Automatic Error Analysis Based on Grammatical Questions*

Tomoki Nagase[a], Hajime Tsukada[b], Katsunori Kotani[c],
Nobutoshi Hatanaka[d], Yoshiyuki Sakamoto[e]

[a]Fujitsu Laboratories,
1-1, Kamikodanaka, 4-chome, Nakahara-ku, Kawasaki, 211-8588, Japan
nagase.tomoki@jp.fujitsu.com
[b]NTT Communication Science Laboratories,
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
tsukada.hajime@lab.ntt.co.jp
[c]Kansai Gaidai University,
16-1 Nakamiyahigashino-cho, Hirakata, Osaka 573-1001, Japan
kkotani@kansaigaidai.ac.jp
[d]Tokyo University of Information Sciences,
4-1 Onaridai,Wakaba-ku,Chiba,265-8501 Japan
hatanaka@rsch.tuis.ac.jp
[e]5-38-7, Akatsutsumi, Setagaya-ku, Tokyo, 156-0044, Japan
ysakanlp@ka2.so-net.ne.jp

Abstract. The present paper proposes automatic error analysis methods that use patterns representing grammatical check points. Our method is comparable to or slightly outperforms conventional methods for automatic evaluation metrics. Different from the conventional methods, our method enables error analysis for each grammatical check point. While our method does not depend on languages, we experimentally show its validity by using a Japanese-to-Chinese test set. Errors in existing Japanese-to-Chinese translation systems are also analyzed.

Keywords: automatic evaluation, machine translation, error analysis

1 Introduction

In the machine translation (MT) field, human evaluators subjectively evaluate the results of MT systems from the viewpoints of adequacy and fluency. However, two problems in subjective evaluations have been noted. The first is that the time and cost consumption of such evaluations is high. The second is that the evaluations have poor reproducibility due to the difficulty of reaching agreement on the scoring criteria among the evaluators. Thus, in the past decade, evaluation methods that automatically evaluate MT quality based on its similarity to a reference translation have become common (Papineni et al., 2002; Doddington, 2002; Banerjee and Lavie, 2005; Snover, et al., 2006; Melamed et al., 2007; Isozaki et al., 2010; Birch and Osborne, 2011).

Error analysis that investigates the strengths and weaknesses of a translation system is required for developing more accurate translation systems. According to the error analysis, system developers improve the translation rules/dictionaries of rule-based MT systems or add bilingual corpora/dictionaries to statistical MT systems. However, all the evaluation methods mentioned that include human subjective evaluation provide less information for error analysis because they give only one score for each document or sentence. Thus, evaluation methods have been proposed that present questions to human evaluators asking whether a sentence contains

* This work was conducted as an activity of **AAMT** (Asia-Pacific Association for Machine Translation) working group

grammatical error(s) or other type of errors (Isahara, 1995; Joans et al., 2007; Uchimoto et al., 2007; Nagase et al., 2009). These question-based evaluation methods have several features that enable the provision of new viewpoints for evaluation, the evaluation of grammatical coverage, and the provision of better metrics for evaluation combined with conventional automatic evaluation metrics.

Joans et al. (2007) used the Defense Language Proficiency Test (DLPT), which was originally developed to measure human language skills in their method. Their method enables MT quality to be evaluated from the viewpoint of "comprehension". Isahara (1995) developed the JEIDA test set, which contains example sentences and questions corresponding to each grammatical item, and proposed using it to find grammatical faults in MT systems. Nagase et al. (2009) developed a test set for a Japanese-to-Chinese MT system that expanded the JEIDA test set, and they showed that the test set was effective for analyzing the grammatical problems in Japanese-to-Chinese MT systems. Uchimoto et al. (2007) showed that an evaluation method combining question-based evaluations with conventional automatic evaluations outperforms the conventional automatic evaluation methods. All these works except that by Uchimoto et al. (2007) require human evaluators to answer the questions. Uchimoto et al. (2007) also showed that it was possible to replace question-based evaluation with matching of grammatical patterns with no performance loss. By using the pattern-matching approach, we can conduct question-based evaluation automatically without human evaluators. Uchimoto et al. (2007) aimed to improve automatic evaluation methods to use in scoring a given set of translated sentences, not for error analysis.

In this paper, we show that automatic question-based evaluation based on pattern matching can be used for error analysis, where achievement of a score for each grammatical check point is displayed. This is not obvious because the patterns for automatic evaluation are approximations of the questions for human evaluation. We reveal that the grammatical question-based evaluation can be replaced with a pattern-matching approach for the purpose of grammatical error analysis. Furthermore, we actually evaluated six existing Japanese-to-Chinese MT systems, and show the performance of our method from the viewpoint of grammatical error analysis.

2 Test Set for Evaluating Japanese-to-Chinese MT Quality

In this section, we describe the process of creating questions that enable the quality of Japanese-to-Chinese MT systems to be automatically evaluated. Section 2.1 describes the questions in previous research on manual evaluation, section 2.2 describes the concretization of the questions to reduce ambiguities, and section 2.3 explains the expansion of the questions for automatic evaluations. In this paper, we propose an evaluation method that includes both concretization and expansion of the questions.

2.1 Test Set

The JEIDA test set (Isahara 1995) is used in a method for MT quality evaluation that sheds light on the grammatical problems of MT. This test set is characterized by its yes/no questions for assessing translation results. An example of a test set sample is given below.

(1-4) Complex Predicates	Grammatical Category
私たちは研究開発をする。[1]	Japanese Sentence
We do research and development.	Reference Translations
We are carrying out research and development.	
The words 「研究[2]」 and 「開発[3]」 should be identified as parallel verbs.	Grammatical Check Point

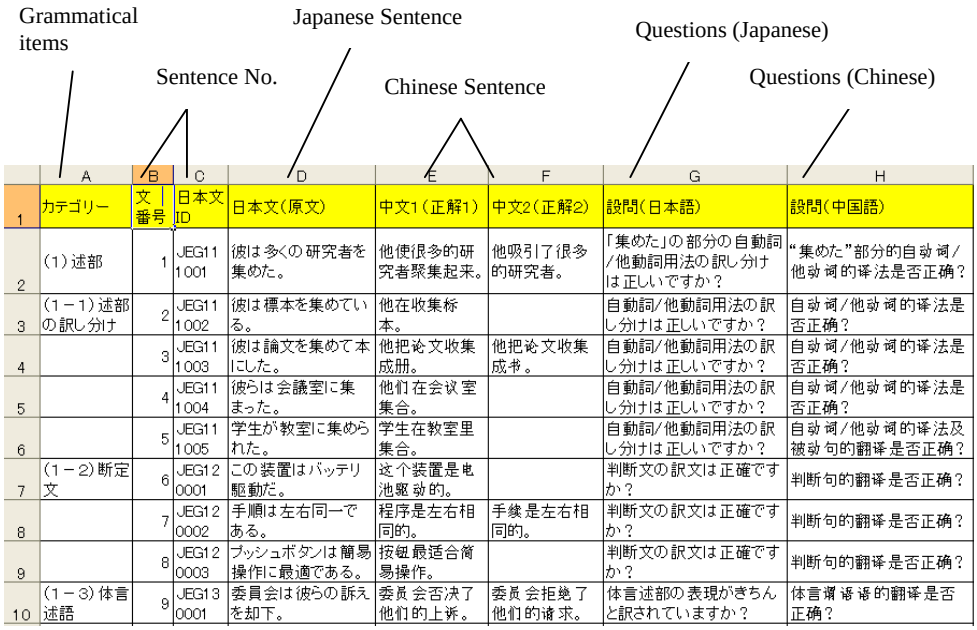
1 "watashitachi wa kenkyukaihatsu wo suru." [we research&development do]
2 "kenkyu" [research]
3 "kaihatsu" [development]

Each example consists of a grammatical category, a Japanese sample sentence, its reference translation(s), and a yes/no question. Here, the yes/no questions are considered to be checked by human evaluators, not by computers.

Our first test set for Japanese-to-Chinese MT evaluation (AAMT 2008/2009, called “ver. 1” in this paper) was made based on the JEIDA test set (Japanese-to-English version) according to the following steps.

1. Translate the Japanese example sentences into Chinese (where a source sentence partially corresponds to two Chinese sentences).
2. Add a yes/no question checking the quality of the Chinese translation to each sample sentence, according to the grammatical point of the example.
3. Translate the yes/no questions into Chinese.

An example from test set ver. 1 is in Fig. 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。	他使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？	体言谓语语的翻译是否正确？

Figure 1: Example of Japanese-to-Chinese test set ver. 1

2.2 Improvement of Test Set

Test set ver. 1 did not contain enough examples; several important Chinese grammatical phenomena could not be checked in the test set because the set was based on a Japanese-to-English test set. For example, the word “not” should be translated as “不 (“bu” in Chinese)” or “没 (“mei” in Chinese)”. Correctly selecting the two words “不” or “没” is important, but this point could not be checked using test set ver. 1. Thus, after review of the test set by Chinese linguists, 251 sentences for 39 newly defined grammatical check points were added to test set ver. 1. The resultant expanded test set is called ver. 2.

Table 1: Size of test set.

Test set	Number of sentences
Ver. 1	325
Ver. 2	576 (additional 251 sentences)

The answer to a question is either “Yes” or “No”, and this answer must be consistent. However, when two native Chinese speakers actually evaluate MT systems using the test set, 25.7% of their answers to the questions differ (AAMT 2008/2009). For this reason, most of the questions are considered to be composed of abstract terms. The question in test set ver. 1 “Is the translation of the idiomatic expression appropriate?” is an example of an abstract question. The answer to this type of question depends on the evaluator’s language skills. Thus, the questions newly added to test set ver. 2 were created based on pattern-matching questions such as “Is the expression xxxxx included in the translation?” When test set ver. 2 is applied, the rate of difference between the two evaluators is reduced to 16.1% from 25.7% (Nagase et al. 2009). This indicates that pattern-matching-based questions help to increase the objectivity of evaluations.

2.3 Expansion of Test Set for Automatic Evaluation

The test set we introduce in this section is an approximate test set with test set ver. 2 Thus, it is not logically ensured that the corresponding questions between an approximate test set and test set ver. 2 are completely consistent. As mentioned in section 2.2, the objectivity of evaluations could be increased through using questions that check whether certain words or phrases are included in the translated text. However, if compatible words (synonyms) are not included in the test set, the evaluators judge the text by their own standards. All synonyms and paraphrases must thus be prepared in advance, and they are added to each question in order to achieve a fully automatic evaluation. Examples of questions before and after rewriting are shown in Table 2.

Table 2: Examples of expanded questions.

Original questions	Expanded questions
「病気で」が「因为生病」で訳されていますか？[4]	「病気で」が「因为生病」または「由于生病」または「因为毛病」または「由于毛病」または「因为病」または「由于病」または「因病」で訳されていますか？[5]
「木で」が「用木头」で訳されていますか？[6]	「木で」が「用木头」または「拿木头」で訳されていますか？[7]

An example of a program (pseudo code) for automatic evaluation is shown in Fig. 2. It represents synonyms and phrases that should exist (or not exist) using regular expressions. Synonyms and paraphrases are exhaustively defined in the program so that an answer to a question is close to a human's evaluation.

One advantage of our method is that tokenization of Chinese sentences, which is required by conventional automatic evaluation methods, is not needed. Chinese word boundaries are very ambiguous and difficult to define, different from in western languages. Therefore, we segmented Chinese sentences by each character to calculate the scores of conventional evaluation metrics in our experiment.

4 Is the phrase “病気で” (byoki-de, [due to illness]) in Japanese translated as “因为生病” (yin wai sheng bin) in Chinese?

5 Is the phrase “病気で” in Japanese translated as either “因为生病”, “由于生病” (you yu sheng bin), “因为毛病” (yin wai mao bin), “由于毛病” (you yu mao bin), “因为病” (in wai bin), “由于病” (you yu bin), or “因病”’(yin bin) in Chinese?

6 Is the phrase “木で” (ki-de, [by wood]) in Japanese translated as “用木头” (yong mu tou) in Chinese?

7 Is the phrase “木で” in Japanese translated as “用木头” or “拿木头” (na mu tou) in Chinese?

```
read a line;
if the line matches /严|严厉|严格|厉害/ then
    print "Yes";
else
    print "No";
endif

read a line;
if the line matches /去买东西|去购物|去逛街|去逛商店/ then
    print "Yes";
else
    print "No";
endif
```

Figure 2: Pseudo codes for automatic evaluation.

3 Experiments and Discussion

3.1 Experimental Method

In this section, we describe the experiment carried out to verify the effectiveness of our automatic evaluation method proposed in section 2. In this experiment, we used the test set consisting of 251 examples (39 grammatical items) that were added for checking grammatical phenomena specific to Chinese. Using our experimental results, we discuss the performance of our automatic evaluation method from the perspective of correlation with a human's objective evaluations, in comparison with those of traditional automatic evaluation methods. Moreover, we discuss the degree of agreement between our proposed method and a human's evaluation (both being question-based). To discuss the above points, the experiment was carried out as follows.

(1) Manual evaluation by human:
- Manual evaluation by human: 2 people (native Chinese speakers)
- Points of evaluation
  - Adequacy: rating on a 1–5 scale by referring to the source sentence and the translation result
  - Fluency: rating on a 1–5 scale by referring to the translation result
  - Question: answering “yes” or “no” to the question by referring to the translation result

(2) Automatic evaluation by program:
- Conventional automatic evaluation metrics based on characters: BLEU, NIST, WER, PER
- Automatic evaluation answering the questions

The two evaluators were native Chinese speakers having linguistic knowledge who had achieved mastery of the Japanese language. Six Japanese-to-Chinese MT systems accessible on the Internet were used as the targets of the experiments. Four out of the six systems use a rule-based MT (RBMT), and two use a statistical MT (SMT).

3.2 Results and Discussion

The coefficients of correlation (Pearson product-moment correlation coefficient) for the MT evaluation scores were examined for six evaluation methods. This was to ensure that our method shows adequate performance compared to conventional automatic evaluation methods. The average score of two evaluators for each sample sentence was used in the comparisons and analyses.

The coefficients of correlation in every combination were very high (Table 3). The reason could be that the lengths of the source sentences were fairly short, in contrast to those in the normal test sets.

Table 3: Coefficients of correlation between automatic evaluations and conventional subjective evaluations.

Methods	Adequacy	Fluency
BLEU	0.9453	0.9588
NIST	0.9783	0.9123
WER	-0.9801	-0.9267
PER	-0.9707	-0.8986
Questions (human)	0.9793	0.9334
Questions (automatic)	0.9902	0.9208

As shown in Tables 3, 4, and 5, our question-based methods highly correlated with subjective evaluations from the viewpoints of fluency and adequacy. The results of a t-test indicate that the correlations between our method and subjective evaluation (fluency and adequacy) achieved statistical significance with a 1% significance level. The correlation coefficients obtained by our methods exceed those of the conventional automatic evaluations. Considering from the above results, our methods may supplement and be a possible alternative to current subjective evaluations. In addition, the performance of our methods compares favorably with that of the conventional automatic evaluations.

Table 4: Frequency distribution of answered pattern (Auto-Human1)

Same	Yes-Yes	425	1181	0.784
	No-No	756		
Different	Yes-No	222	325	0.216
	No-Yes	103		

Table 5: Frequency distribution of answered pattern (Auto-Human2)

Same	Yes-Yes	473	1179	0.783
	No-No	706		
Different	Yes-No	174	327	0.217
	No-Yes	153		

Table 6: Frequency distribution of answered pattern (Human1-Human2)

Same	Yes-Yes	456	1264	0.839
	No-No	808		
Different	Yes-No	170	242	0.161
	No-Yes	72		

Tables 4 and 5 show the number of scoring patterns (“Yes-Yes”, “No-No”, “Yes-No”, and “No-Yes”) scored automatically and manually. Table 6 shows the number of scoring patterns for the two human evaluators. All evaluators must answer a question with either “Yes” or “No”. When the answering pattern is “Yes-Yes” or “No-No”, the pattern type is “the same”; in other cases, it is “different”. Between the automatic evaluation and manual evaluation, the rates of “the same” were 74.8 and 78.3% for “Yes-Yes” and “No-No” respectively. Between the two evaluators, it is 83.9%, which slightly exceeds the rates in Tables 4 and 5. Table 7 shows the

Kappa coefficients calculated using the values in Tables 4, 5, and 6. They are all considerably high, exceeding 0.5.

Table 7: Kappa coefficients between automatic and human evaluation-based questions.

Methods	Kappa coefficient
Automatic : Human (Evaluator-1)	0.5494
Automatic : Human (Evaluator-2)	0.5552
Human (Evaluator-1) : Human (Evaluator-2)	0.6616

The results of a t-test indicate that the agreement of yes/no answering between the automatic evaluation and human evaluation achieved statistical significance with a 1% significance level. This indicates that the automatic question-based method can be used on behalf of human question-based method although the questions of both methods are not exactly the same.

3.3 Evaluation of Grammatical Achievement of Japanese-to-Chinese MT Systems

The question-based evaluation method not only enables MT quality to be quantified but also brings out the strengths and weaknesses of a MT system. The scores of the questions in six Japanese-to-Chinese MT systems (labeled A–F) were averaged, where the answers "Yes" and "No" correspond to the scores "1" and "0", respectively. Figure 3 graphs the evaluation scores with the averages of the six systems according to each grammatical item.

The averages for each system greatly varied among the grammatical items shown in Table 8. For example, most systems were accurate for items No. 9 (sentences including "有" ("*you*" in Chinese)) and No. 10 (adjectival predicates), while almost no systems worked well for items No. 31 (expressions of possibility) and No. 32 (expressions of voluntary) among others. From the low-scored items, there might be problems common among Japanese-to-Chinese MT systems. As a whole, the current Japanese-to-Chinese MT systems are considered to have much room for development because the averages exceeded 0.5 only in 10 items out of the 39.

Comparing the systems, a big difference exists between systems A–D and systems E and F. It is noteworthy that the first- and second-lowest scored systems (E and F) are Statistical MT systems, while the others are rule-based ones. The properties of SMT might limit the improvement of the translation quality. Fifteen and twelve items in systems E and F respectively scored zero, and eleven of these items were the same. In these eleven items, it is possible that the particularly badly scored items, such as No. 15 (expressions of desire), No. 22 (expressions of experience), No. 30 (passive sentences), and No. 37 ("越-越" ("*yue-yue*" in Chinese) according to "すればするほど" ("*sureba suruhodo*" in Japanese) phrases) contain characteristic problems of SMT.

From Fig. 3, we can see that system D obtained a high score for almost all the items, but for No. 34 (honorific) and No. 19 ("了" ("*la*") and "着" ("*zhe*" in Chinese) according to "--ている"("*--teiru*" in Japanese)), the scores were exceptionally below the average for the six systems. In contrast, the scores for systems E and F were relatively bad, but for items No. 16 (prepositions), No. 14 (expressions of permission), and No. 39 (idiomatic phrases), the scores exceeded the average for the six systems.

These data examined above would be quite useful for developers of MT systems. Such data were not obtained by previous automatic evaluation methods. Developers can recognize the strengths and weaknesses of their own systems from the graphs, which can be used as feedback to the developers for deciding the approach to and priority of development candidates.

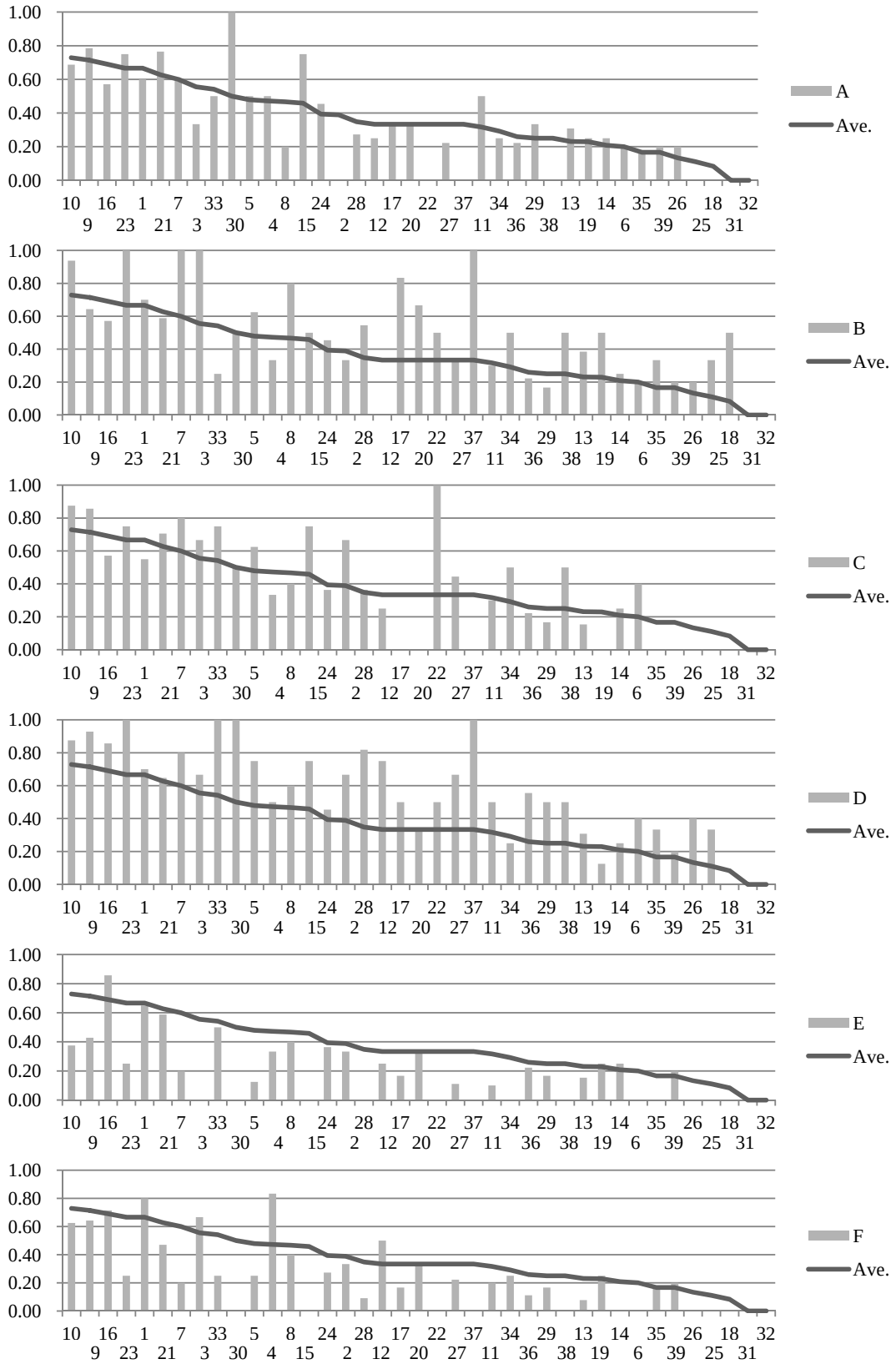


Figure 3: Average scores of question-based evaluation for each grammatical item.

Table 8: Grammatical items.

1	“的” (de) according to “の” (no)	21	Expressions according to “--ている” (--teiru)
2	“的” (de) according to “V た” (da)	22	Expressions of experience
3	“的” (de) according to “もの” (mono)	23	Transitive/Intransitive verbs
4	Location words	24	Particles
5	Interrogatives	25	Sentences ending with “のだ” (noda)
6	Negative sentences	26	Sentences including “把” (ba)
7	Selecting “不” (bu) and “没” (mei)	27	Complements
8	Sentences including “在” (zai)	28	Causative Sentences
9	Sentences including “有” (you)	29	Passive sentences using “被” (bei)
10	Adjectival predicates	30	Passive sentences not using “被” (bei)
11	Pivotal sentences	31	Expressions of possibility
12	Verb phrases with two objects	32	Expressions of voluntariness
13	Auxiliary verbs	33	Respectful phrases
14	Expressions of permission (“可以” (keyi))	34	Honorific expressions
15	Expressions of wish	35	Sentences ending with “ようだ・そうだ” (youda/souda)
16	Prepositions	36	Expressions of solicitation
17	Expressions of comparison	37	“越-越” (yue-yue) according to “すればするほど” (surebasuruhodo)
18	“刚刚” (gang gang) according to “--したばかり” (--shitabakari)	38	Prepositions
19	“着” (zhe) “了” (la) according to “ている” (--teiru)	39	Idiomatic phrases
20	Sentences including “了” (la)		

4 Conclusion and Future Work

We proposed a question-based automatic evaluation method that does not depends on languages, validated out our method in Japanese-to-Chinese translation, and showed its effectiveness by the error analysis of existing six translation systems. Our method has high correlation with subjective evaluations by humans, even when only our method is used (which is comparable to conventional automatic evaluation methods and question-based evaluation by humans). We verified that the results of our automatic evaluation method are consistent with that of human evaluation for each question, and clarified that our method can be used for error analysis of grammatical items. In particular, by visualizing the scores for each grammatical item on a graph, valuable information related to the characteristics of each MT system became visible. We are planning to provide a web service that automatically evaluates MT quality and to provide information on the strengths and weaknesses of an MT system through a cobweb chart and others.

The automatization of question-based evaluation yields not only increased efficiency due to the reduced number of evaluators, but also enhanced reliability of the evaluation by improving objectivity. In an evaluation using a question-based method, 16.1% of the answers of two evaluators differed. However, in the case of automatic evaluation, the answers are always the same. Our experiment indicated that a question-based automatic evaluation method could be closer to a human subjective evaluation than a question-based human evaluation.

Problems of a question-based evaluation method are that sentences must be prepared that include grammatical items for conducting error analysis, and that questions must be prepared that include the words to be checked as well as sets of synonyms and paraphrases. If these preparations could be done automatically, the efficiency of the evaluation process would be

dramatically enhanced. In addition, the grammatical items, which are useful for error analysis, are also defined by a human. In the future, a method for automatically preparing test sentences, the grammatical items, and their questions should be investigated.

References

AAMT Working Group. 2008. Toward Building of AAMT Test-sets for Quality Evaluation of Machine Translation. *AAMT Journal No.42*, 18-24.

AAMT Working Group. 2009. Toward Publication of AAMT Test-sets for Quality Evaluation of Machine Translation. *AAMT Journal No.45*, 29-30.

Banerjee, S. and A. Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.

Birch,. A. and M. Osborne, Reordering Metrics for MT. 2011. *Proc. of ACL-HLT*, 1027–1035.

Doddington G. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-occurrence Statistics. *Proc. of HLT*, 138–145.

Isahara, H. 1995. JEIDA's Test-Sets for Quality Evaluation of MT Systems --Technical Evaluation from the Developer's Point of View--. *Proceedings of MT Summit V*.

Isozaki, H., T. Hirao, K. Duh, K. Sudoh and H. Tsukada. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. *Proc. of EMNLP*, 944–951.

Melamed, D., Green, R., and Turian, J. P. 2007. Precision and Recall of Machine Translation. *Proc. of NAACL-HLT*, 61–63.

Nagase, T., K. Kotani, M. Nagata, N. Hatanaka, Y. Sakamoto, E. Sumita and K. Uchimoto. 2009. Evaliation of Japanese-Chinese MT System using AAMT's Test-Set. *Proc. Of The 5th China Workshop on Machine Translation*, 211-218.

Papineni K., S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proc. of ACL*, 311–318.

Snover, Matthew., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. *Proc. of AMTA*, 223–231.

Uchimoto, K., K. Kotani, Y. Zhang and H. Isahara. 2007. Automatic Evaluation of Machine Translation Based on ate of Accomplishment of Sub-goals. *Proceedings of NAACL HLT*, 33-40.