Evaluation Method for Determining Groups of Users Who Find MT "Useful"

M. Fuji (Fujitsu Labs), N. Hatanaka (Canon), E. Ito (Toshiba), S. Kamei (NEC), H. Kumai (Hitachi), T. Sukehiro (Oki), T. Yoshimi (Sharp), H. Isahara (Communications Research Lab)

Fujitsu Laboratories Ltd,
4-1-1, Kamikodanaka, Nakahara, Kawasaki,
Kanagawa 211-8588, JAPAN.
fuji.masaru@jp.fujitsu.com

Abstract

This paper describes an evaluation experiment designed to determine groups of subjects who prefer reading MT outputs to reading the original text. Our approach can be applied to any language pairs, but we will explain the methodology by taking English to Japanese translation as an example. In the case of E-J MT, it can be assumed that main users are Japanese and that most of them have some knowledge of English. It is often the case, in the case of E-J MT systems, that those people who are comfortable with reading English do not find E-J MT outputs useful, and in many cases, they would rather prefer reading the original English text. On the other hand, E-J MT outputs prove to be useful to those who find it hard to read the original English texts. We have used the reading comprehension part of the Test Of English for International Communication (TOEIC) to determine the threshold English ability level, dividing these two user groups.

Keywords

English-Japanese MT, evaluation, reading comprehension, TOEIC and language ability.

Research body

This evaluation experiment has been designed and carried out as part of the activity for the Technical Research Committee of the Asia-Pacific Association for Machine Translation (AAMT).

Background

A large number of E-J MT products are currently put on market in Japan. One of the main purposes of E-J MT systems is to enable Japanese users to browse, in their own language, through vast quantities of English documents.

Although the output quality of MT systems has been improved recently through continuous research and development effort, the quality is not high enough to satisfy all the potential users. The reality is that those Japanese people who have sufficient command of English would prefer reading original English texts, rather than reading the corresponding Japanese text translated using MT. On the contrary, there are a certain number of people, who are less confident with reading English and who would much prefer reading MT translated Japanese text, even though the output quality is nowhere near perfection.

Objective

The objective of our research work is to devise an experiment for evaluating the quality of MT systems in the form that is easily understood by the general public. The experiment is designed in such a way as to determine the groups of subjects who find MT outputs useful. The methodology we adopt here should be such that it can be applied to any language pairs, though we take English to Japanese translation as an example.

General idea

We make use of the TOEIC as a convenient scale to measure MT quality. As large numbers of people take this test world wide, it is hoped that the scores obtained using this material will given some indication as to the level of English ability.

For example, if subject A scores 500 points from the original TOEIC test, while he/she scores 600 points from the E-J MT results, then the chances are the subject finds the MT system useful. For subject B, the score from the original English test may well be higher, indicating that he does not find the MT output useful.

Our assumption is that whether the user finds MT output useful or not, depends on his ability to understand English. In order to find out if this assumption is reasonable, we have asked test subjects (examinees) to answer the reading comprehension part of TOEIC, but translated into Japanese using MT. The subject is then asked to answer another set of TOEIC reading comprehension questions, this time in original English text, in order to measure his command of English. Having answered these two sets of questions, the subject is asked to give his/her impression as to which of the two sets he/she found more comfortable to answer. The scores obtained in the sets are calculated for all the subjects, and together with their impression judgments, the data is processed with respect to their English ability

Overview of TOEIC test set

The original TOEIC test consists of two main sections, the LISTENING COMPREHENSION section and the READING section. The LISTENING COMPREHENSION section is further divided into four parts, namely, Part I through Part IV, which are designed to test four different aspects of listening ability. The READING section is also divided into three parts, namely Part V through Part VII. The last part, Part VII, is the reading comprehension part that is the only part of the entire test that we judge to be suitable for MT evaluation.

The time allowed to answer the entire READING section is 75 minutes, while the time allocation for each part within the section is open to the examinee. There are 100 multiple-choice questions in the entire READING section, 40 of which belong to Part VII. The questions in this part of the test are based on a variety of reading material (e.g. announcements, paragraphs, and advertisements). The examinee is to choose the one best answer, (A), (B), (C), or (D), to each question.

Preparation for experiment

MT systems

We used two commercial E-J MT systems to prepare MT translated reading texts. We used two systems, in order to find out if the obtained results are system dependent.

Reading material

We have used the reading comprehension part of the TOEIC test as the material for our experiment. Since we do not use the entire test, the scores we obtain from our experiments only give approximation to the real score for the entire test.

Time setting

It is important that the conditions of the experiment should be as close to that imposed by the original TOEIC. Since we only use the reading comprehension part, it is necessary to set our own time limit to this part alone.

Since the number of questions in the entire section is 100 and that in Part VII is 40, Part VII can be seen to constitute 40% of the section. This proportion implies that out of the 75 min. given to the entire section, 30 min. is for Part VII. However, this simple arithmetic does not hold true in real situations, where examinees tend to allocate more time to Part VII than to other parts of the section.

We have found out empirically that the typical time allocated to Part VII is 40 min., and hence we have set our time limit for our experiment to 40 min. Since the experiment involves five sets as described earlier, the total sitting time for an experiment session lasts for 40 min. times 5, namely 200 min.

The subjects were allowed to leave the examination room as soon as they had finished answering questions. They were asked to jot down the time it took to answer questions.

List of test sets

There are two distinct ways of using MT outputs. One is to read the MT translated text on its own. The other is to read the source text alongside with the MT translated text. In order to test these modes of usage, we have extracted the reading comprehension parts from five complete sets of TOEIC test, and processed each set as follows.

Set 1	Original source text only and questions (English)
Set 2	Text and questions translation by MT system A (Japanese)
Set 3	Text and questions translated by MT system A (Japanese) alongside with the source text (English)
Set 4	Text and questions translation by MT system B (Japanese)
Set 5	Text and questions translated by MT system B (Japanese) alongside with the source text (English)

Table 1: List of test sets

It must be noted here that we had to prepare five different sets to carry out the experiment, in order to avoid influence across the sets. This is because once the subject has read a reading material, he/she memorizes the content and the same text processed in a different way will be influenced by the first reading.

Test subjects

It is necessary in our experiment to obtain data from subject groups possessing various English language skills. To achieve this end, subjects are chosen from those who have taken the (original) TOEIC test in the past, and we have set twelve subject groups according to the TOEIC scores of the subjects comprising the group. G1 consists of subjects each of whom claims to have scored under 390 points, and G2 consists of subjects scoring 395-440 points, and so on. The score ranges are adjusted so that they are identical to those appearing the statistical data published by the organisers of TOEIC.

Group name	Score range	No. of subjects
G1	-390	5
G2	395-440	10
G3	445-490	18
G4	495-540	19
G5	545-590	19
G6	595-640	19
G7	645-690	19
G8	695-750	23
G9	745-790	18
G10	795-840	11
G11	845-890	10
G12	895-	12

Table 2: Subject groups and number of subjects

It should be noted here that the TOEIC scores mentioned here are those the examinees have previously attained. Therefore, the scores are for the entire test. The score range here does not have to be accurate, as the aim of using their previous scores is to obtain a roughly uniform distribution of English ability.

Questionnaire for collecting subjects' impression

Having answered reading comprehension tests, the subjects are asked to fill out a questionnaire, designed to collect the subjects' impression on MT outputs.

1-1. Which did you think was more comprehensible?

very slightly equiv. slightly very
source(E) 1---------2---------3---------4---------5 MT(J)

1-2. Which did you think was more awkward to read?

very slightly equiv. slightly very
source(E) 1---------2---------3---------4---------5 MT(J)

2-1. Which did you think was more comprehensible?

very slightly equiv. slightly very
source(E) 1---------2---------3---------4---------5 source + MT(EJ)

2-2. Which did you think was more awkward to read?

very slightly equiv. slightly very
source(E) 1---------2---------3---------4---------5 source + MT(EJ)

Figure 1: Questionnaire

Influencing factors

In our experiment, we have solely focused on the effect of the foreign language ability of the examinee. However, there are various factors that are likely to have influence on the evaluation results.

Tuning level

Prior registration of unknown words will improve the MT quality. Some MT systems also allow users to register sets of verb-object pairs, which may also have significant influence.

As it is difficult to control the tuning level, we used the MT systems as they are supplied by the manufacturers without any tuning.

Functionality of MT systems

Some MT systems have functionality to process itemized or tabular text objects, which may improve the output quality, provided the input text is stored in files along with formatting information.

In order make fair comparison, we have stored the reading comprehension material in HTML format, preserving the itemization and tables. These HTML files are translated by the MT system in WWW browser mode.

Translation of multiple-choice questions

There are two distinct ways of looking at the task of translating TOEIC reading comprehension tests into Japanese.

One way of looking at it is that the multiple-choice questions as well as the reading texts are the essential elements comprising the test. It naturally follows from this viewpoint that every single piece of text appearing in the reading comprehension part has to be translated into Japanese by MT, in order to make fair judgment.

The second way is that the reading material alone is the essential element employed to judge the language ability of the examinee. From this follows that only the reading material is to be translated into Japanese using MT, while the rest (multiple-choice questions, etc) is to be translated manually into Japanese.

For the sake of simplicity we have translated both the reading text and questions using MT, rather than translating the questions by hand.

Results and analyses

Difference in MT systems

We have used two commercially available English to Japanese MT systems, both of which have been put on the Japanese market for some time and have fixed users. Though sentence-by-sentence comparison of the translated outputs from these two systems would yield some differences, are these differences significant enough to influence the comprehension performance?

Table 3 shows the results of test of significance for the two MT systems. The tests have been carried out for each subject group.

Group	MT only	Source + MT
G1	0.7003	0.0123
G2	0.2807	0.2355
G3	0.1203	0.5070
G4	0.7889	0.5134
G5	0.5975	0.0657
G6	0.0251	0.6392
G7	0.7571	0.9254
G8	0.2091	0.5223
G9	0.7464	0.7414
G10	0.8444	0.6120
G11	1.0000	0.3636
G12	0.9481	0.4268

Table 3: T-tests between MT systems

Presentation of MT output only

The column titled "MT only" in Table 3 gives the two-sample t-test results for experiments where only the MT outputs are presented to the subjects. This experiment is carried out for the two MT systems and the distribution of comprehension scores are compared for each subject group using two-sided t-test.

Taking 0.1 as our significance level, no significant difference was observed for all subjects groups except for a small difference in G6. On the whole, the two MT systems can be regarded as performing almost equivalently in this experiment.

Presentation of source text and MT output text

The column titled "Source + MT" corresponds to the experiment where both the source language (English) text and the MT output (Japanese) text are simultaneously presented to the subject. The same tests as for "MT only" were carried out and the test results tabulated. No significant difference was observed for all subject groups except for small differences in G1 and G5. Therefore, the two systems can be regarded to perform almost equivalently in this mode of experiment too.

It can be assumed from the above two sets of results that the difference in the performance of the two MT systems used are almost equivalent in comprehension tests. Hence we will use the results for one of the two MT systems in the later parts of this paper, in order to illustrate the data in a comprehensive manner.

Effect of MT on comprehension performance

Regardless of how the subjects feel about MT outputs, their performance can be measured using reading comprehension tests. The following section describes the results of the performance tests we have carried out.

Original source text (base line)

It is reasonable to expect that those subject groups with high TOEIC scores will obtain high scores in the reading comprehension tests, while those in low TOEIC score groups will obtain low scores. This expected correlation is clearly seen in Figure 2. The number of multiple-choice questions is 40, and hence a value of 40 will be observed on the y-axis for full score. It can be seen from the graph that the an average score of around 20 is obtained among low TOEIC score groups, while an average score of over 30 is obtained among high TOEIC score groups.

Our interest then is to find out how this line is affected by the introduction of an MT system.

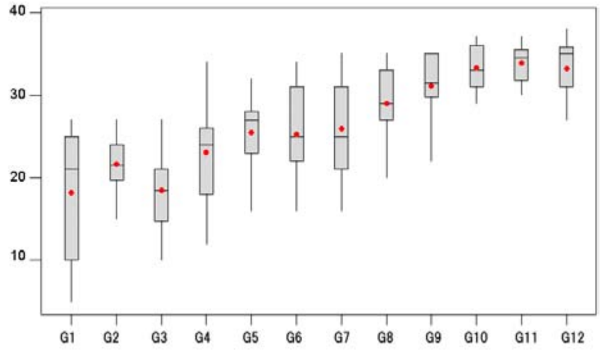

Figure 2: Scores for original source text

Introduction of MT

There are two distinct ways of using MT outputs for reading comprehension. The first way is to read the MT output alone, and the second way is to read the MT output alongside with the original source text. Both of these modes were experimented.

Figure 3 shows the score distribution for each subject group when MT outputs alone were presented. It can be seen from the graph that the average score is almost constant, and an average score of around 20 is obtained among almost all the subject groups.

By comparing this graph with Figure 2, it can be said that the comprehension level observed among low TOEIC score groups seem to be much the same when MT alone is introduced. The comparison also leads to the remark that the introduction of MT alone degrades the comprehension level for high TOEIC score groups. These intuitive remarks visually obtained will be statistically evaluated in later parts.

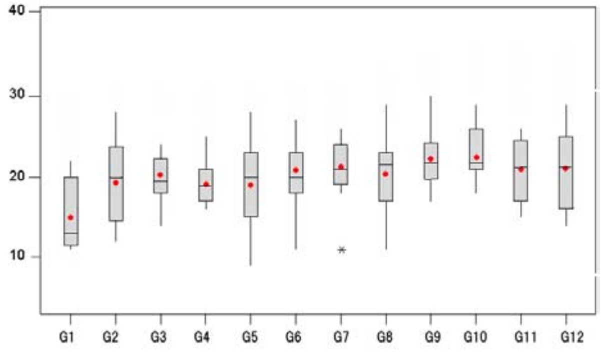

Figure 3: Scores for MT output alone

Figure 4 shows the comprehension performance when the MT output texts as well as their respective source texts were presented to each subject.

It can be seen from the graph that an average score above 30 is obtained among the high TOEIC score groups, while it only goes down to around 25 among the low TOEIC score groups. It can be said by comparing with Figure 2, that not much effect is observed by the introduction of MT among the high TOEIC score groups, while the subjects tend to obtain higher scores by introducing MT among the lower TOEIC score groups.

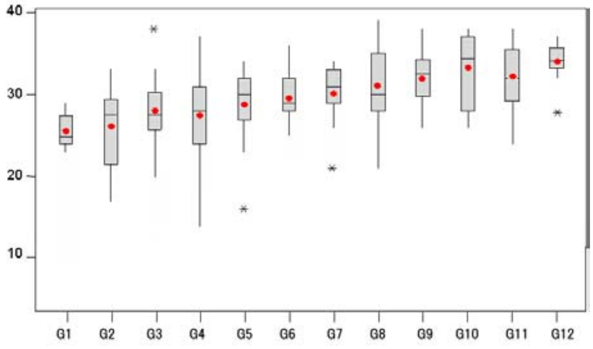

Figure 4: Scores for source text and MT output

Test for significance

The discussion above gives the effect of introducing MT in a qualitative manner. Since we have managed to carry out our experiments using a fairly large number of subjects, it is possible to process the data statistically.

Table 4 shows the t-test for measuring the introduction of MT only and the introduction of source and MT, compared against the performance for the original text. The column titled “MT only” shows the t-test comparison between the distribution for the original text and the distribution for the MT alone. The column titled “Source + MT” shows the t-test comparison between the distribution for the original text alone and the distribution for the source text and the MT output.

We have taken the significance level to be 0.1, and those figures outside this level are regarded to be significant and highlighted by the shaded boxes in the table. It should be noted here that the test only shows whether there are significant differences, but does not show which score is higher. The latter is quite obvious, and should be deducted from Figures 2, 3 and 4.

Groups	MT only	Source + MT
G1	0.4975	0.0958
G2	0.5053	0.0271
G3	0.0013	0.0000
G4	0.4900	0.0051
G5	0.1426	0.0143
G6	0.2715	0.0026
G7	0.2298	0.0087
G8	0.0021	0.7129
G9	0.0003	0.1449
G10	0.0008	0.9255
G11	0.0003	0.4778
G12	0.0000	0.6550

Table 4: T-test for measuring effect of MT introduction

In the case of “MT only”, G8 through G12 obtained significantly higher scores from the source texts than from MT outputs, while the rest of the groups did not see significant differences. This is roughly equivalent to saying that those subjects with scores higher than 695 would prefer the source text to the MT output, while those lower than this score would not benefit from MT outputs alone.

On the other hand, in the case of “Source and MT”, those subject groups with TOEIC scores lower than 695 would obtain significantly higher results from the combination of the source and MT texts. Those with TOEIC scores higher than 695 would see no difference.

Answering time

We set our time limit for answering reading comprehension tests to one minute per question. This implies that 40 minutes are allowed for each test set. However, we have allowed the subjects to leave the examination room as soon as they have finished with the test set they were given. They were also asked to jot down the time it took to finish answering the test set.

Our assumption here is that this time period also indicates the usefulness of MT outputs.

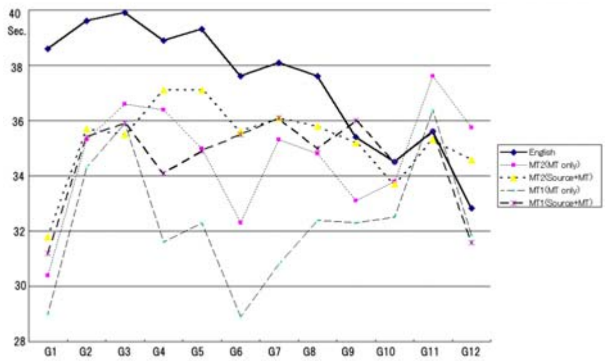

Figure 5: Average answering time

Figure 5 shows the average answering time for each subject group. The results for MT1 and MT2 are plotted separately, since this is the only experiment where some difference was observed between the two MT systems.

It can be seen that the answering time seems to be reduced by the introduction of MT among subject groups with low TOEIC scores. On the other hand, there seems to be little or no effect of MT among the rest of the subject groups.

It should be noted that an inversion of the average scores is observed on the right hand side of the graph, probably indicating that those subjects with high TOEIC scores tend to spend considerable amount of time trying to figure out the meaning from poorly translated texts.

	MT only	Source + MT
G1	0.0276	0.1021
G2	0.0027	0.0016
G3	0.0005	0.0008
G4	0.0002	0.0001
G5	0.0000	0.0009
G6	0.0000	0.0394
G7	0.0004	0.0469
G8	0.0004	0.0301
G9	0.0078	0.9432
G10	0.5897	0.1087
G11	0.5199	0.3486
G12	0.8563	0.4769

Table 5: Significance level for time difference

Table 5 is a table showing the probability that the introduction of MT is of significance, for each subject group. Taking 0.1 as our significance level, significant differences were observed at figures smaller than 0.1, where the figures are indicated by grey boxes in the table.

In the case where only MT is presented, subject groups G1 through G9 shows significant difference, namely, it takes significantly less time to answer questions when MT is introduced. In the case where the source texts as well as the MT outputs are presented, G1 through G5 exhibits significant difference, though with some slight exceptions (G1 and G4). This implies that the introduction of MT has reduced the answering time for these groups.

Subjects’ impression of usefulness

All of the subjects who took place in our performance tests were asked to give their impression on the text material they were presented. It is likely that the impression they get from the tests are closely related to the performance measured in terms of the scores they obtain, but it is of great interest to see to what extent this is true.

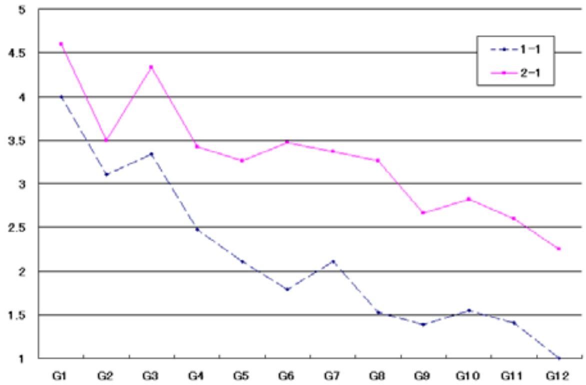

Figure 6: Average scores for comprehensibility

1-1 MT only
2-1 Source and MT

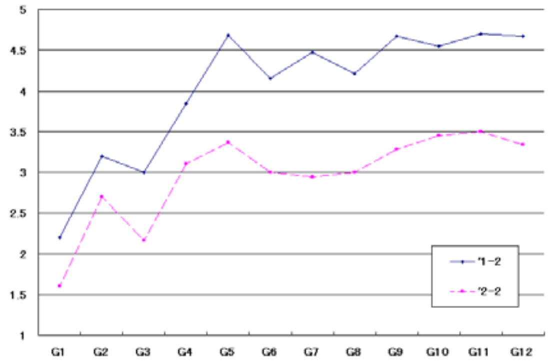

Figure 7: Average scores for awkwardness

1-2 MT only
2-2 Source and MT

Figures 6 and 7 show the variations of average impression scores for each subject group. These figures were obtained in response to the questionnaire shown in Figure 1. It should be noted that the maximum value of impression is 5 and minimum 1, and that 3 is the midpoint. Our evaluation described below is based on the comparison of each individual result against this midpoint score, namely, point 3.

The general trend of these graphs is that the comprehensibility impression of MT against source texts goes down as the TOEIC score goes up, while the awkwardness goes up as the TOEIC scores goes up.

	MT only	Source + MT
G1	0.8193	0.9995
G2	0.5382	0.6610
G3	0.6136	0.9088
G4	0.3223	0.6442
G5	0.1468	0.5741
G6	0.0571	0.6610
G7	0.1776	0.6110
G8	0.1009	0.6236
G9	0.0165	0.4013
G10	0.0015	0.4429
G11	0.0005	0.3597
G12	0.0000	0.2989

Table 6: Comprehensibility

	MT only	Source + MT
G1	0.2464	0.0021
G2	0.5599	0.4065
G3	0.5000	0.2055
G4	0.7620	0.5392
G5	0.9985	0.6184
G6	0.8562	0.5000
G7	0.9750	0.4824
G8	0.7561	0.4119
G9	0.9938	0.5887
G10	0.9990	0.6721
G11	0.9999	0.7324
G12	0.9998	0.6382

Table 7: Awkwardness

Table 6 and Table 7 show the comprehensibility impression and awkwardness impression, tested against the midpoint. The figure gives the probability that the distribution of impression levels among each group goes above midpoint (3). We regard figures above 0.9 and below 0.1 to be significant and the significant figures are shown in grey boxes.

It is noted that, when MT output alone is presented, the impression level for comprehensibility significantly degrades for groups from G9 through G12. When both source and MT texts are presented, the comprehensibility impression significantly improves for groups from group 1 through G3. Much the same results are obtained for the impression of awkwardness.

Conclusions

We have successfully designed and carried out an evaluation experiment for determining the groups of user who would benefit from the introduction of an MT system.

Our approach has proved to give results that are statistically significant, and that are easily understood by the general public.

For the given language pair, namely English and Japanese, and for the given MT systems, it is reasonable to make the following remarks.

- **Comprehension performance**
  Among subjects with TOEIC scores lower than a certain level, namely around 700 points, the comprehension performance where both the source text and MT output are presented prove to be significantly higher than the case where only the source text is presented. The presentation of the MT output alone does not significantly improve comprehension performance among any subject group.
- **Time performance**
  The introduction of MT proves to reduce the time to answer questions among subject groups lower than a certain level.
  However, the time performance seems to be influenced by factors other than the TOEIC scores. It may be influenced by the fact that those subjects who have good command of English tend to spend fair amount of time guessing the intended meaning from MT outputs.
- **Impression**
  Subjects' impression on comprehensibility is closely related to the inverse of awkwardness. When both the source and MT texts are presented, a significant improvement in impression is observed among subjects with TOEIC scores lower than approx. 500 points.

It can be concluded from the above remarks, that a fairly large proportion of subjects benefit from MT outputs, provided MT outputs are presented alongside with the original source text. However, the proportion of subjects whose impression improves is smaller than the above figure, implying that some subjects obtain higher scores when MT is introduced, but they do so with awkwardness.

The figures obtained in our experiment can be used to make rough approximation to the proportion of the entire population who would benefit from MT. This can be achieved by comparing the above results with the statistical data published by the organising bodies of the TOEIC. For example, the percentage of examinees who scored less than 695 TOEIC (IP) in Japan in July 2000 was around 90%, implying that this proportion is likely to benefit from the presentation of both the source and MT output texts.

Acknowledgements

We would like to give our special thanks to Ms. Y. Shibata, Mr. S. Ohno and other staff of IR-Advanced Linguistic Technologies Inc., for preparing and carrying out all the necessary experiments.

Bibliographical References

Hitoshi Isahara, et al., (1995) JEIDA's Test-Sets for Quality Evaluation of MT Systems. In proceedings of MT-Summit V.

Masaru Fuji, (1999). Evaluation Experiment for Reading Comprehension of Machine Translation Outputs. In proceedings of MT Summit VII.