AUTOMATIC EVALUATION METHODS OF A SPEECH TRANSLATION SYSTEM'S CAPABILITY

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
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 automatic schemes of the translation paired comparison method, which was proposed by the authors to precisely evaluate a speech translation system's capability. In this method, the outputs of the speech translation system are subjectively compared with results of native Japanese taking the Test of English for International Communication (TOEIC), which is used as a measure of one's speech translation capability. Experiments are conducted on TDMT, which is a language translation subsystem of the Japanese-to-English speech translation system ATR-MATRIX developed at ATR Interpreting Telecommunications Research Laboratories. The winning rate of TDMT shows a good correlation with the TOEIC scores of the examinees. A regression analysis on the subjective results shows that the translation capability of TDMT matches a Japanese person scoring around 700 on the TOEIC. The automatic evaluation methods use DP-based similarity, which is calculated by DP distances between a translation output and multiple translation answers. The answers are collected by two methods: paraphrasing and query from a parallel corpus. In both types of collection, the similarity shows the same good correlation with the TOEIC scores of the examinees as the subjective winning rate. A regression analysis using the similarity shows that the system's matched point is around 750. In this paper, we also show effects of paraphrased data.

1. INTRODUCTION

ATR Interpreting Telecommunications Research Laboratories earlier developed the ATR-MATRIX speech translation system [1], 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 [2]. To date, 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 [3] conducted end-to-end evaluations to analyze the Verbmobil system. From our experiences, however, it is difficult to enlarge the evaluation target domain/task 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 [4]. 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 [5] 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 obtained with this scheme have been observed to support the design of the target application system and determine its performance. The scheme, however, cannot be applied to present speech translation systems, because their tasks/domains are limited.

We earlier proposed the above-mentioned translation paired comparison method [6], which is applicable to the evaluation of 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 have shown the existence of a matched point where both the capabilities of the system and the humans match. A regression analysis clarifies the precise point.

A major demerit of the translation paired comparison method is its subjective evaluation approach. Such an approach requires large costs and a long evaluation time. In this paper, we propose two automatic evaluation methods to address these issues.

Section 2 explains the proposed evaluation. Section 3 presents evaluation results by the proposed methods and some comparisons. In section 4, we state our conclusion.

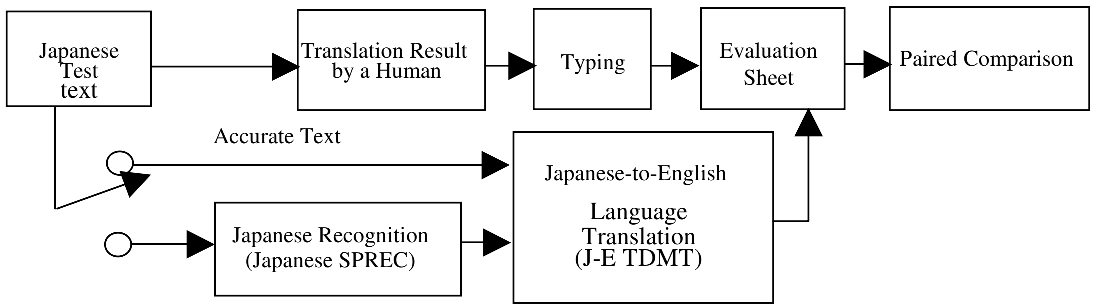

Figure 1: Diagram of translation paired comparison method

2. EVALUATION METHODS

2.1. Translation paired comparison method

Figure 1 shows a diagram of the earlier proposed translation paired comparison method in the case of Japanese to English translation. Using the method, Japanese examinees are asked to listen to Japanese text and provide English translations on a piece of paper. The Japanese text is announced twice 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 the ATR bilingual travel conversation database [7]. The SLTA1 test set is open for both speech recognition and language translation. The answers written on the pieces of paper are typed.

In this 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 Figure 2. The four ranks are the same as those used in [4]. 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.

2.2. Automatic evaluation method

The similarity between a translation output and a correct answer utterance can be calculated by DP matching [8][9] as follows:

$$\sigma = \frac{T - S - I - D}{T} \tag{1}$$

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,

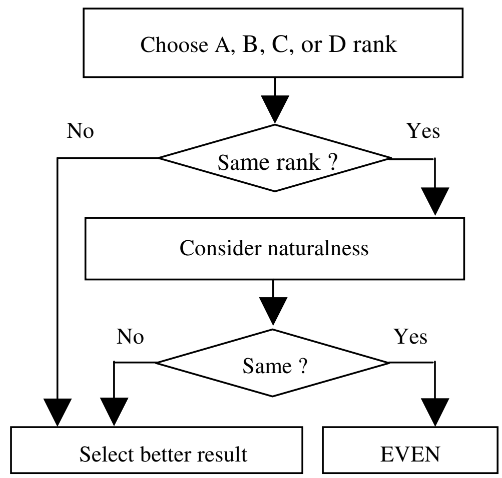

Figure 2: Procedure of comparative selection

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.

For the 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 similarity.

To solve this problem, we consider two approaches for collecting multiple answer candidates as explained in sections 2.2.1 and 2.2.2.

By expanding the number of answer candidates, translation results have better chances of matching the most similar answers. The maximum similarity among paraphrased translation answers for each Japanese sentence is defined as the answer set similarity.

2.2.1. Collection of translation answer set with paraphrasing

We collect English paraphrased translation answers against each Japanese source sentence, by employing Americans (five in this case) with a sufficient Japanese understanding. Each American makes three paraphrased answers for each Japanese sentence. Some of the answers from the different Americans are duplicate sentences. The addition of the target utterance in the parallel corpus to the paraphrased answers produces an average

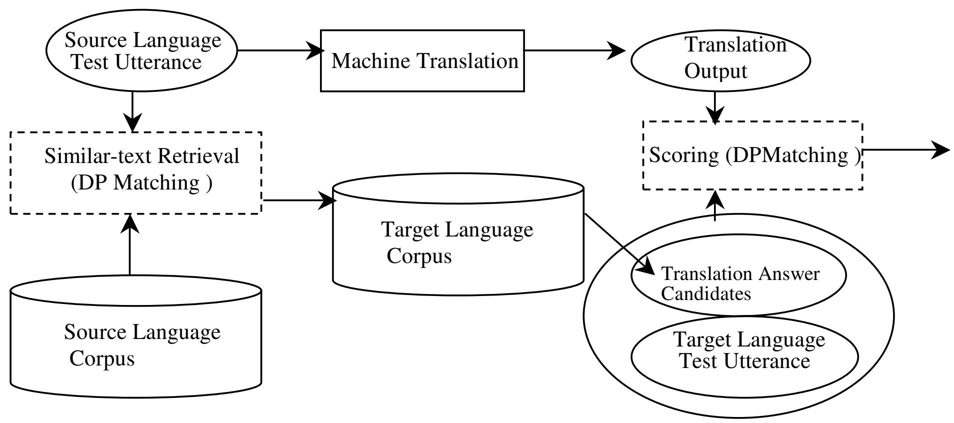

Figure 3: Diagram of the collection of a translation answer set queried from a parallel corpus

number of paraphrased answers of 15.42 utterances for each Japanese sentence, indicating that each American gives diversified utterances in the case of three-sentence generation per person. We use the term "paraphrased answer set" to refer to this answer set.

2.2.2. Collection of translation answer set queried from a parallel corpus

In this sub section, we show an automatic query-based collection method of an answer set. Figure 3 shows a diagram of the translation answer set collection method using a parallel corpus. The source language corpus and the target language corpus have a parallel translation relationship. In addition, the source language test utterance and the target language test utterance have the same relationship.

In this method, translation answer candidates are added to the target language test utterance. Each translation of each utterance in the source language corpus is regarded as a translation answer 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”. In this evaluation experiment, the ATR bilingual travel conversation database, which consists of 16110 utterances in each of the languages employed, is applied to collect the translation answer set. We use the term “queried answer set” to refer to this answer set.

3. EVALUATION EXPERIMENT

3.1. Evaluation results by the translation paired comparison method

Figure 4 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. Only a couple of score holders hit the same point: 895. The horizontal axis in Figure 4 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).

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 4 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 4, 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.

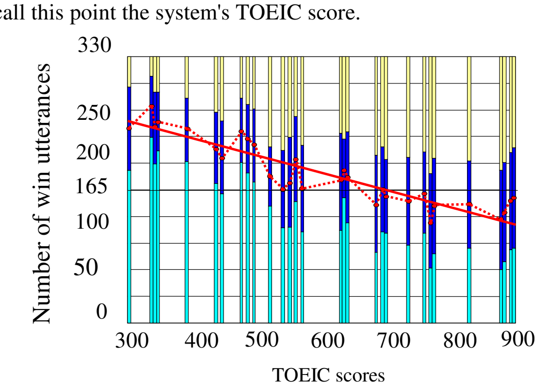

Figure4: Evaluation results for TDMT

3.2. Evaluation results by the automatic evaluation methods

3.2.1. Evaluation results by the automatic evaluation methods

Figure 5 shows the average answer set similarity for individual Japanese examinees taking the TOEIC test. The circle indicates the answer set similarity using the paraphrased answer set. The filled circle indicates the answer set similarity using the queried answer set. A Japanese person with a higher score shows a higher answer set similarity. TDMT’s average

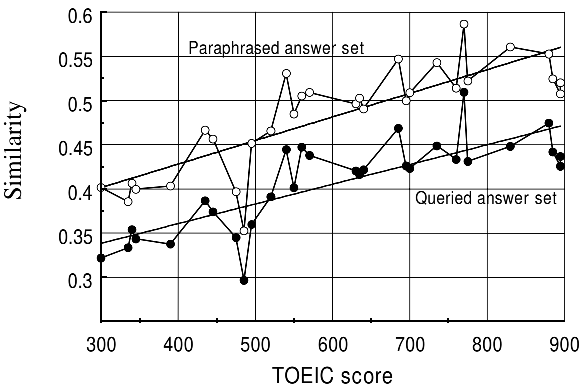

Figure 5: Answer set similarity for various Japanese people taking the TOEIC test

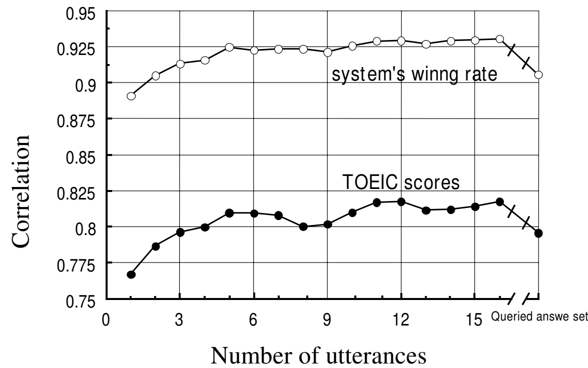

Figure 6: Relationship between correlation and utterance number used for an evaluation

answer set similarity using the paraphrased answer set is 0.52. Using the regression line in Figure 5, the TOEIC score for TDMT is calculated as 751.5. Using the queried answer set, TDMT's average answer set similarity is 0.44 and the TOEIC score is 753.6. Considering the reductions in the evaluation costs and time, this automatic scheme shows a decent performance and is very promising.

3.2.2. Effects of paraphrased utterances

To investigate the effects of paraphrased utterances, we change the utterance number in the paraphrased answer set. Figure 6 shows the correlation between the answer set similarity and the TOEIC score (indicated by the circle), and the correlation between the answer set similarity and the subjective system's winning rate normalized by the total number of test utterances (indicated by the filled circle). The horizontal axis shows the number of paraphrased answers used for the evaluation. The vertical axis shows the correlation. The correlation of the system's winning rate is higher than that of the TOEIC score. One understanding from this is that the evaluators in the translation paired comparison method judged the translation quality accordingly to some scale correlated to the DP based similarity approach.

Figure 6 also shows that the addition of sentences produces higher correlation values. The improvement by the queried answers corresponds to roughly three additional sentences.

4. CONCLUSION

We proposed two automatic methods for a speech translation system. We used the methods to evaluate the language translation subsystem of ATR-MATRIX. The results showed the system's matched points by the automatic methods to be around 750 on the TOEIC, while a subjective method gave a score around 700 points. We hope to apply the automatic methods to system improvement evaluations above this error range, which is usually found in the initial development stages. In the future, we also hope to develop a more precise method applicable to every development cycle.

5. REFERENCES

[1] Takezawa et al. (1998). A Japanese-to-English speech translation system: ATR-MATRIX. In Proceeding of ICSLP (pp. 2779--2782).

[2] Sugaya, F. et al. (1999). End-to-end evaluation in ATR-MATRIX: speech translation system between English and Japanese. In Proceedings of Eurospeech'99 (pp. 2431--2434).

[3] Wahlster, W. (2000). Verbmobil: Foundations of Speech-to-Speech Translation. Springer.

[4] Sumita, E. et al. (1999). Solutions to Problems Inherent in Spoken Language Translation: The ATR-MATRIX Approach. In Proceeding of MT Summit (pp. 229--235).

[5] Tomita, M. et al. (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).

[6] Sugaya, F. et al. (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).

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

[8] Su, K. –Y. et al. (1992). A new quantitative quality measure for machine translation systems. In Proceeding of COLING (pp. 433--439).

[9] Takezawa, T. et al. (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).