An Automatic Evaluation Method of Translation Quality Using Translation Answer Candidates Queried from a Parallel Corpus

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

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

Abstract

An automatic translation quality evaluation method is proposed. In the proposed method, a parallel corpus is used to query translation answer candidates. The translation output is evaluated by measuring the similarity between the translation output and translation answer candidates with DP matching. This method evaluates a language translation subsystem of the Japanese-to-English ATR-MATRIX speech translation system developed at ATR Interpreting Telecommunications Research Laboratories. Discriminant analysis is then carried out to examine the evaluation performance of the proposed method. Experimental results show the effectiveness of the proposed method. The discriminant ratio is 83.5% for 2-class discrimination between absolutely correct and less appropriate translations classified subjectively. Also discussed are issues of the proposed method when it is applied to speech translation systems which inevitably make recognition errors.

Keywords

MT evaluation techniques, Translation quality, Parallel Corpus, DP matching, ATR-MATRIX

1. Introduction

An automatic translation quality evaluation method is required since subjective evaluations require large costs and the turn-around time is very long.

Over the past few years, we have carried out several studies on the evaluation of the speech translation system called ATR-MATRIX (ATR's Multilingual Automatic Translation System for Information Exchange) (Takezawa *et al.*, 1998). Examples of these studies include a rank evaluation method (Sumita *et al.*, 1999), a translation paired comparison method (Sugaya *et al.*, 2000), and an evaluation method using dynamic programming (DP) (Takezawa *et al.*, 1999). The first two methods mentioned above can be considered subjective evaluation methods. The last one is an automatic evaluation method that evaluates the robustness of language translation subsystems against speech recognition errors. This is not to evaluate the translation quality.

Keh-Yin Su *et al*. proposed the most conventional automatic evaluation method of translation quality using DP matching (1992). However, the method appears to lack an evaluation capability to deal with multiple expressions in a target language utterance, since only one translation answer is provided in this method.

The method proposed in this paper solves the problem by using multiple translation answer candidates queried from a parallel corpus by DP matching.

In section 2, the proposed method is explained. In section 3, evaluation results by the proposed method and the most conventional automatic evaluation method are compared from the viewpoint of evaluation performance. To reduce the subjective rank evaluation cost, and improve the evaluation performance, a hybrid method is explained, that is, a combination between the proposed method and the subjective rank evaluation method. Also discussed are issues of the proposed method when it is applied to evaluation of the outputs of speech translation systems, which inevitably make errors in the recognition stage. In section 4, we state our conclusion.

2. Evaluation Method

The similarity between a translation output and a correct answer utterance can be calculated by DP matching as follows:

$$\sigma = \frac{T - S - I - D}{T} \quad (1)$$

where σ is the similarity, T is the total number of words in the correct answer utterance, S is the number of substitution words comparing the correct answer utterance to the translation output, I is the number of inserted words comparing the correct answer utterance to the translation output, and D is the number of deleted words comparing the correct answer utterance to the translation output.

Figure 1 shows a diagram of the proposed method. The source language corpus and the target language corpus have a parallel translation relationship. Also the source language test utterance and the target language test utterance have the same relationship.

In the most conventional automatic evaluation method, the quality of a translation is evaluated by calculating the similarity between the translation output and the target language test utterance. We use the term "conventional DP method" to refer to this evaluation method.

In the proposed method, in contrast, translation answer candidates are added to the target language test utterance. We define a set of test utterance and translation candidates, as an answer set. Each translation of each utterance in the source language corpus is regarded as a translation answer

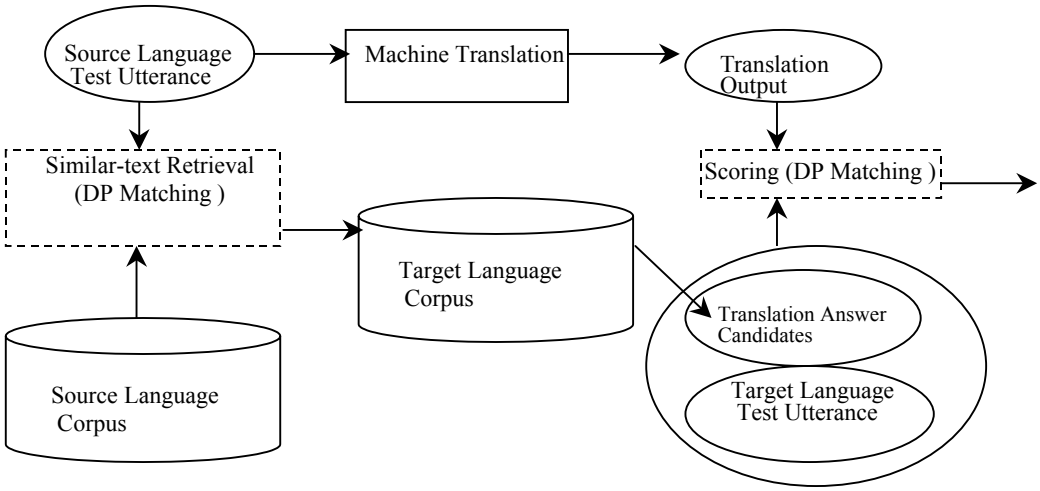

Figure 1: Diagram of the proposed method

candidate only if the similarity between the source language test utterance and the utterance in the source language corpus is larger than a threshold. This threshold is defined as the "retrieval threshold". Then, we calculate the similarity between the translation output and each utterance in the answer set. As a result of these calculations, multiple similarities are obtained. We define the maximum similarity as the answer set similarity, which is the measure of the translation output quality in the proposed method.

3. Evaluation Experiment

TDMT (Transfer Driven Machine Translation) is a language translation subsystem of the Japanese-to-English speech translation system ATR-MATRIX. Our current focus is on automatically evaluating the translation quality. Therefore, we use the TDMT output, which is a translation output from transcription texts, as the evaluation target. Also an evaluation result of the Japanese-to-English ATR-MATRIX is shown in section 3.5.

In this evaluation experiment, the ATR bilingual travel conversation database (Takezawa, 1999), which consists of 16110 utterances in each of the languages employed, was applied to calculate the answer set similarity. The test set was a subset of the database, and consists of 330 utterances in each language.

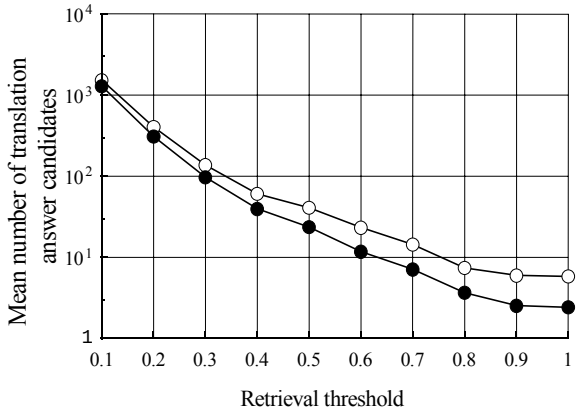

Figure 2: Relationship between retrieval threshold and mean number of translation candidates

3.1 Rank Evaluation Method

In the rank evaluation method, each utterance was assigned one of four ranks for translation quality: (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. We use the term "translation rank" to refer to each result of the rank evaluation method.

3.2 Evaluation Result by the Proposed Method

Figure 2 shows a relationship between the number of translation answer candidates and the retrieval threshold. The ordinate is the number of translation answer candidates. In Figure 2, each plot indicates the mean number of translation answer candidates per one test utterance. Each plot type indicates a different numeration as follows: circle: raw number of translation answer candidates; filled circle: number of unique translation answer candidates.

Table 1 shows examples of translation answer candidates when the retrieval threshold is 0.6.

Figure 3 shows a relationship between the translation rank and the answer set similarity. Each plot indicates the

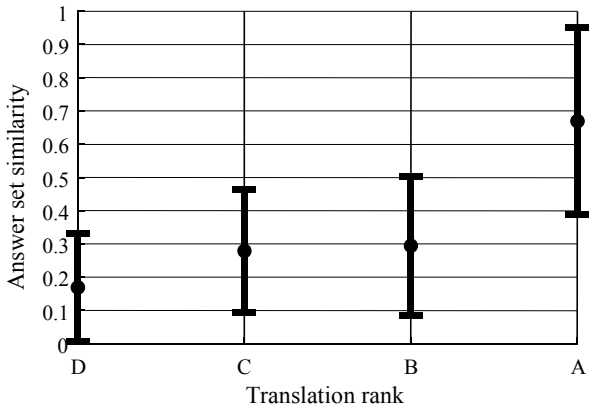

Figure 3: Relationship between the translation rank and answer set similarity evaluating outputs of the

language translation subsystem TDMT

Source language test utterance	Target language test utterance
hai/wakari/masi/ta/o/sirabe/si/masu/node/shoushou/o/machi/kudasai	All right. Please hold the line and I will check.

Translation answer candidates in the source language	Translation answer candidates
kasikomarimasita/o/sirabe/itasi/masu/node/shoushou/o/machi/kudasai	Okay, let me check. Just a moment please.
hai/o/sirabe/si/masu/shoushou/o/machi/kudasai/mase	Okay, could you wait for a moment while I check.
wakari/masi/ta/kakunin/si/masu/node/shoushou/o/machi/kudasai	Okay, I 'll check for you please hold on a moment.
o/sirabe/itasi/masu/node/syousyou/o/machi/kudasai	One moment please. I 'll check on availability.
tadaima/o/sirabe/si/masu/node/shoushou/o/machi/kudasai/mase	Could you hold on a minute while I check please.

Table 1: Examples of translation answer candidates

mean answer set similarity in a translation rank, and each error bar indicates the standard deviation of the answer set similarity in a translation rank. This relationship is obtained by setting the retrieval threshold as 0.6. As shown in Figure 3, the higher the translation rank is, the higher the mean answer set similarity becomes.

3.3 Evaluation Performance

A discriminant analysis was carried out to examine the evaluation performance of the proposed method.

3.3.1 Discriminant Analysis

The translation rank was used for this discriminant analysis. A discriminant rule was the nearest neighbor rule.

We conducted four types of discriminations as follows:

Type1: 2-class discrimination: discriminate between (A) and (B) (C) (D).
Type2: 2-class discrimination: discriminate between (A) (B) and (C) (D).
Type3: 2-class discrimination: discriminate between (A) (B) (C) and (D).
Type4: 4-class discrimination: discriminate among (A), (B), (C), and (D).

The discriminant ratio D can be formulated as follows:

$$D = \frac{N_{correct}}{N_{total}} \qquad (2)$$

where $N_{correct}$ is number of the utterances correctly discriminated, and N_{total}, the total utterance count of the test set.

3.3.2 Result of the Discriminant Analysis

Figure 4 shows a result of the discriminant analysis. The ordinate is the discriminant ratio, and the abscissa is the retrieval threshold or the conventional DP method in the right end, depicted in separation. In Figure 4, although the highest point of the discriminant ratio is different for the different types of discriminations, the discriminant ratio using the answer set similarity is larger than that of the conventional DP method. Namely, the proposed method is superior to the conventional DP method in terms of the evaluation performance. In particular, when the retrieval threshold is 0.6 for the type1 discrimination, the discriminant ratio is fairly high (*i.e.* 83.5%). The gain from the conventional DP method is also high (*i.e.* 15%).

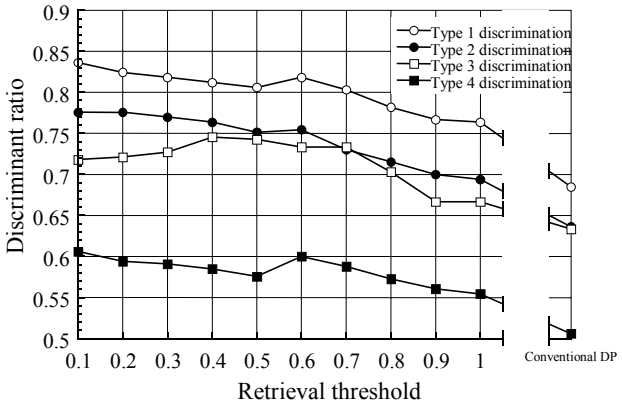

Figure 4: Relationship between retrieval threshold and discriminant ratio

3.4 Combination of the Rank Evaluation Method and the Proposed Method

The proposed method improves evaluation performance as described in the preceding subsection. However, some situations require a higher evaluation accuracy. We therefore made a study on combining the rank evaluation method and the proposed method to settle this issue. Specifically, we discriminate (A) automatically, and evaluate the output, that is, a state is discriminated as not (A), by the rank evaluation method. We use the term "discriminant threshold" to refer to the threshold for discrimination.

First, we define four states: *class1*: state evaluated as (A) by the rank evaluation method; *class2*: state evaluated as (B), (C), or (D) by the rank evaluation method; *CLASS1*: state discriminated as (A) by the automatic discrimination; *CLASS2*: state discriminated as not (A) by the automatic discrimination. Using the four states, four probabilities can be defined as shown in Table 2.

		Rank evaluation method	
		class1	*class2*
Automatic discrimination	*CLASS 1*	*P(CLASS1|class1)*	*P(CLASS1|class2)*
	CLASS 2	*P(CLASS2|class1)*	*P(CLASS2|class2)*

Table 2: States to be defined

In Table 2, $P(CLASS1|class1)$ denotes the correct acceptance ratio: the probability to be discriminated as (A) correctly, $P(CLASS1|class2)$ denotes the false acceptance ratio: the probability to be discriminated as (A) wrongly, $P(CLASS2|class1)$ denotes the false rejection ratio: the probability to be discriminated as not (A) wrongly, and $P(CLASS2|class2)$ denotes the correct rejection ratio: the probability to be discriminated as not (A) correctly.

Figure 5 shows the relationship between the discriminant threshold and the false rejection ratio or the false acceptance ratio. This relationship was obtained from the answer set similarity setting retrieval threshold as 0.6. In Figure 5, the ordinate is the false rejection ratio or the false acceptance ratio, and the abscissa is the discriminant threshold.

Figure 6 shows the same relationship using an evaluation result by the conventional DP method.

Figure 7 shows ROC curves (Receiver Operating Characteristic curves) obtained from the relationships depicted in Figure 5 and Figure 6. In Figure 7, the ordinate is the correct acceptance ratio, and the abscissa is the false acceptance ratio.

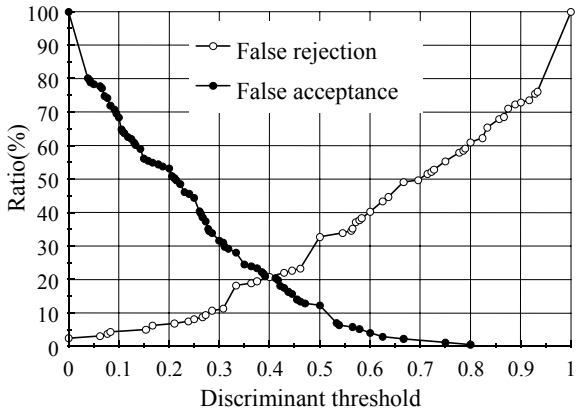

Figure 5: Relationship between discriminant threshold and ratio for false rejection and false acceptance, using the answer set similarity

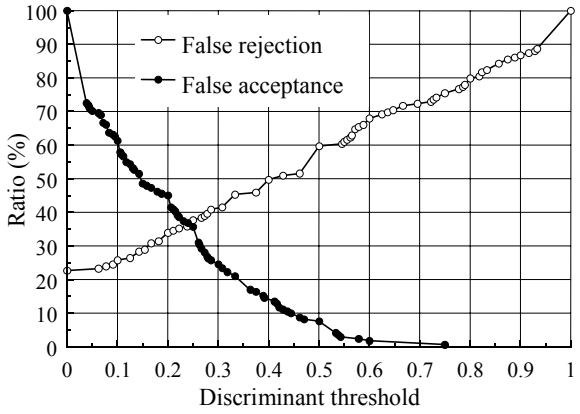

Figure 6: Relationship between discriminant threshold and ratio for false rejection and false acceptance, using an evaluation result of the conventional DP method

For the combined method described in this subsection, a false rejection is not fatal since *CLASS2* is evaluated by the rank evaluation method. A false rejection only provokes an increase in the rank evaluation cost. On the other hand, a false acceptance is fatal since *CLASS1* will not be evaluated again. From this point of view, we define the "cost reduction ratio" R and "error ratio" E as follows:

$$\begin{aligned} R &= P(CLASS1) \\ &= P(CLASS1|\,class1) \times P(class1) \\ &\quad + P(CLASS1|\,class2) \times P(class2) \end{aligned} \tag{3}$$

$$E = \frac{P(CLASS1|\,class2) \times P(class2)}{P(CLASS1)} \tag{4}$$

The cost reduction ratio R is defined as the ratio of (A) discriminated by the automatic discrimination to the test set. In Equation (3), $P(class1)$ is the ratio of (A) evaluated by the rank evaluation method to the test set, $P(class2)$ is the ratio of (B), (C), and (D) evaluated by the rank evaluation method to the test set. The error ratio E is defined as the ratio of (A) discriminated by the automatic discrimination to R, when the evaluation result by the rank evaluation method is (B), (C), or (D). Evaluating TDMT by the rank evaluation method, $P(class1)$ yields 0.48, and $P(class2)$ yields 0.52.

Figure 8 shows the relationship between the cost reduction ratio and the error ratio. Figure 8 is obtained from the relationship depicted in Figure 7. By using the answer set similarity for the discrimination, a 30% cost reduction can be earned when a 5% error is accepted (broken line in Figure 8). By using the evaluation result of conventional DP method, the cost reduction ratio is only 20%.

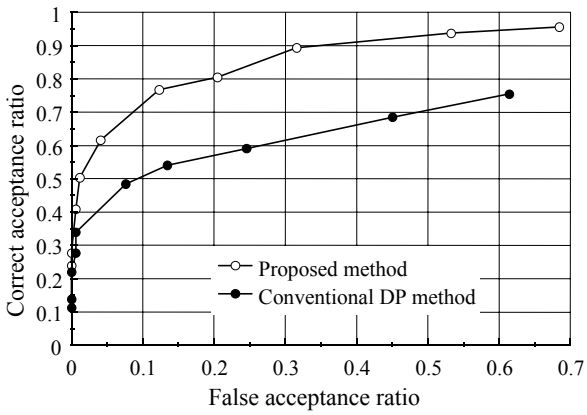

Figure 7: Relationship between false acceptance ratio and correct acceptance ratio

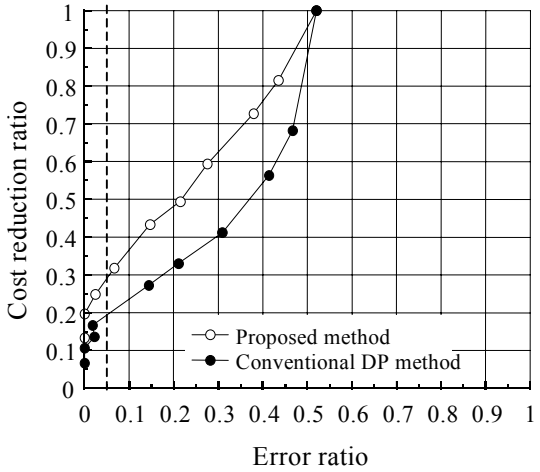

Figure 8: Relationship between error ratio and cost reduction ratio

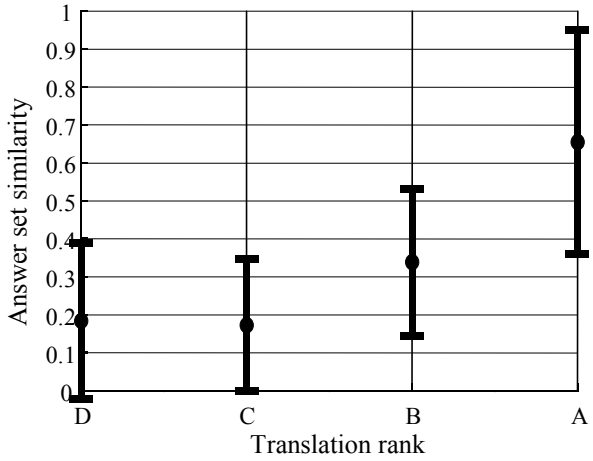

Figure 9: Relationship between translation rank and answer set similarity, evaluating output of speech translation subsystem ATR-MATRIX

3.5 Application to the Evaluation of a Speech Translation System

We used the proposed method to the evaluation of the Japanese-to-English ATR-MATRIX to examine its applicability for speech translation output evaluation.

Figure 9 shows a relationship between the translation rank and the answer set similarity. Each plot indicates the mean answer set similarity in a translation rank, and each error bar indicates the standard deviation of the answer set similarity in a translation rank. Most of the features are almost similar to those in Figure 3. In Figure 9, however, the mean answer set similarity of (D) is larger than that of (C). If we compare Figure 3 and Figure 9, the standard deviation of the answer set similarity of (D) in Figure 9 is larger than that in Figure 3.

These issues are due to recognition errors as shown by the examples in Table 3. The underlines in the table indicate recognition errors. Here, *gomannanasen*[fifty seven thousand] is recognized as *gomarunanasen*[five zero seven thousand], *washinton* [Washington] is recognized as *ashino* [foot], and *ichi* [one] is deleted. All of the translation outputs shown in the table were evaluated as (D) by the rank evaluation method. In these examples, recognition errors did not affect the parser, so the sentence structures of the translation outputs were not broken.

However, some important information was not correctly transferred including the price, hotel name, and number. This was due to a drastic decline in the translation rank, and a slight decline in the answer set similarity.

Example 1	Source language test utterance	konekutinguruumu ga ippaku gomannanasen en to natte orimasu
	Recognition result	konekutingu ruumu ga ippaku gomarunanasen en to natte orimasu
	Translation answer	A connecting room is fifty seven thousand yen per night .
	Translation output	A connecting room is five zero seven thousand yen per night .
	Answer set similarity	0.8
Example 2	Source language test utterance	ima washintonhoteru ni taizai site imasu
	Recognition result	ima ashinohoteru ni taizai site imasu
	Translation answer	I 'm staying at the washington hotel in Washington .
	Translation output	I 'm staying at the foot hotel in Washington now .
	Answer set similarity	0.78
Example 3	Source language test utterance	ni ichi san go yon san no ichi nana gou gou
	Recognition result	ni ___ san gou yon san no ichi nana gou gou
	Translation answer	Two one three five four three one seven five five .
	Translation output	Two three five four three , one seven five five .
	Answer set similarity	0.9

Table 3: Examples of recognition errors

4. Conclusion

We proposed a new automatic translation quality evaluation method. A parallel corpus is used to query translation answer candidates in this method. The translation output is evaluated by measuring the similarity between translation output and the translation answer candidates with DP matching.

The proposed method evaluates TDMT, a language translation subsystem of the Japanese-to-English ATR-MATRIX speech translation system. Then, a discriminant analysis is carried out to examine the evaluation performance of the method. Results of the discriminant analysis show an improvement in the evaluation performance, comparing the proposed method with the most conventional automatic evaluation method.

However, the evaluation results of a speech translation system suggest a weakness in the evaluation capability against recognition errors.

There is room for further investigating for the optimal value for the retrieval threshold and the robustness of the evaluation capability against speech recognition errors.

Acknowledgements

The authors would like to thank all members of the ATR Spoken Language Translation Laboratories. This research was supported in part by a grant from the Academic Frontier Project promoted by Doshisha University.

Bibliographical References

Takezawa, T. (1999). Building a bilingual travel conversation database for speech recognition research. In Proceeding of Oriental COCOSDA Workshop.

Takezawa, T., Morimoto, T., Sagisaka, Y., Campbell, N., Iida, H., Sugaya, F., Yokoo, A. & Yamamoto, S. (1998). A Japanese-to-English speech translation system: ATR-MATRIX. In Proceeding of ICSLP (pp. 2779--2782).

Takezawa, T., Sugaya, F., Yokoo, A. & Yamamoto, S. (1999). A New Evaluation Method for Speech Translation Systems and a Case Study on ATR-MATRIX from Japanese to English. In Proceeding of MT Summit (pp. 299--307).

Su, K. -Y., Wu, M. -W. & Chang, J. -S. (1992). A new quantitative quality measure for machine translation systems. In Proceeding of COLING (pp. 433--439).

Sugaya, F., Takezawa, T., Yokoo, A., Sagisaka, Y. & Yamamoto, S. (2000). Evaluation of the ATR-MATRIX Speech Translation System with Pair Comparison Method Between the System and Humans. In Proceeding of ICSLP (pp. 1105--1108).

Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishilawa, K. & Shirai, S. (1999). Solutions to Problems Inherent in Spoken Language Translation : The ATR-MATRIX Approach. In Proceeding of MT Summit (pp. 229--235).