Evaluation of Rewriting Service in Language Translation Web Services Workflow

Takuro Yamaguchi
Faculty of Science and Engineering
Waseda University
Tokyo, 169–8555 Japan
Email: takuro-yamaguchi@fuji.waseda.jp

Reiko Hishiyama
Faculty of Science and Engineering
Waseda University
Tokyo, 169–8555 Japan
Email: reiko@waseda.jp

Daisuke Kitagawa
Graduate School of Informatics
Kyoto University
Kyoto, 606–8501 Japan
Email: kitagawa@ai.soc.i.kyoto-u.ac.jp

Yuu Nakajima
Department of Information Science
Toho University
Chiba, 274–8510 Japan
Email: yuu.nakajima@is.sci.toho-u.ac.jp

Rieko Inaba
Department of Computer Science
Tsuda College
Tokyo, 187–8577 Japan
Email: inaba@tsuda.ac.jp

Donghui Lin
Graduate School of Informatics
Kyoto University
Kyoto, 606–8501 Japan
Email: lindh@i.kyoto-u.ac.jp

***Abstract*—We discuss here the complementarity effect of rewriting services in a language-translation Web service workflow. The communication mediated by the machine translation service includes mistranslations and changed meanings, which are caused by the quality of the machine translation. To reduce these problems, we propose service reallocation and assemblage of the rewriting service by humans at each stage of the workflow. We set up two rewriting service allocation patterns: 1) a reallocation pattern as a previous rewriting process, and 2) a reallocation pattern as a follow-on rewriting process, for workflows consisting only of machine translations. A team of human judges provide multiple assessments of adequacy and fluency of sample sentences that are translated from English to Japanese using each pattern. Results indicated that the Japanese rewriting task as a follow-on rewriting service provided greater fluency than an English rewriting task as a previous rewriting service, with nearly equal adequacy.**

***Keywords*-intercultural collaboration, machine translation, language grid, web service**

I. Introduction

The opportunities to use multilingual communication over the Internet have increased greatly in recent years due to the popularization of the Internet. To communicate smoothly, we use machine translation (MT) because it can help to achieve their native communication. One example is the YMC (Youth Mediated Communication)-Viet Project [2] that is enforced by the NPO PANGAEA [1].

In the YMC-Viet Project, Vietnamese children and Japanese agricultural specialists share knowledge about rice crops over the Internet. This communication is achieved by combining a Japanese-English MT service and an English-Vietnamese MT service. We call these translation services the "language translation Web service workflow" in this study. Although this workflow allows knowledge to be shared, mistranslations and loss of meaning occur in the MT workflow. In other words, the adequacy and fluency of the input written in Vietnamese are lost through the MT. Therefore, the knowledge communication through this system is not always smooth.

Another problem that arises in the YMC-Viet Project is that the Vietnamese children and the agricultural experts cannot achieve satisfactory interactive communication because the children must choose from fixed phrases when they send information to the experts, whereas the experts can write free text. In the near future, these children will need to explain their agricultural conditions to the experts when the communication becomes active in the YMC-Viet Project. Consequently, in this study, we assume a communication flow in which the Vietnamese children can send information to the experts using free text. To enable this flow, we introduce a bridger, which is a person who provides a rewriting or translation repair service in the language translation Web service workflow. In this paper, we analyze how effective the use of a Japanese native bridger is if he/she is engaged in the workflow. We analyzed the effectiveness in terms of cost and the quality of translation results.

II. Related Works

Language Grid[4] is a service-oriented intelligence platform for language services. End-users can combine many language services that are provided around the world by Language Grid. A user who does not have expert knowledge can combine language services more easily by using the "Multilingual Studio[5]" which is an API (application programming interface) library of various language services that Language Grid provides. Linking these language resources enables the user to achieve advanced knowledge translation communication such as in the YMC-Viet Project (Fig. 1). Kita [3] analyzed the YMC-Viet project as a case study to improve the Language-Communication Model. She finally introduced two bridgers, who are bilingual people that provide translation to achieve the Language-Communication Model. One bridger was a Vietnamese person who was

978-0-7695-5047-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CultureComputing.2013.12

bilingual in English and Vietnamese. The other bridger was Japanese and was bilingual in English and Japanese. Finally she succeeded to obtain high quality sentences because the sentences are translated by hand and the MT services are not used in her Language-Communication Model. But it was difficult to provide a knowledge translation service between Vietnamese children and Japanese experts for long periods because the costs of the bilingual bridgers were very high.

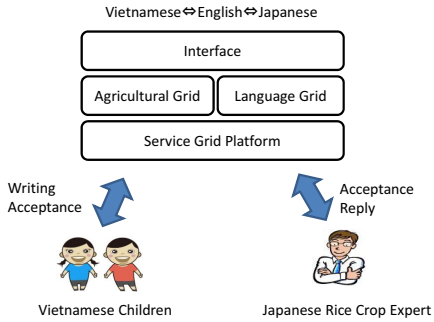

Figure 1. YMC-Viet Project[2]

III. Approach

In applying the language translation Web service workflow, it is important to discuss how to maintain the quality of translation while giving due consideration to the rewriting cost. In this study, we propose a service reallocation and assemblage of the rewriting service by humans at each stage of the workflow. Fig. 2 shows the language translation Web service workflow that we applied. The figure shows a workflow in which a Japanese bridger can repair sentences translated from Vietnamese to English or sentences translated from English to Japanese. In the general workflow, the Japanese bridger repairs the sentences that are written in English using a back translation service. However, English is not the first language for the Japanese bridger, so we can forecast a very high rewriting cost when a Japanese person repairs the input language. To reduce these problems, we propose a service reallocation and assemblage of the rewriting service by humans at each stage of the workflow. We set up two rewriting service allocation patterns: 1) a reallocation pattern as a prior rewriting process, and 2) a reallocation pattern as a follow-on rewriting process, for workflows consisting only of MTs. We prepared two experimental application interface (Fig. 4) to carry out these experiments. One was a translation 'repair' experimental application interface. The other was a 'rewriting' experimental application interface.'Repair Service' enable human to have an updated the sentence by the help of the back-translation, which is to translate text in English to Japanese as another language and then retranslate the result back to English. On the other hand, 'Rewrite Service' enable human to have an updated the sentence without back-translation function.

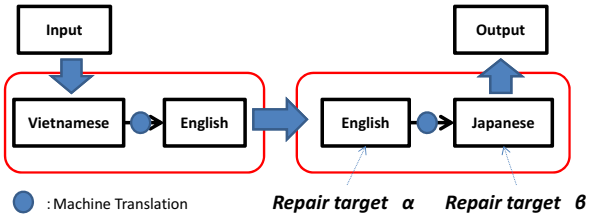

Figure 2. Knowledge communication flow

IV. Experiment

A. Language translation Web service workflow

In this study, we compared the complementarity effect of rewriting services in a language-translation Web service workflow. For comparison, we used two service patterns:
1) a reallocation pattern as a prior rewriting process, and
2) a reallocation pattern as a follow-on rewriting process, for workflows consisting only of MTs (Fig. 3).
In each service, Japanese college students serve as a human service. Fig. 3 shows the experimental patterns. We invited three participants for experiment pattern A and nine participants for experiment pattern B. In this study, we defined the services as follows: the reallocation pattern as a follow-on rewriting process: **Rewriting Service (experiment pattern A)**, the reallocation pattern as a prior rewriting process: **Translation Repair Service (experiment pattern B)**

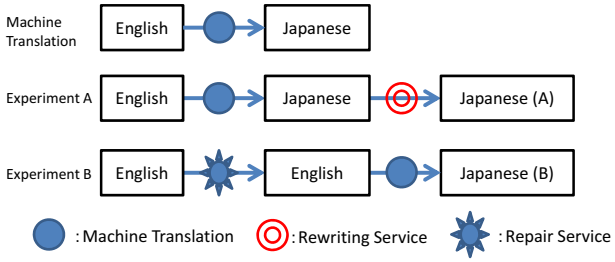

Figure 3. Experimental flow

B. Sample sentences

We chose 96 sentences as sample text. These sentences consisted of parallel texts actually used in the YMC-Viet Project. We chose the Vietnamese versions of them as the input for the MT workflow in order to obtain the English text, which then became the input text for the English-Japanese MT (Fig. 2).

C. Service procedures

The objective of the experiment A service was to rewrite the Japanese sentences obtained directly from the MT workflow. The participants guess the meaning from the context of the sentence and rewrite it better one. The objective of the experiment B service was to repair the input sentences

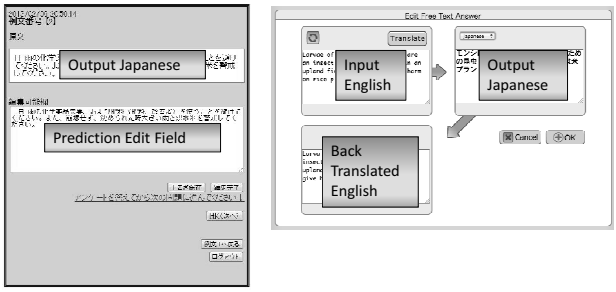

Figure 4. Experimental application interface for rewriting service (left figure) and for repair service (right figure).

of the MT workflow. Subjects can see the translation results Japanese sentence, the input English sentence, and the back-translation English sentence on the device.

V. Results and Analysis

A. Evaluation experiment

A team of human judges provided multiple assessments of adequacy, fluency [6], which were used by Lin et al. [7], and working cost (number of times sentences were rewritten and working time. Each cost was the average of the inclusive sum.) of sample sentences that were translated from English to Japanese using each pattern. The judges were Japanese college students. The judges evaluated the sentences based on the following standard.

Adequacy

How much of the translated text is transmitted in comparison with the parallel text (gold-standard).
5: All 4: Most 3: Much 2: Little 1: None

Fluency

How accurate the grammar of the translated text is.
5: Flawless 4: Good 3: Non-native level 2: Disfluent 1: Incomprehensible

We asked them to evaluate each sentence within 30 seconds. Finally, we standardized the variance of the evaluated value for all sentences and did right-tailed test by a significance level of 10%. Therefore, we excluded the sentences that had statistical variability.

B. Quality of sentences from workflows consisting only of MTs

The quality of sentences from the workflows consisting only of MTs are listed in Table I. The results show that fluency received 1.72 pt and adequacy received 1.98 pt. This is lower than the results of the communication model that Kita [3] proposed (fluency 4.55 pt, adequacy 3.80 pt). This result indicates that this workflow cannot output high quality sentences. We predict that the quality of the input English was low because the input English was already the output of Vietnamese-English MT.

Table I
Quality of sentences consisting only of machine translations (pt)

	Machine Translation
Fluency	1.72
Adequacy	1.98

Machine Translation Results

Fluency Judges: 18, Sentences: 89, Total Points of Evaluation: 2,756
Adequacy Judges: 18, Sentences: 91, Total Points of Evaluation: 3,250

C. Quality of experiment A and experiment B

Table II lists the compared quality of Japanese A and Japanese B. Compared with Table II, each service was effective in raising the quality of the output sentences. Based on this result, Japanese A's fluency was 3.52 pt, which is higher than Japanese B's. This is because the Japanese students were rewriting Japanese as their first language in experiment A. The adequacy results were about even. An adequacy result of 3 pt is 3: Much. Thus, 2.77 pt in experiment A and 2.76 pt in experiment B are both values close to 3 pt. This result shows that the rewriting service is generally effective for translation repair.

Table II
Comparison of sentence quality (pt)

	Experiment A	Experiment B
Fluency	3.52	2.83
Adequacy	2.77	2.76

Experiment A

Fluency Judges: 20, Sentences: 92, Total Points of Evaluation: 6,468
Adequacy Judges: 20, Sentences: 91, Total Points of Evaluation: 5,041

Experiment B

Fluency Judges: 20, Sentences: 91, Total Points of Evaluation: 5,145
Adequacy Judges: 20, Sentences: 92, Total Points of Evaluation: 5,082

Sample sentence

We found that the rewriting service by humans was more effective than the translation repair service by humans. Figure 5 lists the sample sentences that we obtained in experiments A and B. The Japanese output of MT in Figure 5 corresponds to repair target β in Fig. 2. The sentences translated into English by MT (Figure 5) correspond to repair target α in Fig. 2. The evaluation results for these sample sentences (fluency, adequacy) were (1.94, 2.39), (4.35, 4.45), and (2.70, 2.80). As described above, there were many sentences where the quality of Japanese A was higher than Japanese B. This indicates the usefulness of the rewriting service.

Example 21	Japanese Output	Translated sentence in English	(Fluency, Adequacy)
Machine Translation	いくつかの稲種が抵抗力があります ある一定の病気（病気、病気）。病気（病気、病気）のために耐える種稲の植え付けは、防止する測定です。	Some rice plant species are resistant some certain disease (sickness, malady). The planting of species rice plant that resist for disease (sickness, malady), is a preventive measure.	(1.94, 2.39)
Experiment A	いくつかの種類の稲は病気に対し、抵抗力があります。抵抗力のある稲の植え付けは、病気防止に役立ちます。	There are some kind of rice resistance to disease. Planting of rice that is resistant is useful in disease prevention.	(4.35, 4.45)
Experiment B	稲のいくつかの種がある一定の病気に抵抗力があります（病気、病気）。病気（病気、病気）に耐える種稲の植え付けは、防止する対策です。	Some species of rice plant are resistant to some certain diseases (sickness, malady) . The planting of the species rice plant that resist disease (sickness, malady) is preventive measures .	(2.70, 2.80)
Parallel Text (Collective data set)	特定の病気に対して抵抗力を持っているイネが品種もあります。抵抗力を持ったイネを育てることも病気の防除に繋がります。	There is a rice plant breed strong against a particular disease. Growing resistant rice plants also lead to the prevention of disease.	(5,5)

Figure 5. Example: Quality result of experiment A sentence is higher than that for the results of experiments B

Table III
COMPARISON OF REWRITING COST

	Exp A	Exp B
Work Time (Seconds)	162.2	473.3
Rewrites (Number of Times)	1.2	9.5

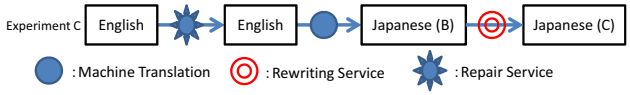

Figure 6. Experiment C flow

Table IV
COMPARISON OF QUALITY OF ALL EXPERIMENTS

	MT	Exp A	Exp B	Exp C
Fluency	1.72	3.52	2.83	4.13
Adequacy	1.98	2.77	2.76	3.51

D. Rewriting cost

Table III compares the rewriting cost of experiments A and B. As indicated in the Table, the amount of time and the number of rewrites in experiment A were smaller than in experiment B. In this study, repair services and rewriting services were all in Japanese, so we found that the costs of experiment A were lower than those of experiment B.

E. Experiment C

After obtaining the results of experiment A and experiment B, We conducted experiment C as an extension of experiment B. This experimental flow is shown in Fig. 6. This flow involved allocating a rewriting service for Japanese B that was obtained in experiment B. As a result, there was a prior rewriting process and a follow-on rewriting process for workflows consisting only of MTs. The aim of this experiment was to obtain sentences with higher quality than those obtained in experiments A and B. In addition, we designed the workflow without the people who have a translation technology, in other words, the workflow is conducted only Japanese bridger. The participants in experiment C were three Japanese college students, who were responsible for rewriting the results of experiment B. The evaluation results for sentences obtained in experiment C are given in Table IV.

Experiment C

Fluency Judges: 18, Sentences: 88, Total Points of Evaluation: 6,536

Adequacy Judges: 18, Sentences: 91, Total Points of Evaluation: 5,751

1) Quality of Japanese C: It is evident that the results of experiment C had higher quality than compared with those of Japanese A and Japanese B. The evaluation results for fluency were 4.13 pt out of 5 pt. A fluency result of 4 pt is good on the scale mentioned previously (4: Good). This shows that it was possible to obtain very fluent Japanese. In addition, the result of adequacy was 3.51 pt, where 3 pt for adequacy is 3: Much. This means that the sentence of Japanese C retains much of the original meaning. From the above results, we confirmed that is possible to achieve knowledge communication that retains the appropriate meaning and has a high level of fluency.

Sample sentences

Figure 7 indicates that the sample sentence in experiment C obtained a high evaluation. The Japanese output of the MT (Figure 7) corresponds to repair target β in Fig. 2. The translated sentence in English of the MT listed in Figure 7 corresponds to repair target α in Fig. 2. The quality of the text of example No. 62 in Figure 7 was (1.39, 1.94), (2.40,

Example 62	Japanese Output	Translated sentence in English	(Fluency, Adequacy)
Machine Translation	米多様性に依存している稲、持っている種々の特徴(品種、種) : よい成長、風味、高い生産高	Rice plant have different characteristics depending on the rice variety (breed, seed) : good growth , flavor, high yield.	(1.39, 1.94)
Experiment A	種々の特徴(品種、種)は、米や稲の多様性に依存している:よい成長、風味、高い生産高	Characteristics of species are dependent on variety of rice : for example, enough growth, flavor, high production.	(2.40, 2.65)
Experiment B	稲、持っている種々の特徴。それらの特徴は米の多様性に依存します(品種、種) : 例えば、十分な成長、風味、高い生産高	Rice plant have different characteristics. Those characteristics depend on variety of rice (breed, seed) : for example, enough growth , flavor, high yield.	(2.70, 3.20)
Experiment C	個々の稲には様々な特徴があり、それは米のできに大きな影響を与えます。例えば、成長のしやすさ、風味のよしあし、高い生産高などが挙げられます。	Rice of individual has various characteristics and characteristics affect quality of rice. For example, ease of growth, the good and bad of rice, high production.	(4.61, 4.11)
Parallel Text (Collective data set)	稲の品種選びですが、丈夫なもの、味の良いもの、量がたくさんとれるものなど特徴があります。	Rice has different characteristics depending on its breed variety: some are strong, others have good taste, still others promise yield volume.	(5, 5)

Figure 7. Example: Quality result of experiment C sentence is higher than that for the results of experiments A and B

Table V
COST COMPARISON

	Exp A	Exp B	Exp C	Entire C
Work Time (Seconds)	162.2	473.3	133.4	606.6
Rewrites (Number of Times)	1.2	9.5	1.2	10.7

Table VI
COMPARISON OF SELF-EVALUATION AND OTHERS' EVALUATION FOR EXPERIMENT A

	Self-evaluation	Others' Evaluation
Fluency	4.64	3.52
Adequacy	3.88	2.77

Table VII
COMPARISON OF SELF-EVALUATION AND OTHERS' EVALUATION IN EXPERIMENT C

	Self-evaluation	Others' Evaluation
Fluency	4.32	4.13
Adequacy	4.03	3.51

2.65), (2.70, 3.20), and (4.61, 4.11) for machine translation, experiment A, experiment B, and experiment C. The text obtained in experiment C had higher quality than the text obtained in experiments A and B.

2) Rewriting cost: Table V compares the working costs of all experiments. For experiment C, the cost is indicated for just the follow-on rewriting service of the experiment C flow, as well as the cost of the entire C flow involving the prior rewriting process and the follow-on rewriting process for workflows consisting only of MTs.

Table V indicates that the average working time of experiment C was 133.4 seconds. It seems that the working time of experiment C is shorter by 28.8 seconds than for experiment A at 162.2 seconds. We consider that this result is because the sentences used in experiment C have already been repaired in experiment B. The participants rewrote the Japanese B, which was repaired in experiment B, in Fig. 6. It is easier for participants to rewrite the Japanese B because the quality of Japanese B is higher than repair target β in Fig. 2. Table V also indicates that the working time of the entire C flow was 606.6 seconds. This is much higher than the results of the other experiments. This result demonstrates that allocating a prior rewriting process and a follow-on rewriting process for workflows consisting only of MTs can produce sentences that have high fluency and adequacy, but it requires a high cost. We found that there is a trade-off between high quality translation and the working cost.

F. Self-evaluation

In this experiment, participants evaluated Japanese sentences they had rewritten themselves. The points to be evaluated were the *quality of output Japanese sentences (fluency, adequacy)* and *quality of rewritten Japanese sentences (fluency, adequacy)*. The results of evaluation are in Table VI for experiment A. The average self-evaluations for the participants in experiment A were (fluency, adequacy) = (4.64, 3.88). By contrast, the results of the third-party evaluation for experiment A were (fluency, adequacy) = (3.52, 2.77). These results show that it is possible that knowledge communication is not performed as intended by the participants in the rewriting service.

In the same way, participants in experiment C evaluated Japanese sentences they had rewritten. The results of evalua-

tion are given in Table VII. The average self-evaluations for the participants in experiment C were (fluency, adequacy) = (4.32, 4.03). On the other hand, the third-party evaluation results for experiment C were (fluency, adequacy) = (4.13, 3.51). Unlike the evaluation results of experiment A, the self-evaluation values were close to the evaluation values by others. We think that the participants in experiment C can predict sentences easily because the sentences used in experiment C were repaired once in experiment B. Consequently, the knowledge communication might have occurred as the participants intended.

G. Questionnaire

We then gave a questionnaire on paper to the participants. The content consisted of the question, "Would you be able to tolerate experiment A or experiment B if you were reallocated as a rewriting service or a translation repair service to raise the quality of knowledge communication?". Opinions obtained from the participants were as follows:

Experiment A

- Rewriting Japanese would be acceptable because I am Japanese.
- The rewriting work is troublesome, but it is more acceptable in comparison with experiment B.
- It is more comfortable to rewrite the Japanese A than to rewrite the English.
- There was a variation in quality of the sentences to be rewritten.
- The results might be good if I repair the English because I know English grammar to some extent.

Experiment B

- We have to read three types of sentences (input sentence, translation result, and back translation result), so I would not like to do this.
- I would not like to do this because English is not my native language.
- I would not like to repair the English even though the Japanese is visible.
- I would not like to do this because I do not have confidence in my English ability.

Thus, we obtained many opinions indicating that participants would prefer to carry out experiment A than experiment B. However, there were few opinions indicating that the participants would carry out experiment B because the quality of some of the output Japanese sentences was very low, and they could therefore not predict the meanings of the input sentences even though each human rewrite/repair service was provided by Japanese people. We found some interesting opinions. For example, "The results might be good if I repair the English because I know English grammar to some extent." It is necessary to do more study to determine whether we can obtain high quality text and reduce the burden of the work.

VI. Conclusion

We focused on multi-language communication by a language translation Web service workflow represented by the YMC-Viet Project. We analyzed how Japanese bridger should intervene in the flow to raise the translation quality. The results of the experiment indicated that by allocating a rewriting service for the output Japanese, we were able to get a fluent Japanese result, in contrast to the case where we allocated a translation repair service for the input English. In addition, the rewriting service made it possible to significantly reduce the working costs. Further, the rewriting service achieved sentences that retained the original meaning, which was comparable to the results for sentences achieved using a repair service. As a future challenge, we will conduct more experiments to validate the results of this study in the real communication field.

Acknowledgment

This work was supported by the Service Science, Solutions and Foundation Integrated Research Program in the Research Institute of Science and Technology for Society, Japan Science and Technology Agency (JST/RISTEX).

References

[1] NPO PANGAEA :
URL : http://www.pangaean.org/web/english/general/generaltop_en.html, Finally accessed April 2, 2013

[2] YMC-Viet Project, Ministry of Agriculture and Rural Development in Vietnam,
URL : http://www.agroviet.gov.vn/en/Pages/news_detail.aspx?NewsId=923, Finally accessed April 2, 2013

[3] Kaori Kita, Toshiyuki Takasaki, Donghui Lin, Yuu Nakajima, Toru Ishida : Case Study on Analyzing Multi-Language Knowledge Communication, *Culture and Computing 2012*, 2012.

[4] Toru Ishida : Language Grid: An Infrastructure for Intercultural Collaboration, *IEEE/IPSJ Symposium on Applications and the Internet(SAINT-06)*, pp.96-100, keynote address, 2006.

[5] Multilingual Studio, URL : http://langrid.org/developer/en/, Finally accessed April 2, 2013

[6] Linguistic Annotation Specification : Assessment of Fluency and Adequacy in Translations, URL : http://projects.ldc.upenn.edu/TIDES/Translation/TransAssess04.pdf, Finally accessed April 2, 2013

[7] Donghui Lin, Yoshiaki Murakami, Toru Ishida, Yohei Murakami, Masahiro Tanaka : Composing Human and Machine Translation Services : Language Grid for Improving Localization Processes, *Proceedings of the International Conference on Language Resources and Evaluation*(LREC 2010), pp.17-23, 2010.