Experimental Comparison of MT Evaluation Methods: RED vs. BLEU

Yasuhiro Akiba[†,‡], **Eiichiro Sumita**[†], **Hiromi Nakaiwa**[†],
Seiichi Yamamoto[†], **and Hiroshi G. Okuno**[‡]

† ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Keihana Science City, Kyoto 619-0288, Japan
‡ Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
{yasuhiro.akiba, eiichiro.sumita, hiromi.nakaiwa}@atr.co.jp
seiichi.yamamoto@atr.co.jp, okuno@i.kyoto-u.ac.jp

Abstract

This paper experimentally compares two automatic evaluators, RED and BLEU, to determine how close the evaluation results of each automatic evaluator are to average evaluation results by human evaluators, following the ATR standard of MT evaluation. This paper gives several cautionary remarks intended to prevent MT developers from drawing misleading conclusions when using the automatic evaluators. In addition, this paper reports a way of using the automatic evaluators so that their results agree with those of human evaluators.

1 Introduction

This paper addresses the problem of automating the human evaluation of machine translation (MT) systems, especially speech-to-speech machine translation (S2SMT) systems. With advances in techniques for automatically learning translation knowledge from parallel corpora, MT developers are requiring rapid MT evaluation methods now more than ever. A rapid MT evaluation method makes it easier for the MT developers to evaluate whether their new ideas for constructing MT systems are effective. As a criterion for human evaluation, this paper adopts the ATR standard of MT evaluation[1], which was proposed for the human evaluation of S2SMT systems.

Two promising automatic evaluators were simultaneously proposed: BLEU (bilingual evaluation understudy (Papineni et al., 2001)) and RED (Ranker based on Edit Distances (Akiba et al., 2001)) as rapid MT evaluation methods. BLEU was designed to evaluate the performance of MT systems on a test suite. On the other hand, RED was designed to evaluate a translation segment such as an utterance or a sentence. As mentioned in (Akiba et al., 2001), RED can also be used to evaluate the performance of MT systems by summing all of the evaluation results of translations for the test suite.

Both of these automatic evaluators were reported to be able to closely simulate the human evaluation of MT systems. BLEU (Papineni et al., 2001) was shown to correlate highly to two human evaluation measures of translation, adequacy and fluency, which are graded from 1 (very bad) to 5 (very good). RED (Akiba et al., 2001) was shown to approximate the performance of an MT system, as measured by the ATR standard of MT evaluation.

Although many researchers have insisted on the quality of their novel approaches for developing MT systems by the evaluation results of BLEU (Yamada and Knight, 2002; Och and Ney, 2002; Marcu and Wong, 2002) in their papers, we have doubts over whether the automatic evaluators, including BLEU, are perfectly or significantly reliable as a basis to form conclusions in research papers. As the regression results in (Papineni et al., 2001) indicate, for example, a small but distinct gap exists between the evaluation results by BLEU and those by human evaluators. This gap highlights the possibil-

[1] ATR's standard of MT evaluation (Sumita et al., 1999) is defined as follows: (A) Perfect: no problems in either information or grammar; (B) Fair: easy-to-understand, with either some unimportant information missing or flawed grammar; (C) Acceptable: broken, but understandable with effort; (D) Nonsense: important information has been translated incorrectly.

ity that BLEU may cause some misleading conclusions. To prevent either developers or researchers of MT from drawing misleading conclusions when using the automatic evaluators, it is essential that they learn either the proper usage or the functional limits of ability of the automatic evaluators by conducting as many experiments as possible.

Accordingly, the authors compared the three evaluation methods: BLEU, RED, and human evaluation according to the ATR standard of MT evaluation. In this comparison, we evaluated eighteen MT systems: nine Japanese-to-English (JE) MT systems and nine English-to-Japanese (EJ) MT systems, which are subsystems of S2SMT systems. The human evaluations were carried out by nine native speakers of the target language who are also familiar with the source language. Each native speaker assigned to each translation one of four ranks[1]: A, B, C or D. Sixteen reference translations (Papineni et al., 2001) or multiple standards (Akiba et al., 2001; Thompson, 1991) were prepared for this experiment.

The main lessons from the comparison are as follows:

- The evaluation results by RED are close to the average evaluation results by humans. The ratio at which RED agrees with the average evaluation by humans is in the range of 90s%, even when different types of MT systems are compared.
- The ratio at which BLEU agrees with the average evaluation by humans reaches 100%, but only when the same type of MT systems are compared by using some sets of reference translations, because BLEU is very sensitive to the choice of reference translations.

In Section 2, the authors outline both of the two automatic evaluators and briefly compare and contrast their basic features. Experimental results are shown and a discussion is provided in Section 3. Finally, our conclusions are presented in Section 4.

2 Outline of Automatic MT Evaluation Methods

This section outlines the two automatic evaluators: BLEU and RED. We then briefly compare and contrast their basic features.

2.1 RED

RED (Akiba et al., 2001) is an automatic ranking method based on edit distances to multiple reference translations[2]. RED consists of a learning phase and an evaluation phase. In the learning phase, in order to estimate a rank[1] from edit distances, RED learns a Decision Tree for the ranking (hereafter, "ranker") from ranking examples by a Decision Tree learner (Quinlan, 1993). In the evaluation phase, RED assigns a rank to each MT output by using the ranker.

Each ranking example is encoded by using multiple edit distances and a median rank[3] among the ranks assigned by three or more human evaluators. On the other hand, each translation to be ranked is encoded by using only multiple edit distances before being assigned a rank.

Each edit distance is measured by one of sixteen variations of the basic edit distance measure, ED1, with three edit operators: insertion, deletion and Replacement. For ED1, two morphemes are regarded as being matched if and only if the base form of each morpheme is the same and each POS tag is the same. For the remaining edit distances, their definitions are changed due to a combination of the following four changing policy. The first policy is that whether the swap edit operator is additionally used. The second is that whether semantic codes of content words are referred instead of the base forms of the content words. The third is that whether the editing units are restricted to only content words. The last is that whether the editing units are restricted to only keywords. The keywords is defined[4] as words that appear in two or more of the reference translations. Both the Ranking examples and the translations to be ranked are encoded by using all the sixteen variations.

2.2 BLEU

BLEU (Papineni et al., 2001) is an automatic scoring method based on n-gram matching with multiple reference translations. BLEU measures the pre-

[2]This is called "multiple standards" in (Thompson, 1991).

[3]This is described in (Akiba et al., 2001) as "majority rank".

[4]This definition is changed from the original one (Akiba et al., 2001) such that RED works better. Furthermore, note that even functional/non-content words can become keywords.

Table 1: Differences between RED and BLEU

	RED	BLEU
Evaluation unit	An utterance	A segment
Evaluation target	An utterance	A document
Evaluation results	Ranking	Scoring
Learning strategy	Supervised	Not learning
Approach	Edit distances	N-gram matching
Robustness to replacing or swapping words	Relatively weak	Strong
Long-distance co-occurrence	Strong	Weak

cision of unigrams, bigrams, trigrams and 4-grams with respect to a whole set of reference translations with a penalty for sentences that are too short. High BLEU scores are better.

2.3 Basic Features of RED and BLEU

In this section, we briefly compare and contrast the basic features of the two automatic evaluators. A similar feature of the automatic evaluators is that they both use reference translations.

Table 1 shows the different features of the automatic evaluators. The evaluation unit of RED is a sentence or an utterance, whereas that of BLEU is a segment such as a sentence or a paragraph. Thus, the evaluation unit of BLEU can be relatively longer. The evaluation target of RED is a sentence or an utterance, whereas that of BLEU is a test suite or a document.

RED expresses each evaluation result by a rank, whereas BLEU does so by a score. In the case where ranks follow an evaluation standard, such as the ATR standard of MT evaluation, the result of RED can be interpreted. On the other hand, BLEU's result, other than zero or one, cannot be interpreted in any case. The learning strategy of RED is supervised learning, while BLEU is not learning-based.

RED's approach is based on edit distances, whereas BLEU's is based on n-gram. Consequently, RED can handle long-distance co-occurrence; however, its robustness for replacing or swapping words is relatively weak, whereas BLEU's is strong. On the other hand, BLEU cannot handle long-distance co-occurrence.

3 Experiment

The authors evaluated the two automatic evaluators to answer the following questions:

1. How accurately does each automatic evaluator order the different types of MT systems in the same way as average evaluation by humans does?

2. How does an increase in the number of reference translations affect the accuracy of the ordering of MT systems?

3. How accurately does each automatic evaluator order the same type of MT systems?

3.1 Experiment Resources

This section describes the experiment resources used: test suite, evaluated MTs, reference translations and human evaluation results.

Test Suite: The test suite used consists of three hundred and forty-five pairs of English and Japanese sentences, which were randomly selected from the Basic Travel Expression Corpus (BTEC) (Takezawa et al., 2002). BTEC contains a variety of expressions used in a number of situations related to overseas travel.

Reference Translations: The authors asked five native speakers of Japanese who are familiar with English to translate English sentences in the test suite in three ways. Consequently, there were sixteen reference translations in Japanese, including Japanese sentences in the test suite. The reference translations in English for the test suite were prepared in the same way as those in Japanese.

Evaluated MTs: The MT systems used were the nine EJ MT systems of three types and the nine JE MT systems of three types described below. The EJ MT systems were three versions of EJ TDMT (0103, 0110, 0203)[5], three versions of EJ HPAT (0110, 0204, 0209), and three versions of EJ SAT (0110, 0204, 0209). The JE MT systems were three versions of JE TDMT (0103, 0110, 0203), three ver-

[5]Numbers such as '0103' indicate the time when the MT system was released. For example, '0103' means the MT system was released in March of 2001.

Table 2: Correlation between two independent vectors of median ranks. The correlation is calculated as follows: in the case where N equals three, (1) randomly select two disjoint subsets of three human evaluators, (2) for each subset of three human evaluators, give each MT output the median of the ranks that the three human evaluators assigned to the MT output, and (3) for each subset, gather the resulting median ranks and make a vector so that each element corresponds to the same MT output.

	# of human evaluators (N)			
	1	2	3	4
Correlation	0.736	0.896	0.901	0.946

sions of JE D^3(0110, 0204, 0209), and three versions of JE SAT (0110, 0204, 0209).

TDMT, HPAT, SAT, and D^{3}are different types of MT systems (Sumita, 2002). Each version of **TDMT** (Transfer Driven Machine Translation) is a **pattern-based** MT system[6] using **hand-coded** syntactic transfer rules (Furuse and Iida, 1996). Each version of **HPAT** (Hierarchical Phrase Alignment-based Translation) is a **pattern-based system** using **automatically-generated** syntactic transfer (Imamura, 2002). Each version of **SAT** (Statistical ATR Translator) is an **SMT** (Statistical machine translation) system using hierarchical phrase alignment (Watanabe et al., 2002). Each version of **D**^{3}(DP-match Driven transDucer) is an **example-based** MT system using **online-generated** translation patterns, which are close to **templates** (Sumita, 2001).

Human evaluation results: Nine English translations of each Japanese sentence by the JE MTs were simultaneously shown to each of nine JE translators that were native speakers of English, to keep the evaluation results consistent. These translations were then evaluated according to the ATR standard of MT evaluation. In the same way, nine Japanese translations were evaluated by nine EJ translators. Each translation was finally assigned the median rank among its nine ranks. Table 2 shows how the median ranks become more consistent as the number of human evaluators increased. Each fractional number is the correlation coefficient between two independent vectors of median ranks, each of which were calculated from one of two disjoint sets of human evaluators selected randomly. Even in the case where the number of human evaluators is four, the correlation coefficient reached the large number of 0.946. The median ranks among the nine ranks are used in this paper, and have a high correlation coefficient of at least 0.946.

[6]Some researchers classify these as example-based MT systems.

Figure 1: Results of a paired comparison (Sugaya et al., 2001) between translations by an MT system and those by a person whose TOEIC score is known. "$X < Y$" denotes that Y is superior to X. Each number, such as 420, denotes that person's TOEIC score.

SAT-0203	$<$	Person-420	$<$	Person-540	$<$
Person-685	$<$	TDMT-0203	$<$	Person-820	$<$
D^3-0203	$<$	Person-965			

3.2 Design of Experiment

This section gives three critical points to consider in the comparison of the two automatic evaluators. The first point is related to the dependency of the automatic evaluation methods on both the set of reference translations used and the test suite used. The second is related to the way of selecting and applying a statistical test. The third point is related to the way of judging whether an automatic evaluator is correct.

Let us consider the evaluation of three MT systems: TDMT-0203, SAT-0203, and D^3-0203. Figure 1 shows the superiority or inferiority of each MT system to a person whose TOEIC score is known. The superiority or inferiority is judged by using a paired comparison (Sugaya et al., 2001), which is commonly used in the discipline of psychology as a stable human evaluation method.

The TOEIC[7] (Test of English for International Communication) as well as TOEFL (Test of English as a Foreign Language) tests were both created by ETS[8] (Educational Testing Service). A TOEIC score's standard error of measurement is known to be 25 points. SAT-0203 is inferior to a person that has a TOEIC score of 420 and TDMT-0203 is superior to a person that has a TOEIC score of 685. The difference in the TOEIC scores, 420 points and 685 points, is more than 200, which is significantly different. In addition, a comparison of the number of correct translations by each person and TOEIC score yields a high correlation coefficient of 0.97.

[7]http://www.toeic.com

[8]http://www.ets.org

Table 3: Dependency of BLEU score on reference translations. Fractional numbers in the ith line give the BLEU scores of the corresponding MT systems that were calculated for the test suite by using the ith set of reference translations. The number of reference translations is the same, i.e., four. Each reference translation is randomly selected from among the whole set of sixteen reference translations. The integers in parentheses express the order of BLEU scores within each line.

Ref. subset	MT systems		
	D^3-0203	SAT-0203	TDMT-0203
No.1	0.3333 (1)	0.2350 (2)	0.2267 (3)
No.2	0.3546 (1)	0.2392 (2)	0.2260 (3)
No.3	0.3439 (1)	0.2331 (**3**)	0.2347 (**2**)
No.4	0.3545 (1)	0.2414 (2)	0.2286 (3)
No.5	0.3487 (1)	0.2417 (2)	0.2353 (3)

This means that the test suite was designed well enough that the people are ordered in the same order as the TOEIC score. Therefore, we are justified in claiming that TDMT-0203 is significantly superior to SAT-0203.

Further discussion of experimental design focuses on three critical points as follows.

First point: Table 3 shows the dependency of BLEU on the set of reference translations used. Even when the number of elements in a set of reference translations was the same, BLEU ordered the three MT systems in different ways.

Let us consider a situation in which Mr. A prepared the third set of reference translations by chance. Mr. A would plan to carry out a statistical test in order to test the difference between the BLEU scores of SAT-0203 and TDMT-0203. There exists the possibility of rejecting the NULL hypothesis and of concluding that TDMT-0203 is significantly superior to SAT-0203.

Let us consider another situation in which Mr. B prepared the first set of reference translations by chance. Mr. B would plan to carry out a statistical test in order to test the difference between the BLEU scores of SAT-0203 and TDMT-0203. There is no possibility of concluding that TDMT-0203 is significantly superior to SAT-0203 because the BLEU score of TDMT-0203 is less than that of SAT-0203.

Mr. A would think that BLEU worked well. On the other hand, Mr. B would think that BLEU was capable of making a mistake. The lesson from the above consideration is that one ought to consider the dependency of BLEU on the set of reference translations used when testing the difference in the BLEU scores of two MT systems. For the same reason, one ought to consider the dependency of BLEU on the test suite used when testing the difference in the BLEU scores of two MT systems. These lessons also hold true for RED.

Therefore, one ought to consider the dependency of the automatic evaluation methods on both the set of reference translations used and the test suite used. From this perspective, the authors carried out an experimental comparison through a statistical test in consideration of various combinations of a subset of the test suite and a subset of the whole set of reference translations.

Second point: To compare the two automatic evaluators, one needs to carry out multiple statistical tests. It is well known that repeating a pairwise test, such as a simple t-test, multiple times decreases the total confidence level. For example, even if the confidence level of each statistical test is equal to 0.95, if the t-test is repeated 10 times, the total confidence level drops to around 0.6 (= 0.95^{10}).

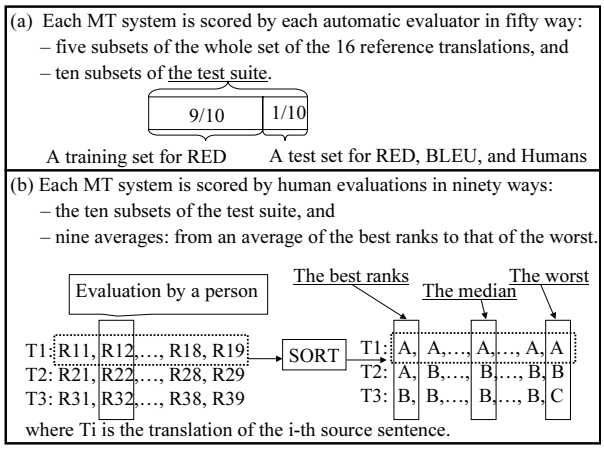
To maintain the total confidence level, the authors used a non-parametric multiple comparison test, the Tukey-Kramer-type multiple comparison test with the Kruskal-Wallis test (Hochberg and Tamhane, 1983). The multiple comparison test is designed to avoid a decrease in the confidence level. The Kruskal-Wallis test is a non-parametric one-way analysis of variance (ANOVA). This test does not assume the data distribution.

Third point: The basic criterion of judgment for whether the automatic evaluation results are correct is based on the agreement with the average evaluation results by humans. That is, if and only if the automatic evaluation results agree with the average evaluation results by humans, the automatic evaluators are regarded as correct.

One approach to obtaining the average evaluation results repeatedly mentioned in this paper is to correspond the ranks A, B, C, and D to 4, 3, 2, and 1, respectively, and to calculate the difference of the averages of the median ranks assigned to the translations of the MT systems.

As we learned from the discussion on the correlation coefficient in Section 3.1, the variance of the median rank is very small, but maybe not zero. To

Figure 2: The scoring of each MT system, (a) by the automatic evaluators and (b) the human evaluators



be fail-safe, the difference in the averages was also statistically tested by using the multiple comparison test instead of just checking whether the difference is positive, zero, or negative.

3.3 Experimental Procedure

To answer the question mentioned at the beginning of Section 3, the author used the following experimental procedure on the basis of the above considerations of the three critical points.

1. Prepare ten disjointed subsets of the test suite according to ten-fold cross validation (Mitchell, 1997). The subsets are the shared test sets for all of the evaluation methods.

2. For each i (i = 1, 4, 7, 10, or 13), five subsets of i reference translations are randomly selected from the sixteen reference translations (See Section 3.1). These subsets are shared by the automatic evaluators.

3. For each pair of a subset of the test suite and a subset of reference translations, both the BLEU score and the RED score of each MT system are calculated (Figure 2 (a)). The RED score is calculated as follows: (i) assign each translation by the MT system to a rank by using the relevant ranker corresponding to the MT system. (ii) To correspond A, B, C, and D to 4, 3, 2, and 1, respectively, calculate an average of values corresponding to the ranks assigned the ranker. The ranker is learned from the complementary set of the subset of the test suite (Section 2.1). The class label of each translation in the complementary set is the assigned median rank (Section 3.1).

4. For each size of the set of reference translations, statistically test the differences of the median of the fifty resulting scores of each MT system at a confidence level of 95% by using the multiple comparison test.

5. Count the number of agreements with the average evaluation results by humans. The average evaluation results by humans are calculated as follows (Figure 2 (b)): (i) Sort the nine ranks of each translation in the best-to-worse order. (ii) Calculate nine averages, from the average of the best ranks to the average of the worst ranks on a set of translations corresponding to each of ten disjointed subsets of the test suite. (iii) Statistically test the differences in the median of the ninety resulting averages for each MT system at a confidence level of 95% by using the multiple comparison test.

3.4 Experimental Results and Discussion

This section answers the three questions presented at the beginning of Section 3.

Table 4 shows the comparison results of the average evaluation by humans and each automatic evaluator. This setting corresponds to the situation where researchers compare different types of MT systems.

Note that in these table captions, all of the MT systems are described as code names, such as MT_1, rather than their system names, in order that we get rid of our prejudice against the MT systems. Each column corresponds to the counts of agreement when the number of reference translations is equal to a certain integer. The fractional number is the ratio of agreement.

With respect to the first question, the ratio of agreement of RED with the average evaluation by humans was in the range of 90s%, whereas that of BLEU was in the range of 30%-60%. The ratio of agreement for JE MT systems tends to be equal to or better than that for EJ MT systems. This is because the target language being more flexible in word order needs more reference translations. In fact, the

Table 4: # of agreements between average evaluation by humans and automatic evaluation both for English-to-Japanese (EJ) nine MT systems and for Japanese-to-English (JE) nine MT systems. The value of X indicates how the multiple comparison test judges significant differences between the performances of MT_i and MT_j ($i > j$) based on average evaluation by humans. When X is equal to 0, MT_i is significantly superior to MT_j. When X is equal to 1, MT_i is not significantly different from MT_j. When X is equal to 2, MT_i is significantly inferior to MT_j. The value of Y indicates how the multiple comparison test judges significant differences between the performances of MT_i and MT_j based on an automatic evaluation.

	RED						BLEU					
		# of reference translations						# of reference translations				
	XY	1	4	7	10	13	XY	1	4	7	10	13
	00	21	20	19	20	20	00	0	0	0	0	0
	11	10	12	11	11	11	11	10	10	10	10	10
	22	2	2	2	2	2	22	1	1	1	1	1
EJ	Ratio	91.6	94.4	88.9	91.6	91.6	Ratio	30.6	30.6	30.6	30.6	30.6
	00	17	16	17	17	17	00	6	9	10	11	11
	11	9	10	9	9	9	11	9	5	6	5	5
	22	7	9	9	8	8	22	5	7	7	7	8
JE	Ratio	91.7	97.2	97.2	94.4	94.42	Ratio	55.6	58.3	63.9	63.8	66.7

Table 5: # of agreements between average human evaluation and automatic evaluation on each type of the three MT systems.

RED						BLEU					
	# of reference translations						# of reference translations				
3MTs	1	4	7	10	13	3MTs	1	4	7	10	13
Three version of EJ TDMTs	**3**	**3**	**3**	**3**	**3**	Three version EJ TDMTs	**3**	**3**	**3**	**3**	1
Three version EJ HPATs	**1**	1	1	1	2	Three version EJ HPATs	**3**	2	2	2	2
Three version EJ SATs	**3**	2	**2**	2	1	Three version EJ SATs	1	1	**3**	1	2
Three version JE TDMTs	**3**	**3**	**3**	**3**	**3**	Three version JE TDMTs	**3**	**3**	**3**	**3**	**3**
Three versionJE D^3	2	**3**	**3**	**3**	**3**	Three versionJE D^3	2	1	0	0	0
Three versionJE SAT	**3**	**3**	2	**3**	**3**	Three versionJE SAT	1	2	2	2	2

word order in Japanese is more flexible than that in English.

The high ratio of agreement of RED is due to RED's function that edit distances used by RED can refer the base forms, POSs, or semantic codes and that the editing unit used by RED can be restricted to the content words or keywords, as described in Section 2.1. On the other hand, BLEU has no such functions and just considers word n-grams. Consequently, when particles in Japanese translation or articles in English translation are wrong or missed, the bigram precision as well as BLEU score tend to be low, even if content words are correctly translated. In such a case, BLEU tends to evaluate MT systems lower than humans do.

The high ratio of agreement of RED is also due to training data; however the cost of generating training data is expensive. When a large-scale evaluation is expected, RED has the potential to help the evaluation in the following ways: 1) carry out a reasonably scaled subjective evaluation, 2) learn the ranker for RED, and 3) carry out the actual evaluation by using RED. Note that the accuracy of BLEU is worse than that of RED, although the advantage of BLEU over RED is that it needs no training data.

With respect to the second question, the increase in the number of reference translations does not necessarily improve the agreement ratio of RED and BLEU uniformly. However, the agreement ratio in the case of multiple reference translations is better than that in the case of only one reference translation. Four reference translations are reasonable for both cost and effect.

With respect to the third question, Table 5 shows the comparison results of the average evaluation by humans and each automatic evaluator for each type of the three MT systems. This setting corresponds to the situation of developing a series of MT systems. In this situation, MT system developers are interested in the improvement of the MT system. RED tends to work better than BLEU in the same way as in the comparison of the different type of MT systems. The agreement ratio of BLEU becomes 100% in the case that BLEU uses some sets of reference translations, because BLEU is very sensitive to the choice of reference translations.

4 Conclusions

This paper experimentally compared the automatic evaluators RED and BLEU to determine how close the evaluation results by each automatic evaluator are to the average evaluation results by human evaluators, following the ATR standard of MT evaluation. The main lessons from the experiments are:

- The evaluation results by RED are close to the average evaluation results by humans. The ratio at which RED agrees with the average evaluation by humans is in the range of 90s%, even when different types of MT systems are compared.
- The ratio at which BLEU agrees with the average evaluation by humans reaches 100%, but only when the same type of MT systems are compared by using some sets of reference translations, because BLEU is very sensitive to the choice of reference translations.

So far, the methodology of constructing reference translations has not been extensively examined, although it is a very important issue. In the future, the authors plan to pursue a study that will establish this methodology as a research topic.

Acknowledgment

This research was supported in part by the Telecommunications Advancement Organization of Japan. The authors thank Kadokawa-shoten for providing them with a Ruigo-shin-Jiten dictionary.

References

Y. Akiba, K. Imamura, and E. Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. In *Proc. MT Summit VIII*, pages 15–20.

O. Furuse and H. Iida. 1996. Incremental translation utilizing constituent boundary patterns. In *Proc. COLING96*, pages 412–417.

Y. Hochberg and A. C. Tamhane. 1983. *Multiple Comparison Procedures*. Wiley.

K. Imamura. 2002. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT. In *Proc. TMI2002*, pages 74–84.

D. Marcu and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. ACL2002 Workshop on EMNLP*, pages 133–139.

T. M. Mitchell. 1997. *Machine Learning*. New York: The McGraw-Hill Companies Inc.

F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proc. ACL2002*, pages 295–302.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, IBM research report rc22176 (w0109-022). Technical report, IBM Research Division, Thomas, J. Watson Research Center.

J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

F. Sugaya, K. Yasuda, T. Takezawa, and S. Yamamoto. 2001. Precise measurement method of a speech translation system's capability with a paired comparison. In *Proc. MT Summit VIII*, pages 345–350.

E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai. 1999. Solutions to problems inherent in spoken-language translation: The atr-matrix approach. In *Proc. MT Summit VII*, pages 229–235.

E. Sumita. 2001. Example-based machine translation using DP-matching between work sequences. In *Proc. the ACL 2001 Workshop on DDMT*, pages 1–8.

E. Sumita. 2002. Corpus-centered computation. In *Proc. the ACL 2002 Workshop on Speech-to-Speech Translation: Algorithms and Systems*, pages 1–8.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. LREC2002*, pages 147–152.

H. S. Thompson. 1991. Automatic evaluation of translation quality: Outline of methodology and report on pilot experiment. In *Proc. the Evaluator's Forum*, pages 215–223.

T. Watanabe, K. Imamura, and E. Sumita. 2002. Statistical machine translation system based on hierarchical phrase alignment. In *Proc. TMI2002*, pages 188–198.

K. Yamada and K. Knight. 2002. A decoder for syntax-based statistical MT. In *Proc. ACL2002*, pages 303–310.