Evaluation Experiment for Reading Comprehension of Machine Translation Outputs

Masaru Fuji
Document Processing Lab
Fujitsu Laboratories Ltd.

Abstract

This paper proposes evaluation methods for reading comprehension of English to Japanese translation outputs. The methods were designed not only to evaluate the performance of current systems, but to evaluate the performance of future systems had the current problems been solved.

The experiments have shown that the proposed methods are capable of producing results that are statistically significant, and that improvement in certain linguistic aspects will result in significant improvement in the comprehension level.

1 Introduction

In recent years, machine translation (MT) has created a large market in Japan. Currently, the mam users of commercial MT systems are those who wish to browse through information expressed in a foreign language, and NOT professional translators whose task is to produce complete translation documents. Those users who wish to seek information written in a foreign language will be interested to know *how much information can be retrieved from the translated text and how easily it is retrieved.*

Under these circumstances, the Japanese MT providers are interested to find out. on what sort of linguistic aspects they need to concentrate in their development processes, in order to improve browsability. They are also interested to know how much improvement can be expected from the development effort they put in. This is exactly the motivation behind our research.

1.1 Related works

We have classified traditional MT evaluation methods into the following three categories. Since our target is to devise evaluation methods for prospective end users, we have adopted the last category - usability test - for our experiments.

- *Evaluation using test sets*
  Test set is a set of sample sentences with their ideal translation sentences [3]. The sample sentences are fed into the MT system and the output compared to the ideal translations. This is probably the most popular test for MT developers to monitor their system development, though the test sets are usually used within their private development environment and much of the results are unpublished to academic conferences. It offers systematic evaluation and gives measurable evaluation results. The results are unlikely to represent the comprehension level of end users.

- *Inspection test by trained evaluators*
  MT output sentences are judged by linguists or MT developers and scores given, such as "good", "poor" , etc. This is a quick and easy way of estimating the translation quality produced by the MT system. Some researchers have pointed out that, in inspection test, there is a number of independent measurable aspects such as informativeness, comprehensiveness, etc. However, the result may not exactly correspond to the impression received by general users.

- *Usability test*
  The target MT users are used as the subjects of evaluation. The results obtained closely correspond to the general users' impression. However, it is necessary to use a large number of subjects in order to obtain statistically significant results. Concerning this test, there are several different measurable aspects, such as informativeness, comprehensiveness, and fluency.

Works by Tomita [1][2] are former works classified as usability tests. His experiments makes use of a test of English, namely, the Test of English as a Foreign Language (TOEFL), in order to compare various MT systems currently available. The reading comprehension tests of TOEFL are translated into Japanese by three different MT systems, and 70 native Japanese students are asked to take the test.

This experiment has proved effective for evaluating currently existing MT systems. Moreover, his exper-

iment only dealt with informativeness. and did not take into account other aspects of usability.

1.2 Object

The object of our experiment is to design evaluation methods that give suggestions as to what kinds of improvement should be made to the MT system in order to attain the translation quality required by the end users. It is not our intention just to measure the current performance of MT systems.

The results obtained will be highly dependent on the language pair, the MT system, the texts used, and the subject sets. However, the evaluation method itself is quite transportable, and can be applied to any language pairs, MT systems, texts, etc.

2 Experiment procedures

We have decided to employ usability test, as this is the most relevant test for dealing with end users' requirements. Out of the many measurable aspects of usability test, we carried out evaluation experiments for "informativeness". "comprehensiveness", and "fluency". Then the experimental results were compared in order to see which of these methods are suitable for finding clues for development.

In order to achieve our final object, such as to find clues for development, we have applied the above mentioned experiments to successively enriched MT outputs. We then observed the change in the comprehension level, as these enrichments were successively made. Our hypothesis was that this change in the comprehension level would represent the change in the comprehension level had the system been improved in the future.

2.1 Successively enriched MT outputs

In preparing MT outputs for the experiments, we set the raw MT output as the starting point, and we successively made manual correction to this raw output, in order to obtain various versions of the original text. We set the human-written ideal translation to be the goal. The following table shows the list of texts used in the experiments.

Table 1: Successively enriched MT outputs

output 1	raw MT output
output 2	output 1 with words and noun compounds corrected
output 3	output 2 with other compounds corrected
output 4	output 3 with attachments corrected
output 5	output 4 with temporal and referential modifiers corrected
output 6	output 5 with other contextual aspects improved
output 7	human-translated, ideal translation output

It should be noted, that the earlier corrections (e.g. correction of words and noun compounds) are those that are relatively easy to incorporate into the development of MT systems, while the later corrections (e.g. correction of temporal and referential modifiers) are those that are relatively more difficult to incorporate.

The corrections are made in this order for practical reasons. For example, it is not practical to correct attachments without correcting noun compounds before hand.

2.2 Texts used for evaluation

For the use of the experiment, several texts are taken from one of the official examinations of English as a foreign language (“Eigo Kentei” , Jun-ikkyu). designed for Japanese examinees. Each text is accompanied by a set of multiple choice questions, designed to test the comprehension of the texts. The texts are originally taken from English newspaper articles and magazines, on the subjects of politics, medical science, etc. The original multiple choice questions, as well as original texts, are all written in English.

This English test is adopted to evaluate an English-Japanese MT system. The English texts are translated into Japanese using MT, and the successive enrichments are applied to the translated Japanese texts. The accompanying multiple choice questions are manually translated into natural Japanese.

The text corresponding to the results shown in the following section, is taken from a newspaper article titled “Coffee and heart attacks”. The original English text consists of approx. 340 words in 17 lines. The human-translated Japanese version of this text contains approx. 1600 characters. The text is accompanied by five multiple choice questions, each of which contains four choices. An example of such a multiple choice question is as follows.

Why was this study undertaken?

1. Because heart disease is predominant.
2. Because lifestyles increase the risk of heart disease.
3. Because previous studies were inconclusive.
4. Because coffee drinkers needed new evidence.

2.3 Experiment material prepared for each evaluation method

Evaluation of informativeness

This experiment makes use of the set of enriched MT outputs (output 1 to output 7), as well as the sets of multiple choice questions associated with the original English texts. The multiple choice questions are manually translated into Japanese. Each multiple choice question contains four choices. An average of five questions are set for each text.

For each text, the level of informativeness is calculated to be the ratio of the number of correct answers to the total number of questions, expressed in percent.

Evaluation of comprehensiveness

This experiment only makes use of the set of enriched MT outputs (output 1 to output 7). The identical set is used for all the three tests for the sake of comparison. The multiple choice questions used for informativeness are not used in this experiment.

Each MT output is printed on an answer sheet. Having read through the text, the subject is asked to intuitively judge the comprehensiveness of the text by choosing one from the four choices, “"very comprehensive"', “"slightly comprehensive”, “"slightly incomprehensive” and “very incomprehensive”. An even number of choices are given, in order to avoid the convergence of responses on the middle choice.

Evaluation of fluency

The preparation of this evaluation is quite similar to that for comprehensiveness. It makes use of the sets of MT outputs. The choices given to the subjects for this test are “very fluent”, “slightly fluent"', “slightly clumsy”, and “very clumsy”.

2.4 Test subjects

The subjects who took part in the experiments are approximately 500 Japanese university students. Each student is asked to write down their names and other details on the answer sheet, so as to avoid irresponsible answers. Those sheets with all the answers marked on the same choice (e.g. all the answers for comprehensiveness marked as "incomprehensible") are judged unreliable and are not taken into account in later statistical analysis.

The students are divided into seven groups, corresponding to the seven enriched MT outputs. No subject is asked to read more than one output. It is important to ensure this, since the subject quickly grasps the meaning of the entire text having read through the text for the first time, and this would influence the result for the second text.

2.5 Environment for experiments

All the experiments were conducted in lecture rooms, under the supervision of university staff, who were given instructions for conducting experiments. The experiments were carried out in limited time span, in order to ensure uniform quality.

In the case of informativeness, we have allowed 8 minutes to read through the text and answer the questions. In the cases of comprehensibility and fluency, this time limit was not crucial, though we set a limit of 5 minutes.

3 Results

3.1 Evaluation of informativeness

There are five multiple choice questions associated with this comprehension text, and a score of 100% is given when the subject correctly answers all of the five questions, while 0% is given if he/she answers them all wrongly. In Figure 1, the Y-axis shows the average score over the test subjects, for the given comprehension test.

The level of informativeness for output 1 and output 2 lies below approx. 60%, while that for output 3 through output 7 lies above approx. 75%,.

It is observed that, in the case of informativeness, a statistically significant improvement solely arises, when moving from output 2 to output 3 in the enrichment phase. No statistically significant improvement is observed elsewhere. This result holds true for all the texts used in the experiments.

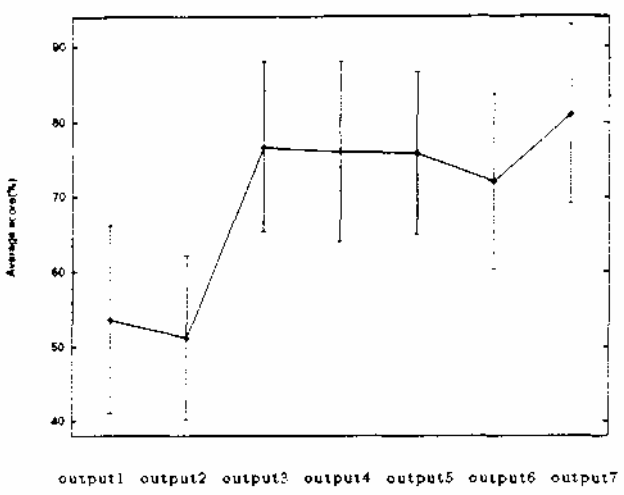

Figure 1: Results for informativeness

3.2 Evaluation of comprehensiveness

Score "1" is given to the answer of "very incomprehensive" and "4" is given to the answer of "very comprehensive".

The level of comprehensiveness for output 1 lies around 2.1, while that for output 7 lies around 3.0. A gradual improvement is observed over the entire course of enrichment.

Unlike the case of informativeness. no statistically significant improvement is observed at any unique point in the course of enrichment. However, a significant improvement is observed over the entire enrichment phase. It is also observed that the contextual enrichment is more effective compared to the case of informativeness.

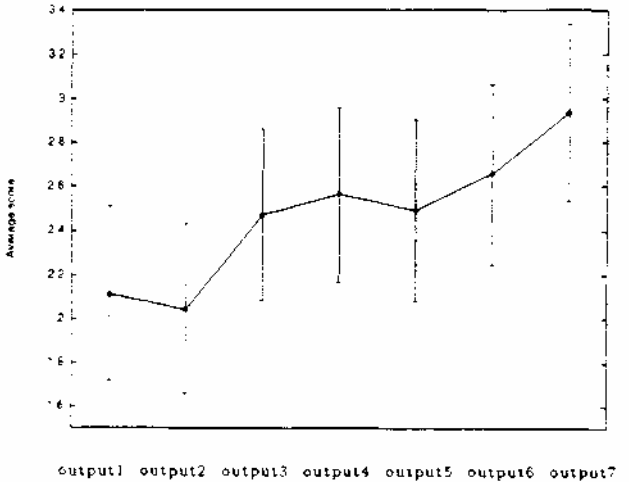

Figure 2: Results for comprehensiveness

3.3 Evaluation of fluency

Score "1" is given to the answer of "very clumsy" and "4" is given to the answer of "very fluent".

The level of comprehension for output 1 lies around 2.0, while that for output 7 lies around 3.0. Like the case of informativeness, a marked improvement arises between output 2 and output 3. although the improvement is not large enough to be statistically significant, in the case of fluency. Like the case of comprehensiveness, the contextual contribution seems to be large.

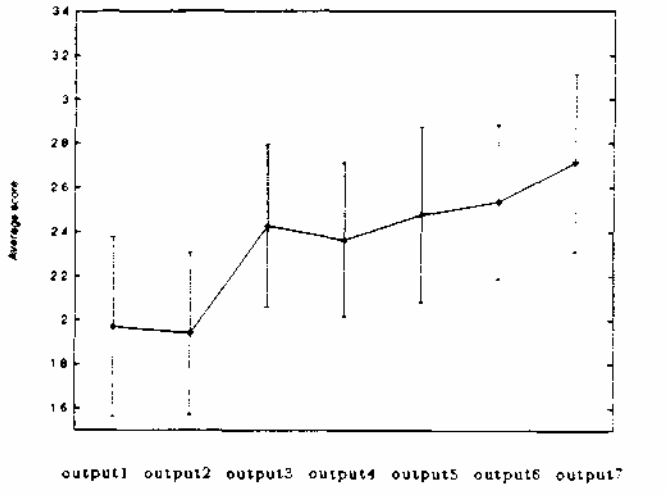

Figure 3: Results for fluency

3.4 Subjects' impression on the experiments

Having participated in the experiments, some of the test subjects have had the impression that comprehensiveness and fluency are closely related. They feel that incomprehensive texts cannot be fluent.

4 Conclusions

As stated above, the results we have obtained are highly dependent on the language pair, MT system used, test subjects, text domain, etc., and therefore they cannot be over generalised. However, the results give some indications as to the current status of English to Japanese MT systems.

For the given experimental environment, we have succeeded in obtaining clues for future development.

4.1 Validity of the evaluation experiment

The evaluation experiment using enriched outputs, prove? to produce statistically significant results, for informativeness, comprehensiveness, and fluency.

4.2 Distribution of improvements

In the evaluation of informativeness. *a* statistically significant improvement is observed solely at the point,

where all the non-noun compounds are corrected, having corrected all the words and noun compounds. On the other hand, in the cases of comprehensiveness and fluency, a significant improvement occurred over the entire enrichment phase.

There seems to be a difference between how the test subject feels comprehensive, and how much he/she actually comprehends. This is depicted by the fact that for output 3 through output 5, the many of the subjects feel that the outputs are incomprehensive and unnatural, but he can obtain considerable amount of information from these outputs.

4.3 Choice of evaluation method

Since evaluation of comprehensiveness and that of fluency do not produce significant improvement until all the enrichments are made, they are not suitable for measuring current MT systems, most of which do not have adequate capability to deal with contextual information. The tests for comprehensiveness and fluency are too sensitive to the contextual aspects.

On the other hand, evaluation for informativeness can be used for current system development as significant improvement arises at earlier enrichment phases.

4.4 Current status of English-Japanese MT systems

The only enrichment facility commonly available on practical MT systems is the registration of words and noun compounds. It is seen from the three different evaluation methods, that no statistically significant improvement is observed when this first enrichment alone is made.

Looking from a different point of view, a significant improvement can be expected if non-noun compounds are adequately registered, on the condition that noun compounds are adequately registered. Some of the newly published versions of English to Japanese MT systems are able to handle registration of such compounds.

References

[1] Masaru Tomita et al., “Evaluation of MT Systems by TOEFL.” Proceedings of the 5th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'93). 1993

[2] Masaru Tomita. “Application of the TOEFL Test to the Evaluation of Japanese-English MT”', Proceedings of MT Evaluation workshop, AAMT, Nov. 1992

[3] Hitoshi Isahara. et al., “JEIDA’s Test-Sets for Quality Evaluation of MT Systems”. MT-Summit V, July 1995.

[4] John S. White, “The Primacy of Core Technology MT Evaluation”, AMTA. Oct. 1996.