translations on a piece of paper. The Japanese text is announced twice in a minute, and there is a pause in-between. To measure the English capabilities of the Japanese natives, the TOEIC score is used. The examinees must each present an official TOEIC score certificate showing that he/she has officially taken the test within the past six months.

The test text is from the SLTA1 test set, which consists of 330 utterances in 23 conversations from a bilingual travel conversation database (Morimoto et al., 1994). The SLTA1 test set is open for both speech recognition and language translation. The answers written on the pieces of paper are typed.

In the proposed method, the typed translation results by the examinees and the outputs of the system are merged to make evaluation sheets, and are compared by native Americans. The evaluation sheets show two translation results: the results of the examinees and those of the system in random order to eliminate discrimination by the native Americans. The native Americans are asked to follow the procedure in Fig. 2. The four ranks are the same as those used in (Sumita et al., 1999). The meanings of ranks A, B, C, and D are as follows: (A) Perfect: no problems in both information and grammar; (B) Fair: easy-to-understand with some unimportant information missing or flawed grammar; (C) Acceptable: broken but understandable with effort; (D) Nonsense: important information has been translated incorrectly.

3. Evaluation of Language Translation Subsystem

3. 1 Evaluation Results for Language Translation Subsystem

Figure 3 shows a comparison between the language translation subsystem TDMT and examinees. The inputs

for TDMT included accurate transcriptions. The total number of examinees was thirty, with five people in every range of one hundred TOEIC points between 300 and 900. Although we advertised for five examinees in every range of one hundred TOEIC points, only a couple of score holders hit the same point: 895. The horizontal axis in Figure 3 shows the TOEIC scores and each vertical bar against a TOEIC score shows an evaluation result. Each bar consists of three parts. From the horizontal line, we have the number of TDMT won utterances, the number of even (non-winner) utterances (which indicates no difference between the TDMT and examinee utterances), and the number of examinee won utterances. These three numbers sum to 330 (the total number of test utterances).

English native speakers able to understand Japanese judged the evaluation sheets. Figure 3 shows that the TDMT system won for TOEIC scores around 300-400. The examinees, in contrast, won for scores around 800. The capability-balanced area was found to be around 600-700. To precisely get the balanced point, we applied regression analysis.

To prepare the regression analysis, the number of even utterances was divided and put into the number of TDMT won utterances and examinee won utterances. The dotted line in Figure 3 shows this modified number of TDMT won utterances. The straight line shows the regression line. The capability balanced point between the TDMT subsystem and the examinees was determined to be the point where the regression line crossed half the number of all test utterances (330/2=165).

In Figure 3, the point is 707.6. Consequently, the translation capability of the language translation system equals that of an examinee with a score of around 700 points on the TOEIC. We call this point the system's TOEIC score.

3.2 Feature of Language Translation Subsystem

The number of TDMT won utterances is larger than that of examinee won utterances for lower TOEIC scores. In this area, the system dominates the match. The dominance rate (R) is defined by

$$R = \left(N_{TDMT} + \frac{N_{EVEN}}{2} \right) \Big/ \left(N_{HUMAN} + \frac{N_{EVEN}}{2} \right) \quad (1)$$

where N_{TDMT} is the number of TDMT won utterances, N_{HUMAN} is the number of examinee won utterances, and N_{EVEN} is the number of even utterances. A rate of more than one indicates that TDMT's capability is superior to the examinees'. In contrast, a rate of less than one indicates that TDMT's capability is inferior to the examinees'. If the rate equals one, this indicates that

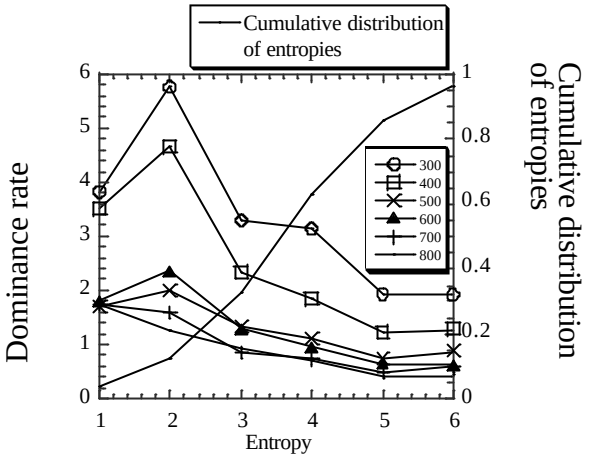

Figure 4: Dominance rate vs. entropy for a language translation subsystem

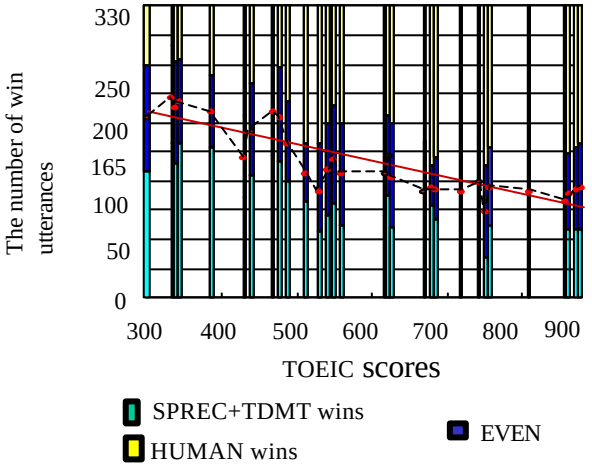

Figure5: Evaluation results for a language translation subsystem with a speech

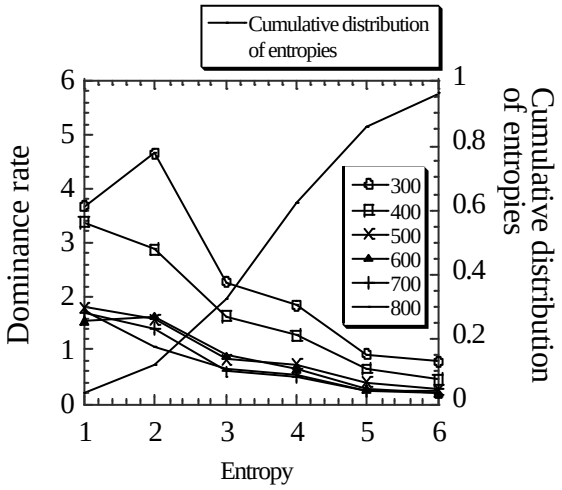

Figure 6: Dominance rate vs. entropy for a language translation subsystem with a speech recognizer

TDMT's capability matches that of the examinees. Figure 4 shows the dominance rate according to the average word entropy. The entropy is a logarithm of the perplexity, which is calculated by a language model using a variable-order N-gram algorithm (Masataki & Sagisaka, 1996).

In Figure 4, the dominance rate is averaged for every range of hundred points (TOEIC scores). The results of every TOEIC range (300, 400, 500, 600, 700, 800) are shown in Figure 4. The dominance rate curves of every TOEIC range decrease along the entropy axis. Around lower values of the entropy, the rate is large, which means that the TDMT system dominates in capability, while around higher values of the entropy, the rate is less than 1, which indicates that the human capability is superior to that of the TDMT system. Through a corpus analysis on our bilingual travel conversation database (Morimoto et al., 1994), the percentage of utterances with an entropy of 4 or less is so large, i.e., 62.5%, that TDMT works very effectively for the travel conversation task/domain.

4. Evaluation Results for Language Translation with Speech Recognition

Figures 5 and 6 show evaluation results of the TDMT language translation capability using speech recognition. All of the characteristics in the figures are almost similar to those in Figs. 3 and 4. However, the system's TOEIC score drops to 548.0, which is lower by 150 compared with the case of accurate transcriptions as the TDMT inputs. This degradation is due to the speech recognizer's performance.

Here, we define the dominance degradation rate as the ratio of the dominance rate of TDMT with a speech recognizer, and that of TDMT with accurate transcriptions, as R(SPREC+TDMT)/R(TDMT). Figure 7 shows that the speech recognition rate (WA: Word Accuracy) drastically falls with an increase in the entropy. The dominance degradation rate also shows a similar decrease. The correlation between the speech recognition rate and the dominance degradation rate is very high, i.e., 0.91. The dominance degradation rate is approximated as 1-(1-WA)*2.15.

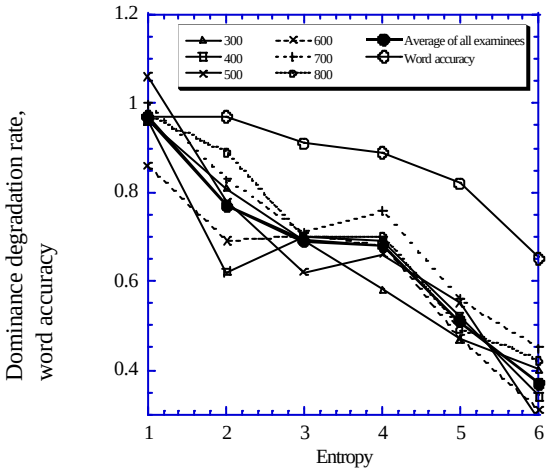

Figure 7: Dominance degradation rate, word accuracy vs. entropy

5. Errors in the System's TOEIC Score

The number of modified winning utterances for the system (X_i) and TOEIC scores for the examinees (Y_i) are assumed to satisfy the population regression equation:

$$Y_i = b_1 + b_2 X_i + e_i \quad (i = 1,2,...,n) \qquad (2)$$

where β_1, β_2 are population regression coefficients. The error term (ε_i) is assumed to satisfy the following conditions:

$$\begin{aligned} &(a) \quad E(e_i) = 0 \\ &(b) \quad V(e_i^{\,2}) = s^2, \quad i = 1,2,...,n \\ &(c) \quad Cov(e_i, e_j) = E(e_i e_j) = 0 \quad if \ i \neq j \\ &(d) \quad e_i \cong 0 \end{aligned} \qquad (3)$$

Under the above conditions, the standard deviation of the system's TOEIC score is calculated by

$$s_T = \left| \frac{s}{b_2} \right| \sqrt{\frac{1}{n} + \frac{(C_0 - \bar{X})^2}{\Sigma (X_i - \bar{X})^2}} \qquad (4)$$

where n is the number of examinees, C_0 is the system's TOEIC score, and $\bar{X}$ is the average of the examinees' TOEIC scores. Equation (4) indicates that the minimum error is given when the system's TOEIC score equals the average of the examinees' TOEIC scores.

By using a t-distribution, the confidence interval of the system's TOEIC score with confidence coefficient 1-α is given by

$$\left[C_0 - t_{a/2}(n-2) s_T, C_0 + t_{a/2}(n-2) s_T \right] \qquad (5)$$

6. Errors Among Evaluators

For the evaluation of the language translation subsystem with the speech recognizer, another evaluator also compared the evaluation sheets. The results had the same tendency. A regression analysis showed that the system's TOEIC score was 544.1, while the score was 548.1 for the other evaluator. The difference was as small as 3.7. In Table 1, numerical data calculated from the evaluation results and half of the confidence interval are summarized. Half of the confidence interval was 49.45, 56.73, or 45.92 for three cases when the confidence coefficient was 0.99.

7. Automatic Evaluation

In the above comparison between the system and humans, the comparison was subjective. In this section, an automatic comparison method is proposed. The basic idea is to use dynamic programming (DP)-based similarity. In this similarity, a translation result is compared with a translation answer by DP matching. After the DP matching, the distance between the two translations is scored by the following equation:

$$\text{Similarity} = (\text{Total} - \text{Sub} - \text{Ins} - \text{Del})/\text{Total} \quad (6)$$

where Total is the total number of words in the translation answer, Sub is the number of substituted words, Ins is the number of inserted words, and Del is the number of deleted words.

		SPREC+TDMT	
	TDMT	Eva. I	Eva. II
β_1	307.60	265.64	252.97
β_2	-0.20	-0.18	-0.16
system's TOEIC score	707.56	548.05	544.41
σ	17.21	19.63	13.90
average TOEIC score of examinees	606.05		
variance of TOEIC scores	32056.92		
the number of examinees	30		
σ_T	17.90	20.53	16.62
confidence interval/2	49.45	56.73	45.92

Table 1: Summary of evaluation results

7. 1 Collection of Translation Answer Set with Paraphrasing

In the DP-based similarity, the translation quality is measured as the distance between the translation result and its answer. In most cases, the translation answer is not limited to only one expression, but probably has many variants against each Japanese source sentence. Because of this, a good translation without the appropriate answer will lead to a bad score by the DP-based similarity.

To solve this problem, we collected English paraphrased translation answers against each Japanese, by five Americans with a sufficient Japanese understanding. Each American made three paraphrased answers for each Japanese. Some of the answers from the different American were duplicate sentences. The average number of paraphrased answers was 14.42 sentences for each Japanese, indicating that each American gave diversified sentences in the case of three sentence generation per person.

7. 2 Answer Set Similarity

In the expansion of answer candidates, translation results have better chances to match the most similar answers. The maximum similarity among paraphrased translation answers for each Japanese sentence is defined as the answer set similarity. Figure 8 shows the average answer set similarity for individual Japanese examinees taking the TOEIC test. A Japanese person with a higher score shows a higher answer set similarity. The average answer set similarity for TDMT and TDMT with the SPREC speech recognizer are 0.48 and 0.45. Using the regression line in Figure 8, the TOEIC score for TDMT is calculated as 682.9 and that for TDMT with the recognizer is 547.3. The difference of the TOEIC scores between the subjective and objective evaluations is small. The error is within the limit of Table 1. Considering the reductions in the evaluation cost and time, this automatic scheme shows a decent performance and is very promising.

7. 3 Effects of Paraphrased Sentences

Figure 9 shows the relationship between the similarity and a subjective system's winning rate normalized by the total number of test sentences. The data in Figure 9 is the same as in Figure 5. The correlations between the similarities and the winning rates are as high as 0.9 or larger, especially for the case of the answer set similarity. The higher correlation of the answer set similarity shows a greater effectiveness due to the expansion of similar translation answers. The volume of the paraphrased sentences and its relation to the precision of the similarity are remaining big issues.

8. Conclusion

We proposed a translation paired comparison method for a speech translation system. This method is applicable to wider tasks/domains without additional labor like dialog tests. We evaluated the ATR-MATRIX system. The results showed that the system's capability equals that of native Japanese scoring around 548 points on the TOEIC. According to public information on the TOEIC, the average TOEIC score of university students in Japan is 568 points. Even considering the confidence interval, ATR-MATRIX nearly approaches the average university student's speech translation capability, which is achieved at tremendous costs even with the limited task/domain involved.

The system's performance in language translation and speech recognition, and its effects on language translation, are strongly related to the entropy. This entropy is a measure of information, and should not be directly related to the task/domain. Accordingly, the performance dependency on the entropy can be expected to be valid in

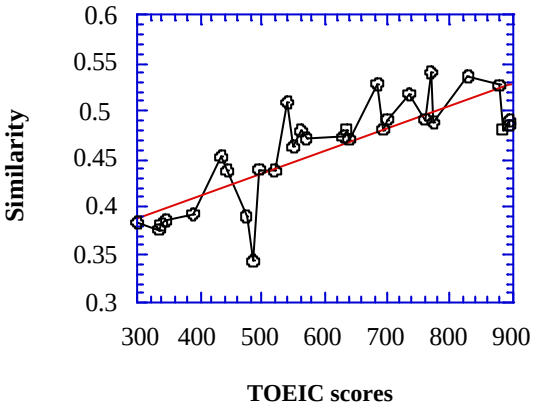

Figure 8: Answer set similarity for various Japanese people taking the TOEIC test

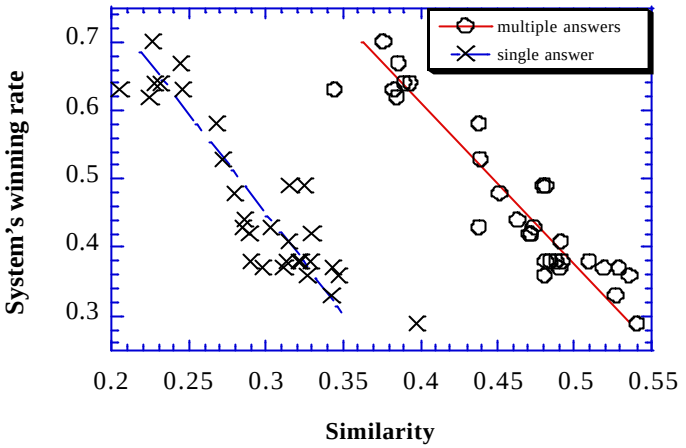

Figure 9: Effect of paraphrased sentences

other tasks/domains. The results in this paper are taken and analyzed in a limited domain/task, but the entropy dependency of the performance can hopefully be used effectively to select new tasks/domains. A promising automatic evaluation method is also proposed.

References

1. Masataki, H. & Sagisaka, Y. (1996). Variable-Order N-gram Generation by Word-Class Splitting and Consecutive Word Grouping. In Proceedings of ICASSP'96 (pp. 188--191).
2. Morimoto, T., Uratani, N., Takezawa, T., Furuse, O., Sobashima, Y., Iida, H., Nakamura, A., Sagisaka, Y., Higuchi, N. and Yamazaki, Y. (1994). A speech and language database for speech translation research. In Proceedings of ICSLP '94 (pp. 1791--1794).
3. Sugaya, F., Takezawa, T., Yokoo, A. and Yamamoto, S. (1999). End-to-end evaluation in ATR-MATRIX: speech translation system between English and Japanese. In Proceedings of Eurospeech'99 (pp. 2431--2434).
4. Sugaya, F., Takezawa, T., Yokoo, A. and Yamamoto, S. (2000). Evaluation of the ATR-MATRIX speech translation system with a paired comparison method between the system and humans. In Proceedings of ICSLP'2000, Vol. III (pp. 1105--1108).
5. Sumita E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K. and Shirai, S. (1999). Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach. In Proceedings of MT Summit '99 (pp. 229--235).
6. Takezawa, T., Morimoto, T., Sagisaka, Y., Campbell, N., Iida, H., Sugaya, F., Yokoo, A. and Yamamoto, S. (1998). A Japanese-to-English speech translation system: ATR-MATRIX. In Proceedings of ICSLP 1998 (pp. 2779--2782).
7. Tomita, M., Shirai, M., Tsutsumi, J., Matsumura, M. and Yoshikawa, Y. (1993). Evaluation of MT Systems by TOEFL. In Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'93) (pp. 252--259).
8. Wahlster, W. (2000). Verbmobil: Foundations of Speech-to-Speech Translation. Springer.