Can Indicating Translation Accuracy Encourage People to Rectify Inaccurate Translations?

Mai Miyabe[1] and Takashi Yoshino[2]

[1] Graduate School of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama, Japan
miyabe@yoslab.net
[2] Faculty of Systems Engineering, Wakayama University,
930 Sakaedani, Wakayama, Japan
yoshino@sys.wakayama-u.ac.jp

Abstract. The accuracy of machine translation affects how well people understand each other when communicating. Translation repair can improve the accuracy of translated sentences. Translation repair is typically only used when a user thinks that his/her message is inaccurate. As a result, translation accuracy suffers, because people's judgment in this regard is not always accurate. In order to solve this problem, we propose a method that provides users with an indication of the translation accuracy of their message. In this method, we measure the accuracy of translated sentences using an automatic evaluation method, providing users with three indicators: a percentage, a five-point scale, and a three-point scale. We verified how well these indicators reduce inaccurate judgments, and concluded the following: (1) the indicators did not significantly affect the inaccurate judgments of users; (2) the indication using a five-point scale obtained the highest evaluation, and that using a percentage obtained the second highest evaluation. However, in this experiment, the values we obtained from automatically evaluating translations were not always accurate. We think that incorrect automatic-evaluated values may have led to some inaccurate judgments. If we improve the accuracy of an automatic evaluation method, we believe that the indicators of translation accuracy can reduce inaccurate judgments. In addition, the percentage indicator can compensate for the shortcomings of the five-point scale. In other words, we believe that users may judge translation accuracy more easily by using a combination of these indicators.

Keywords: multilingual communication, machine translation, back translation.

1 Introduction

The Internet has increased the opportunity for multilingual communication. However, communicating in non-native language is complicated, because the language barrier hampers mutual understanding [1, 2]. Machine translation is used in order to overcome the language barrier when communicating using non-native language [3].

Despite recent advances in machine translation technology, obtaining highly accurate translations is still very difficult. The probability of inaccurate machine translations increases as messages become longer. As we have pointed out, inaccurate

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2011, LNCS 6762, pp. 368–377, 2011.

translations impede mutual understanding. Therefore, for smooth communication, users need to create messages with very few translation errors.

Translation repair plays an important role in multilingual communication when using machine translation, as it can be used to create messages with very few translation errors [4]. Translation repair is typically only performed when a user assumes that his/her message is inaccurate. As a result, translation accuracy suffers, because people's judgment in this regard is not always accurate, and many inaccurate messages are not repaired effectively [5]. Therefore, it is necessary to develop a method that helps to reduce inaccurate judgments on the part of users.

There are differences among users' judgments of translation accuracy. Inaccurate judgments may occur because of irresponsible users. We believe that people can more correctly judge the accuracy of a translation if they are provided with an indicator of the accuracy of that translation. In this study, we propose a method that indicates the accuracy of a translation in order to reduce users' inaccurate judgment. In particular, the method uses an automatic method to evaluate the accuracy of the translation, and then indicates this value to the user. This paper verifies the effect of indicating the translation accuracy using the proposed method.

2 Translation Accuracy Indication of Back-Translated Sentences

In order to indicate translation accuracy, we have to measure the accuracy of the translated sentences. In this study, we use an automatic evaluation method of translation accuracy, called Bilingual Evaluation Understudy (BLEU). This method is one of a number of automatic evaluation methods for translation accuracy. In the area of natural language processing, various studies use BLEU as an automatic evaluation method.

Many researchers have proposed different automatic evaluation methods of translation accuracy, including BLEU [6], NIST [7], and so on. These methods calculate the accuracy of a translation by comparing a translated sentence to human reference translations. Uchimoto et al. proposed a method that calculates translation accuracy by comparing a back-translated sentence with its input sentence [8]. We assume that people use back translation for creating multilingual messages. Therefore, we use the latter method here.

2.1 Accuracy Indicators

In this study, we propose the following three methods for indicating translation accuracy.

Method (A): Translation accuracy is expressed in **percentage** terms.

Method (B): Translation accuracy is expressed using the following **five-point scale**[1]: "It is correctly translated," "It is translated correctly, but is not fluent," "It is not entirely comprehensible," "It is partially translated, but does not express the meaning of the input sentence," and "It is incorrectly translated."

Method (C): Translation accuracy is expressed using the following **three-point scale**: Correct, Neutral, and Wrong.

[1] We defined the five-point scale on the basis of the adequacy evaluation method developed by Walker [9].

2.2 Evaluation Method of Translation Accuracy

In this study, we use BLEU [6] to measure translation accuracy. We derive the values for each of the three indicators from the BLEU score. The BLEU score is calculated using the following equations:

$$BLEU = BP \times \exp(\log P) \tag{1}$$

$$BP = \begin{cases} 1 & (c \geq r) \\ \exp(1 - r/c) & (c < r) \end{cases} \tag{2}$$

$$P = \frac{N_s}{c} \tag{3}$$

Here,

c is the number of words in a back-translated sentence.

r is the number of words in an input sentence.

N_s is the number of matching words between an input sentence and its back-translated sentence.

A BLEU score gives a value between 0 and 1. Therefore, using method (A), the BLEU score is provided to users. For method (B), we convert the BLEU score to values on a five-point scale using the following equation.

$$Value = 2.856 \times BLEU + 1.183 \tag{4}$$

Table 1 shows the relationship between the BLEU score, and the values of the five-point scale and the three-point scale. For method (C), we convert the values on the five-point scale to values on a three-point scale using the relationship on table 1.

Table 1. The relationship between the BLEU score, the five-point scale, and the three-point scale

BLEU score	Five-point scale	Three-point scale
$1 \leq Value < 2$	"It is incorrectly translated."	Wrong
$2 \leq Value < 3$	"It is partially translated, but does not express the meaning of the input sentence."	
$3 \leq Value < 4$	"It is not entirely comprehensible."	Neutral
$4 \leq Value < 5$	"It is translated correctly, but is not fluent."	Correct
$Value = 5$	"It is correctly translated."	

3 Experiment

We performed an experiment to verify the effect of each indicator. The subjects of the experiment were 20 university students. Their ages ranged from 18 to 24 years, with an average age of 21 years.

In this experiment, we used 40 Japanese sentences containing conversational expressions. These sentences required translation repair because their back-translated sentences were inaccurate. We decided to limit the number of letters in a sentence to between 20 and 30.

3.1 Evaluation Points

The points of evaluation were as follows:

[Point 1] Can indicating translation accuracy encourage people to rectify inaccurate translations?

[Point 2] Which method is most effective?

3.2 Experimental Condition

In order to verify the evaluation points, the experiment was performed under the following four conditions:

[Condition 1] Without a translation accuracy indicator
[Condition 2] With the translation accuracy indicator using Method (A)
[Condition 3] With the translation accuracy indicator using Method (B)
[Condition 4] With the translation accuracy indicator using Method (C)

The subjects repaired 10 translation sentences under each condition.

3.3 Experimental Tool

Figure 1 shows a screenshot of our experimental tool. When a subject inputs a sentence into the input area, its back-translated sentence is shown in the back translation area. Under experimental conditions 2, 3, and 4, the value of the calculated accuracy indicator is shown at the end of the back-translated sentence. The calculated accuracy indicator is highlighted in red.

This experimental tool used a J-Server with Language Grid [10] as the machine translation system. In this experiment, the source language for the translation was Japanese, and the target language was Chinese. In order to calculate the BLEU score, this tool carries out the Japanese morphological analysis using the morphological analyzer MeCab [11].

3.4 Experimental Procedure

The experimental procedure was as follows:

(1) Repair an input sentence to ensure that the meaning of the back-translated sentence conforms to that of the input sentence.
(2) Fix the back-translated sentence when a subject concludes that its meaning is the same as that of the input sentence.
(3) Repeat steps (1) and (2) 10 times.
(4) Repeat steps (1) to (3) for the other experimental conditions.

When the subjects were unable to repair a sentence after 5 min, we allowed them to abandon the translation repair.

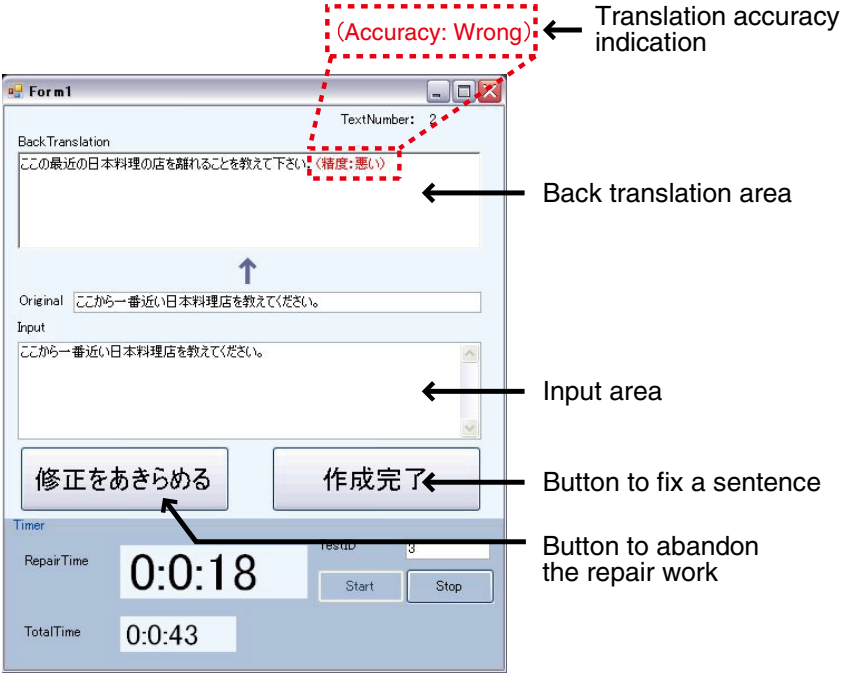

Fig. 1. Screenshot of experimental tool

Table 2. Evaluated accuracy in each experimental condition

	Unrepaired sentences	Repaired sentences			
		Condition 1 (Without indication)	Condition 2 (Method (A))	Condition 3 (Method (B))	Condition 4 (Method (C))
Average	1.6	3.2	3.2	3.2	3.3
S.D.	0.5	0.6	0.5	0.6	0.7
Significance probability	<0.001				

Table 3. Number of the sentences abandoned in the experiment

	Condition 1 (Without indication)	Condition 2 (Method (A))	Condition 3 (Method (B))	Condition 4 (Method (C))
Number of sentences abandoned	41 sentences	34 sentences	40 sentences	51 sentences
Abandonment rate	20.5 %	17.0 %	20.0 %	25.5 %

In each condition, the total number of repaired sentences was 200.

4 Results

4.1 Accuracy of Repaired Back-Translated Sentences

We evaluated the accuracy of both the unrepaired back-translated sentences and the repaired back-translated sentences by using the adequacy evaluation method developed by Walker [9]. In this method, a five-point scale is used to evaluate the translation accuracy. The evaluation method asks the question: "How much of the meaning expressed in the input message is also expressed in the back-translated message?" The method then grades the accuracy value on the following scale: 5: All, 4: Most, 3: Much, 2: Little, and 1: None. In this evaluation, three evaluators evaluated the accuracy of back-translated sentences.

Table 2 shows the evaluated accuracy in each experimental condition. As can be seen from the results of the evaluation, there was a significant difference among unrepaired sentences and repaired sentences under the four conditions. In addition, we see from the results of the multiple comparisons that there was no significant difference between unrepaired sentences and repaired sentences under each condition.

4.2 Number of Sentences Abandoned

Table 3 shows the number of the sentences abandoned in the experiment. The purpose of this study is to verify how well the indicators reduce the number of inaccurate judgments. The subjects abandoned the repair work when they had not improved the accuracy of a sentence enough after 5 minutes. Therefore, we consider that the abandoned sentences occurred as a result of the subjects' accurate judgments.

Table 4. Number of the inaccurate judgments

	Condition 1 (Without indication)	Condition 2 (Method (A))	Condition 3 (Method (B))	Condition 4 (Method (C))
Number of inaccurate judgments	51 sentences	58 sentences	49 sentences	40 sentences
Inaccurate judgment rate	25.5 %	29.0 %	24.5 %	20.0 %

In each condition, the total number of repaired sentences was 200.

Table 5. Results of questionnaire

Questions	Average (S.D.)
(1) I checked the translation accuracy when the accuracy was indicated.	4.2 (1.1)
(2) I think that the method (A) was useful for judgments of translation accuracy.	3.9 (1.2)
(3) I think that the method (B) was useful for judgments of translation accuracy.	3.9 (0.9)
(4) I think that the method (C) was useful for judgments of translation accuracy.	2.9 (0.9)

We used a five-point Likert scale for the evaluation: 1: Strongly disagree, 2: Disagree, 3: Neutral, 4:Agree, and 5: Strongly agree.

Table 6. Frequency distribution of ranking of indicators by subjects

	Method (A)	Method (B)	Method (C)
Rank 1	8	11	1
Rank 2	9	7	4
Rank 3	3	2	15
∑ar	35	31	54

∑ar is the sum of the values calculated by multiplying the frequency of appearance by the rank value.

4.3 Number of Inaccurate Judgments

In this experiment, we evaluated the accuracy of the repaired back-translated sentences using a five-point scale. If the accuracy evaluation is greater than or equal to 3, the back-translated sentence is passed on. If the value is less than 3, the back-translated sentence is not passed on. Therefore, if the accuracy of a repaired sentence is less than 3, we regard it as an inaccurate judgment. We counted the number of inaccurate judgments of 200 repaired sentences.

Table 4 shows the number of the inaccurate judgments. From table 4, we see that Condition 4 had the least number of the inaccurate judgments. Similarly, Condition 2 had the most number of inaccurate judgments.

4.4 Results of Questionnaire

Table 5 shows the result of the questionnaire. We used a five-point Likert scale for the evaluation: 1: strongly disagree, 2: disagree, 3: neutral, 4: agree, and 5: strongly agree.

Moreover, in the questionnaire, we asked subjects to rank the three methods for indicating translation accuracy. Table 6 shows the results of the ranking. The smaller the value of ∑ar in table 6, the better is the ranking. From the results of the ranking, we see that method (B) has the smallest ∑ar, and method (C) has the largest ∑ar. Therefore, according to the subjects, method (B) was evaluated as the best method for indicating translation accuracy.

4.5 The Accuracy of the Automatic Evaluation Method

In this experiment, the experimental tool indicated the accuracy of back-translated sentences calculated using the equations described in section 2.2. However, the calculated accuracy is not necessarily accurate. Therefore, we verified the correlation between a subjective evaluation and the BLEU score.

The correlation coefficient between the subjective evaluation and the BLEU score was 0.335, and the significance probability was less than 0.001. Although there was a positive correlation between the subjective evaluation and the BLEU score, it was not very high, and so there may have been a mismatch between them.

Table 7. Number of the inaccurate judgments with an incorrect indication of translation accuracy

	Condition 3 (Method (B))	Condition 4 (Method (C))
Number of inaccurate judgments	23 sentences	18 sentences

5 Discussion

5.1 Reducing the Effect of Inaccurate Judgments

In this section, we discuss evaluation point 1: "Can indicating translation accuracy encourage people to rectify inaccurate translations?"

In section 4.3, we showed that Condition 4 had the least number of the inaccurate judgments and that Condition 2 had the most inaccurate judgments. However, the difference between the highest number and the lowest number was 18. Therefore, there was no large difference between each of the conditions.

In the free description section of the questionnaire, subjects commented that "with translation accuracy indications using methods (B) and (C), I finished repair work when the experimental tool indicated 'It is not entirely comprehensible' or 'Neutral'." Recall that Method (A) provides a percentage indicator of the accuracy of the translated sentence. In this method, the translation accuracy criteria vary depending on subjects. In contrast, Methods (B) and (C) provide a scale defined clearly in words. In these latter two methods, when the subjects had repaired a sentence to a certain level of accuracy, they may have finished the repair work.

In section 4.5, we showed that a mismatch between the subjective evaluation and the BLEU score may have occurred. The incorrect BLEU score may have affected subjects' judgments.

We counted the number of inaccurate judgments that occurred when the BLEU score was incorrect. Table 7 shows the number of inaccurate judgments with an incorrect indication of the translation accuracy. From table 7, we see that approximately half of inaccurate judgments occurred when an incorrect indication was provided. We think that we can prevent the occurrence of these inaccurate judgments by improving the accuracy of the automatic evaluation method.

5.2 Appropriate Method for Accuracy Indication

In this section, we discuss evaluation point 2: "Which method is most effective?"

In section 4.4, we showed that subjects evaluated Method (B) as the best method for indicating the accuracy of a translation. From Table 6, Method (A) had the second smallest value of $\sum ar$, and the difference between the values for Methods (A) and (B) was small.

From Tables 5 and 6, Methods (A) and (B) received a high evaluation. However, in the free description section of the questionnaire, subjects provided the following comments: "Although Method (B) provides a clear scale, it is difficult to judge small differences," "Method (A) makes it is easy to check variation of accuracy, but it is difficult to set evaluation standards." From these comments, we found that a combination of Methods (A) and (B) was most effective, as each covered for the others shortcomings. Therefore, we think that a combination of these indicators may help users in judging translation accuracy more easily.

6 Conclusion

In this paper, we proposed methods to indicate the accuracy of a translation in order to encourage people to rectify inaccurate translations. We used an automatic evaluation method to measure the accuracy of the translated sentences by using the following three indicators: a percentage, a five-point scale, and a three-point scale. Moreover, we verified the effects of these indicators in reducing inaccurate judgments.

The results of the experiment were as follows.

(1) The indications of translation accuracy did not significantly affect the inaccurate judgments of users. However, the automatically evaluated values in this experiment were not always accurate. We think that the incorrect automatic-evaluated values may have led to some inaccurate judgments. If the accuracy of the automatic evaluation method improves, the translation accuracy indicators can help to reduce inaccurate judgments.

(2) The indication using a five-point scale obtained the highest evaluation and that using the percentage obtained the second highest evaluation. Further, the percentage indicator can cover for the shortcomings of the five-point scale. We believe that a combination of these indicators can help users in judging the translation accuracy more easily.

In the future, we will need to improve the accuracy of the automatic evaluation of the translation accuracy. Moreover, we will need to consider a method that reduces users' inaccurate judgment.

Acknowledgments. This work was partially supported by a Grant-in-Aid for Scientific Research (B), No. 22300044, 2010-2012.

References

1. Aiken, M.: Multilingual Communication in Electronic Meetings. ACM SIGGROUP, Bulletin 23(1), 18–19 (2002)
2. Tung, L.L., Quaddus, M.A.: Cultural differences explaining the differences in results in GSS: implications for the next decade. Decision Support Systems 33(2), 177–199 (2002)
3. Inaba, R.: Usability of Multilingual Communication Tools. In: Aykin, N. (ed.) HCII 2007. LNCS, vol. 4560, pp. 91–97. Springer, Heidelberg (2007)
4. Miyabe, M., Yoshino, T., Shigenobu, T.: Effects of Undertaking Translation Repair using Back Translation. In: Proceedings of the 2009 ACM International Workshop on Intercultural Collaboration (IWIC 2009), pp. 33–40 (2009)
5. Miyabe, M., Yoshino, T., Shigenobu, T.: Effects of Repair Support Agent for Accurate Multilingual Communication. In: Ho, T.-B., Zhou, Z.-H. (eds.) PRICAI 2008. LNCS (LNAI), vol. 5351, pp. 1022–1027. Springer, Heidelberg (2008)
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318 (2002)
7. NIST: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, Technical report, NIST (2002)
8. Uchimoto, K., Hayashida, N., Ishida, T., Isahara, H.: Automatic Rating of Machine Translatability. In: 10th Machine Translation Summit (MT Summit X), pp. 235–242 (2005)

9. Walker, K., Bamba, M., Miller, D., Ma, X., Cieri, C., Doddington, G.: Multiple-Translation Arabic (MTA) Part 1. Linguistic Data Consortium (LDC) catalog number LDC2003T18 and ISBN 1-58563-276-7
10. Ishida, T.: Language Grid: An Infrastructure for Intercultural Collaboration. In: IEEE/IPSJ Symposium on Applications and the Internet (SAINT 2006), pp. 96–100 (2006)
11. Kudo, T., Yamamoto, K., Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), pp. 230–237 (2004)