Evaluation of Japanese-Chinese MT System using AAMT's Test-Set

Tomoki NAGASE[1], Katsunori KOTANI[2], Masaaki NAGATA[3], Nobutoshi HATANAKA[4] ,
Yoshiyuki SAKAMOTO, Eiichiro SUMITA[5], Kiyotaka UCHIMOTO[5]

中文摘要：
亚太机器翻译协会的机器翻译课题调查委员会，开发了面向日汉翻译系统开发人员的翻译质量评价测试集（AAMT 测试集）。通过利用 AAMT 测试集，评测人员仅需要对译文句法相关的提问以 yes/no 进行回答，就可以获得比以往人工评价更高效的评价结果。本文中介绍了利用 AAMT 测试集与目前方法实际进行了日汉机器翻译引擎的评测情况，并通过目前评测方法和 yes/no 方法结果相关性、利用各种评测方法时评测人员间的差异等观点的分析，总结了评测以及分析结果。

1. Introduction

The methods for evaluating the quality of a machine translation (MT) system have two major problems. One is the establishment of criteria which are independent of evaluator objectivity and the other is reducing man-hour costs needed to evaluate huge sets of translation sentences. With these aims, JEIDA (Japan Electronic Industry Development Association) had developed JEIDA's Test-Sets [1] which are applied to English-Japanese and Japanese-English translation evaluation. One of the features of their Test-Sets is translation examples that include grammatical checkpoints, which are answered "Yes" or "No" by evaluators. Compared with conventional evaluating methods, such as grading translation on its quality into 3 or 5 levels, the method based on answering yes/no questions is more objective.

Since 2007, we have been developing the new Test-Set for evaluating quality of Japanese-Chinese MT systems, expanding JEIDA's Test-Sets [2]. In June 2009, our new Test-Set was completed and made publicly available [3]. The work described in this paper has been developed by the Special Working Group of AAMT (Asia-Pacific Association for Machine Translation) (we call our Test-Set "AAMT's Test-Set"). In 2008, we evaluated the quality of several actual Japanese-Chinese MT systems, applying AAMT's Test-Set. In order to compare with conventional methods, we also carried out subjective evaluation focused on fluency and adequacy.

In this paper we would like to describe the merits of the evaluation method using AAMT's Test-Set comparing with conventional methods. In Section 2, we will overview AAMT's Test-Set showing examples of yes/no questions. Section 3 describes the details of evaluation experiments we carried out. Chapter 4 describes variability of evaluation results depending on evaluators, and analyzes correlation among evaluating methods such as fluency evaluation, adequacy evaluation, yes/no evaluation and statistical evaluation (BLEU)[4]. Moreover, we will briefly discuss strong and weak points of MT systems, comparing Japanese-Chinese and Japanese-English MT systems.

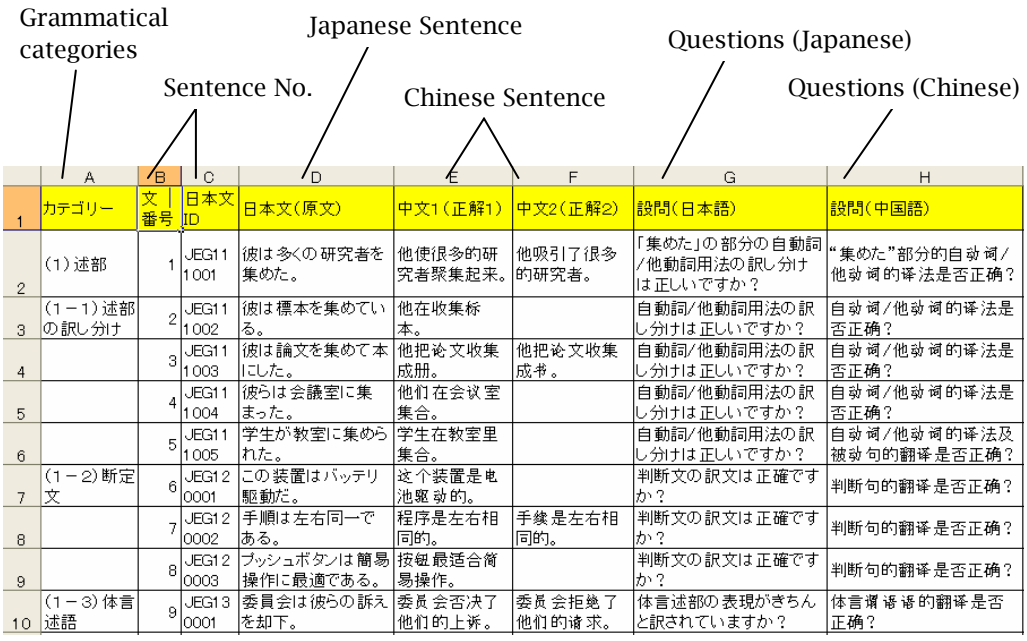
[1] Fujitsu Laboratory Ltd.
[2] Kansai Gaidai University
[3] NTT Communication Science Laboratories
[4] Tokyo University of Information Sciences
[5] MASTAR Project/National Institute of Information and Communications Technology

2. Test-Set for Japanese-Chinese MT systems

Our Test-set for Japanese-Chinese MT systems consists of source sentences (Japanese), target sentences (Chinese), grammatical classification labels and yes/no questions for checking grammatical points. The details of each of these items are as follows:

- Source language sentences (Japanese)
  378 Japanese sentences which are the same as the examples of JEIDA's Test-Set for Japanese-English MT systems.
- Target language sentences (Chinese)
  We made new Chinese sentences by translating the source sentences. In some Japanese sentences, it is difficult to choose only one translation. As a result, 35 of 378 Japanese sentences have two corresponding Chinese sentences.
- Grammatical classification labels
  In JEIDA's test-sets, linguistic phenomena in Japanese were classified into 45 categories. Each Japanese-Chinese example pair in our Test-Set has a label indicating the classified categories.
- Yes/no questions for checking grammatical points
  These were mainly redefined from the point of view of Japanese analysis. They were prepared in Chinese as well as in Japanese, so that any Chinese person can achieve evaluation with the Test-Set, even if he/she can't understand Japanese.

The example of AAMT's Test-Set is shown in Figure 1.



	A	B	C	D	E	F	G	H
1	カテゴリー	文番号	日本文ID	日本文(原文)	中文1(正解1)	中文2(正解2)	設問(日本語)	設問(中国語)
2	(1)述部	1	JEG111001	彼は多くの研究者を集めた。	他便很多的研究者聚集起来。	他吸引了很多的研究者。	「集めた」の部分の自動詞/他動詞用法の訳し分けは正しいですか？	"集めた"部分的自动词/他动词的译法是否正确？
3	(1－1)述部の訳し分け	2	JEG111002	彼は標本を集めている。	他在收集标本。		自動詞/他動詞用法の訳し分けは正しいですか？	自动词/他动词的译法是否正确？
4		3	JEG111003	彼は論文を集めて本にした。	他把论文收集成册。	他把论文收集成书。	自動詞/他動詞用法の訳し分けは正しいですか？	自动词/他动词的译法是否正确？
5		4	JEG111004	彼らは会議室に集まった。	他们在会议室集合。		自動詞/他動詞用法の訳し分けは正しいですか？	自动词/他动词的译法是否正确？
6		5	JEG111005	学生が教室に集められた。	学生在教室里集合。		自動詞/他動詞用法の訳し分けは正しいですか？	自动词/他动词的译法及被动句的翻译是否正确？
7	(1－2)断定文	6	JEG120001	この装置はバッテリ駆動だ。	这个装置是电池驱动的。		判断文の訳文は正確ですか？	判断句的翻译是否正确？
8		7	JEG120002	手順は左右同一である。	程序是左右相同的。	手续是左右相同的。	判断文の訳文は正確ですか？	判断句的翻译是否正确？
9		8	JEG120003	プッシュボタンは簡易操作に最適である。	按钮最适合简易操作。		判断文の訳文は正確ですか？	判断句的翻译是否正确？
10	(1－3)体言述語	9	JEG130001	委員会は彼らの訴えを却下。	委员会否决了他们的上诉。	委员会拒绝了他们的请求。	体言述部の表現がきちんと訳されていますか？	体言谓语的翻译是否正确？

Figure 1: ***Example of AAMT's Test-Set***

3. Experimental Process

A major aim of our experiment is to determine whether or not a yes/no based evaluating method yields results that are comparable to those of the conventional method, in evaluating quality of MT.

Therefore we carried out not only the yes/no based method, but also a conventional scoring method from the point of view of adequacy and fluency. Moreover we assigned two Chinese people as evaluators, so that we were able to compare variability of evaluation result between them.

Our experiments of evaluation for the quality of Japanese-Chinese MT system were conducted as follows.

(1) Each sentence in the Test-Set was translated using six different Japanese-Chinese translation engines that can be used on the Internet.
(2) Two evaluators independently answered “yes” or “no” to the question on each example by referring to the translation result.
(3) Two evaluators independently evaluated adequacy rating on a 1-5 scale by referring to the translation result.
(4) Two evaluators independently evaluated fluency rating on a 1-5 scale by referring to the translation result.
(5) BLEU, a common statistic MT evaluating method, was applied to examples in the test-set and evaluated the translation result.

The rating criteria in evaluating adequacy and fluency conformed to that of the DARPA TIDES (Translingual Information Detection, Extraction and Summarization) program.

- Fluency

How do you judge the fluency of this translation?

It is: 5:Flawless Chinese 4: Good Chinese 3: Non-native Chinese
2: Disfluent Chinese 1: Incomprehensible

- Adequacy

How much of the meaning expressed in the gold-standard translation is also expressed in the target translation?

5: All 4: Most 3: Much 2: Little 1: None

4. Experimental Results and Discussion

4.1 Correlation between yes/no evaluation and conventional subjective evaluation

Table 1: *Evaluation Results*

	Evaluator A						Evalator B						Automatic Evaluation	
Engine	Adequacy		Fluency		Yes/No		Adequacy		Fluency		Yes/No		BLEU	
JC1	3.52	2	2.96	4	0.33	4	2.52	3	2.45	3	0.53	3	**32.80**	1
JC2	3.50	3	3.03	3	0.35	3	2.47	4	2.42	4	0.50	4	28.90	5
JC3	2.90	6	2.36	6	0.17	6	1.98	6	1.98	6	0.33	6	24.20	6
JC4	3.20	5	2.72	5	0.28	5	2.38	5	2.39	5	0.49	5	29.40	4
JC5	3.42	4	**3.08**	1	**0.39**	1	**2.62**	1	**2.60**	1	**0.57**	1	30.70	3
JC6	**3.60**	1	**3.08**	1	**0.38**	2	2.61	2	2.58	2	0.55	2	32.20	2
Average	3.36	-	2.87	-	0.31	-	2.43	-	2.40	-	0.47	-	29.70	-

Table 1 summarizes the results of evaluating quality of Japanese-Chinese MT systems. Both evaluator A and B are native Chinese who can understand Japanese. Values in Adequacy and Fluency

columns are averages of scores rated on a 1-5 scale. Values in yes/no columns are averages of scores "1" (answered yes) or "0" (answered no). We also listed evaluation results with automatic evaluation BLEU which is one of the most common among statistical evaluation methods, for reference.

As shown in Table 1, ranking orders in yes/no evaluations are completely equivalent to that of fluency by both evaluator of A and B. As for adequacy and yes/no ranking, they are equal in evaluator B but slightly different in evaluator A. As a whole, yes/no evaluation results correlate with those of both fluency and adequacy. The correlation coefficients between yes/no and adequacy are shown in Table 2, which are calculated according to following expression:

$$r_{aq} = \frac{\sum_{i=1}^{n} (a_i - \overline{a})(q_i - \overline{q})}{\sqrt{\sum_{i=1}^{n} (a_i - \overline{a})^2} \sqrt{\sum_{i=1}^{n} (q_i - \overline{q})^2}}$$

Here, a_i and q_i correspond to the scores of adequacy and yes/no evaluation for the i-th example, and $\overline{a}$ and $\overline{q}$ correspond to the average scores of adequacy and yes/no evaluation. The value n indicates the number of examples. The correlation coefficients between yes/no and fluency can be calculated in the same way.

Table 2: ***Correlation coefficients between yes/no and adequacy fluency***

Evaluator A

	yes/no
Adequacy	0.66
Fluency	0.65

Evaluator B

	yes/no
Adequacy	0.48
Fluency	0.49

4.2 Variability of evaluation depend on evaluators

Table 3 shows the ratio of conflict of yes/no evaluation between the two evaluators. Yes/no evaluation should be designed so that anybody can evaluate an MT engine and return the same answer ("Yes" or "No") for questions belonging to the same examples. However, in about a quarter of the questions, conflicts are occurring. We wonder if a very significant number of ambiguous yes/no questions are included in the test-set. For example, the following question:

「副词"ゆっくり"的翻译是否正确？」

Reading this question, some evaluators may think translation of "ゆっくり" should be the same word used in given examples, other evaluators may judge correctness based on their own knowledge. We are expecting to revise questions to be less ambiguous, so that the conflict ratio may be reduced.

Table 3: ***Conflict of yes/no evaluation between evaluators***

	number of examples	
Both "Yes"	518	23%
Both "No"	1157	51%
Conflict	587	26%

Figure 2 and Figure 3 indicate distribution of adequacy and fluency scores (1-5), respectively, for six

MT systems by evaluators. In these figures, the vertical axis represents frequency of sentences, and the horizontal axis represents scores.

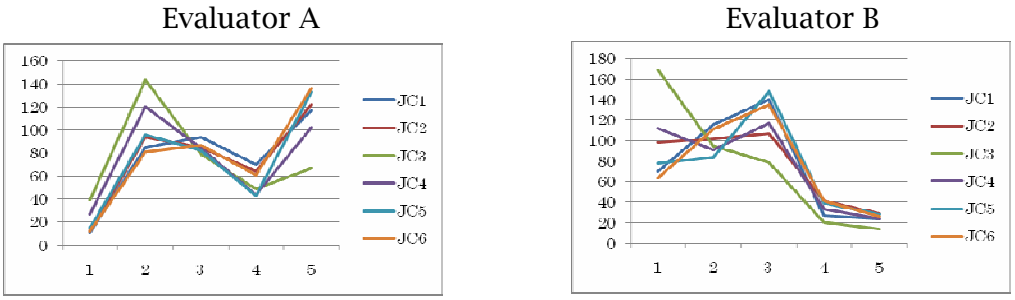


Figure 2: ***Distribution of Adequacy score by evaluators***

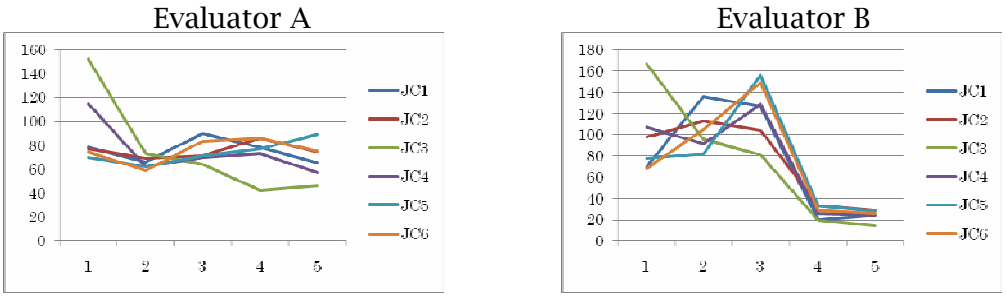


Figure 3: ***Distribution of Fluency scores by evaluators***

As shown in Figure 2, evaluator A puts “2” and “5” a lot (“1” seldom), whereas evaluator B puts “3” a lot (“4” and “5” seldom). Scoring criteria for subjective evaluation are so unclear that they can be interpreted in various ways by evaluators. In *fluency* evaluation, Figure 3 indicates a big difference of distribution between two evaluators.

4.3 Efficiency of Evaluation Process

The efficiency of the evaluation is also an important point for a MT evaluation method. We found that the evaluation time of yes/no answering in our test-set takes less than half the time compared to that of fluency or adequacy scoring. On the other hand, when discussing the efficiency of evaluation process, we should also consider the productivity of the process to compose yes/no questions. Since hundreds of questions must be attached to each example, it takes a lot of time and effort to go through it.

To enable the evaluation approach based on yes/no question to be more efficient, Uchimoto et al. have proposed a method which is able to automatically determine the answer to each question as yes or no [5]. Moreover, Uchimoto have tried to develop a method that can automatically generate yes/no questions. Since their work could be applied for English into Japanese MT, we have to arrange for Japanese into Chinese MT, which is our future work.

4.4 Weak points for Japanese-Chinese MT Engines

The other merit of using our test-set is that MT developers can recognize strong/weak points of their systems. They can find unsupported grammatical phenomena for their MT systems.

In this work, we tried to clarify the relative position of current Japanese-Chinese MT systems, compared to Japanese-English systems. In order to do this, we also evaluated six Japanese-English

MT systems in the same way as the process mentioned in chapter 3. The results of yes/no questions are shown in Table 4. It is right to presume that the quality of Japanese-English translation is much higher than that of Japanese-Chinese translation, as a whole.

Table 4: ***Scores of yes/no Evaluation***

Language Pair	Average Score of six Engines	
Japanese to Chinese	0.40	0.31 (Evaluator A)
		0.47 (Evaluator B)
Japanese to English	0.72	0.75 (Evaluator C)
		0.68 (Evaluator D)

Figure 4 shows details of scores evaluated in each of the grammatical categories which are defined in AAMT's test-set. That indicates the scores of Japanese-English translation are higher in almost every grammatical category, but the variability of scores shows different tendencies according to the translation language pair. The grammatical categories can be classified into three groups based on the yes/no evaluation results. The three groups are as follows:

- Good at both Japanese-Chinese and Japanese-English : Group A
- Good at Japanese-English but bad at Japanese-Chinese : Group B
- Bad at both Japanese-Chinese and Japanese-English : Group C

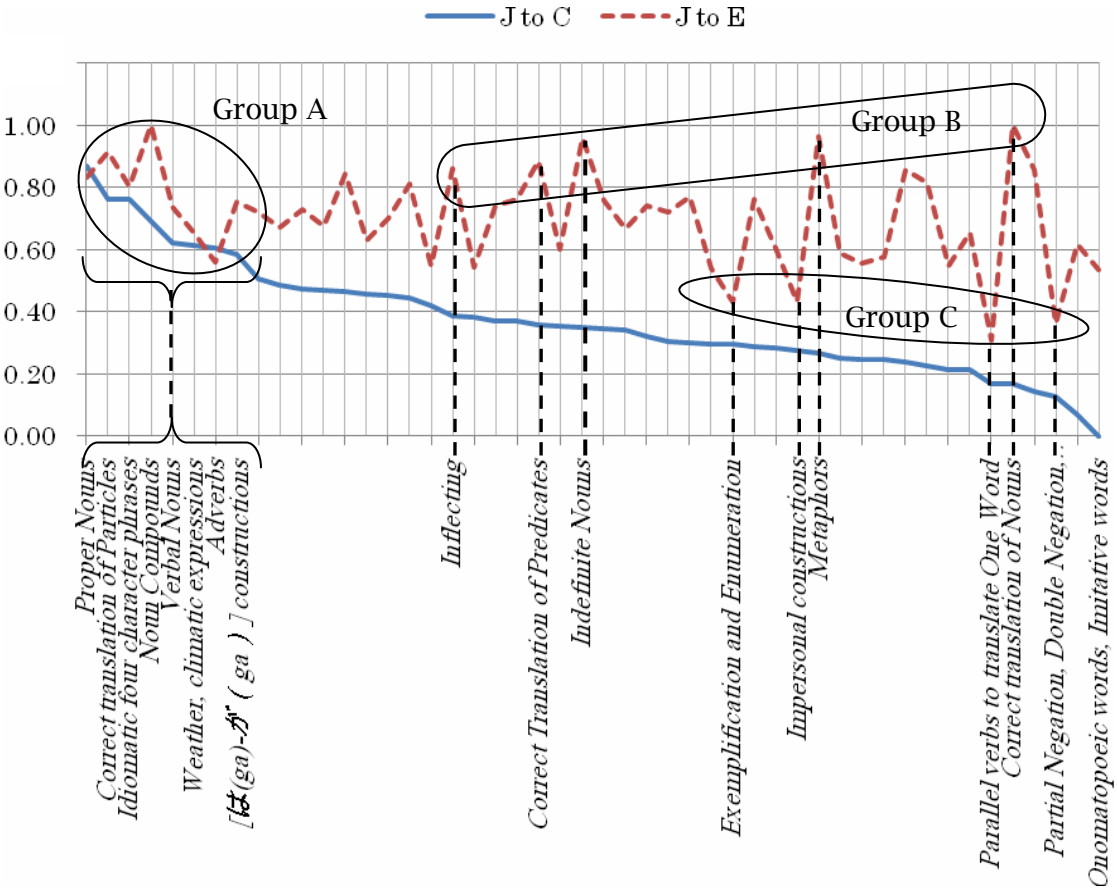


Figure 4: ***Result of yes/no evaluation of each of checking items***

(1) Good at both Japanese-Chinese and Japanese-English

This group contains item which can only be treated by registering words to *a dictionary*, such as "*Proper Nouns*", "*Idiomatic four character Phrases*", "*Noun Compounds*", and "*Verbal Nouns*". In noun phrase translation, translating into English is often more difficult than the case into Chinese. For example, in case of translating "変化量" into English, translation engine should properly analyze and construct a phrase (e.g. "the amount of change"). By contrast, when translating into Chinese, translation engine should simply convert characters (e.g. "变化量"). Therefore, it can be argued that the quality of Japanese-Chinese engines could exceed that of Japanese-English engines in noun phrase translation if the Japanese-Chinese dictionary grows as large as the Japanese-English one.

(2) Good at Japanese-English but bad at Japanese-Chinese

The items which belong in this group are *"Inflecting", "Correct translation of Predicates", "Indefinite nouns", "Metaphors" and "Correct translation of Nouns"*. These items have examples that are difficult to be precisely translated without meaning-based processing. For example, Japanese conjunctive "ため (tame)" has two meanings.

私はプログラムを実行するためディスクを拡張する。
(我为了执行程序增大了磁盘。 I enlarge the disk in order to install a program.)
プログラムを実行したためディスクが満杯になる。
(由于执行程序磁盘满了。 The disk is filled up because a program is installed.)

There is one Japanese-Chinese and five Japanese-English systems which can precisely choose meaning of "ため" for the above two examples among the six systems evaluated respectively. Since Japanese-Chinese translation system development does not have a very long history, word selection process, especially those needing meaning-based programs, pales compared to Japanese-English translation systems.

(3) Bad at both Japanese-Chinese and Japanese-English

There were several items in which examples could not be accurately translated in both Japanese-Chinese and Japanese-English translation. The items are *"Exemplification and Enumeration", "Impersonal construction", "Parallel verbs to translate One Word", and "Parallel Negation, Double Negation, Inversion"*. The following is an example of *"Double Negation "*:

それは望みがないこともない。
(那不是完全没有希望。 That is not entirely hopeless.)

Neither Japanese-English nor Japanese-Chinese systems can properly translate the above sentence, because even for a human translator it is often difficult to translate *Double Negation* sentences. The following is an example of "*Parallel verbs to translate One Word* ":

ナットA を軸に組み付ける。
(在轴上装配螺帽A。 Assemble nut A on the shaft.)

This kind of problem is solved by registering words to a dictionary, but since complex-verbs are

uncontrollably generated, techniques for automatic dictionaries-maintenance are expected. Statistic approaches may be effective against such a problem.

5. Conclusion

In this paper, we have evaluated six Japanese-Chinese MT systems by the AAMT's Test-Set, and discussed the evaluation process and its results from the point of objectivity, efficiency and usefulness for MT developers.

The evaluation results from answering yes/no questions strongly correlated with adequacy/fluency based evaluation results. In other words, the yes/no based evaluation method reflects a human's intuitive feel for MT translation. According to the results of this experiment, a yes/no based evaluating method yields results that are comparable to those of conventional method, in evaluating quality of MT. However, it is necessary to verify the comparability by a lot more evaluators. Answers to questions should not vary among evaluators, but in our experiments, conflicts were actually occurring in a quarter of the examples. We suppose the reason is that some questions are unclear and evaluators didn't share criteria for yes/no judgments. We are expecting to revise questions to be clearer, so that objectivity may be improved.

From the view point of efficiency, because evaluators need only to answer the question in our test-set, the efficiency of the evaluation process was greatly improved. We found that the evaluation time of yes/no answering takes less than half that of fluency or adequacy scoring.

As a whole, the quality of Japanese-Chinese translation is still much lower than that of Japanese-English translation. Evaluation by yes/no question clarified the defects of Japanese-Chinese translation quality on each grammatical item. Especially semantic processing, such as word sense disambiguation, should be improved in Japanese-Chinese engine.

The questions in the test-set are originally designed to cover grammatical points of Japanese-English MT. There are enough sentences for checking Japanese analysis but insufficient sentences for checking Chinese generation. We are going to add examples and questions for the next version of the test-set, as an activity of AAMT.

Our test-set (shown in Figure 1) is freely available to the world, downloadable from the internet site. (http://corpus.aamt.info/)

References

[1] H. Isahara : "JEIDA's Test-Sets for Quality Evaluation 0f MT Systems –Technical Evaluation from the Developer's Point of View—", Proceedings of MT Summit V, 1995.

[2] AAMT Working Group: "Toward Building of AAMT Test-sets for Quality Evaluation of Machine Translation", AAMT Jounal No.42, 2008.

[3] AAMT Working Group: "Toward Publication of AAMT Test-sets for Quality Evaluation of Machine Translation", AAMT Jounal No.45, 2009.

[4] K. Uchimoto et al.: "Automatic Evaluation of Machine Translation Based on ate of Accomplishment of Sub-goals", Proceedings of NAACL HLT 2007.

[5] K. Papineni et al. : " BLEU: a method for automatic evaluation of machine translation" Proceedings of 40th Annual meeting of the ACL, 2002.