Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum

Hiroshi Echizen-ya
Dept. of Electronics and Information
Hokkai-Gakuen University
S 26-Jo, W 11-Chome, Chuo-ku
Sapporo, 064-0926 Japan
echi@eli.hokkai-s-u.ac.jp

Kenji Araki
Graduate School of Information Science
and Technology
Hokkaido University
N 14-Jo, W 9-Chome, Kita-ku
Sapporo, 060-0814 Japan
araki@media.eng.hokudai.ac.jp

Abstract

In this paper, we propose a new automatic evaluation method applicable to machine translation. Our method specifically examines the length and position of the common parts between two sequences. First, the common parts continuum is determined using the length and position information of common parts in two sequences. That is, the most intuitive common parts continuum is obtained using this process. Moreover, our method recursively repeats this process to control the difference of the common part order. In this repetition process, a greater penalty is assessed on the intuitive common parts continuum as the number of repetitions increases. We call this method Recursive Acquisition of **I**ntuitive com**M**on **PA**rts **C**on**T**inuum (IMPACT). The evaluation results show that the IMPACT score better correlates with human judgment in both adequacy and fluency than the scores of some other automatic evaluation methods.

1 Introduction

Various approaches have been proposed for automatic evaluation of machine translation. Two features are salient among the studies of the literature. We can use the proposed evaluation system facilities using some references. In addition, we can apply it to translation of sentences of various languages. The scores obtained using previous methods correlate with those produced by human judgment. However, from the perspective of the word order and sentence level structure, their methods are insufficient. One of those methods, BLEU (Papineni et al., 2002), is an n-gram co-occurrence based approach. Using a unigram-based measure, GTM (Turian et al., 2003) specifically examines the lengths of common words. Neither BLEU nor GTM can deal sufficiently with the difference between the correct position words and wrong position words, even though they can use all common parts. In addition, ROUGE-L and ROUGE-W are methods based on the longest common subsequence (LCS). These methods can better distinguish the correct position words and wrong position words compared to BLEU and GTM. However, they are insufficient because they cannot use all the common parts. In contrast, ROUGE-S (Lin and Och, 2004) deals with any pair of words in their sentence order by allowing for arbitrary gaps. However, that method cannot accommodate common words that are less than a sufficient S skip distance from the perspective of the word order and sentence level structure.

We propose a new automatic evaluation method of machine translation to resolve those problems presented by other methods. Our method uses the length and position information of common parts between two sequences. First, our method uses LCS to obtain several candidates of the common parts continuum between two sequences. Using the length and position information of the each common part, our method determines only that with the most intuition common parts continuum among the several candidates of common parts continuum between two sequences. Moreover, it recursively obtains a new intuition common parts continuum after the determined common parts con-

tinuum has been excluded. In such a case, a greater penalty is assessed upon the new common parts continuum as the number of repetition increases. Consequently, when the wrong position common parts are included in a sequence, our method can distinguish them. That is, our method can accommodate the word order and sentence level structure without ignoring common parts. We call our method Recursive Acquisition of **I**ntuitive Com**M**on **PA**rts **C**on**T**inuum (IMPACT). Experimental results indicate that the IMPACT score better correlates with human judgment in terms of both adequacy and fluency than some other automatic evaluation methods.

2 Related Work

In previous methods, it is difficulty dealing with cases of word order that uses all common parts. For example, we can discuss this matter based on the following sentences.

A. doctor cured the Japanese
B. doctor cure the Japanese
C. the Japanese cure doctor
D. the Japanese doctor cured
E. Japanese cured the doctor

For those sentences, BLEU-2, which is based on bigram matches, does not distinguish B and C. That is, the score between A and B is equal to the score between A and C because the common words are "*doctor*", "*the*", and "*Japanese*" in the unigram and the consecutive common word is "*the Japanese*" in bigram. Neither ROUGE-L nor ROUGE-W distinguishes C and D; both are based on LCS. In A and C, they use only "*the Japanese.*" Therefore, the LCS between A and C is 2. The other common word "*doctor*" is ignored because the word order of "*the Japanese*" is different from the word order of "*doctor*" between A and C. In A and D, they use only "*the Japanese*" or "*doctor cured.*" Consequently, the LCS between A and D is also 2. Ultimately, ROUGE-L and ROUGE-W do not distinguish C and D. Actually, GTM, which uses the "maximum matching" concept, does not distinguish B and C. The common parts between A and B are "*doctor*" and "*the Japanese;*" the common parts between A and C are also "*doctor*" and "*the Japanese,*" which means that the word order information is not used sufficiently in GTM. Moreover, ROUGE-S2, which is based on the two skip-bigram co-occurrence statistics measure, does not distinguish C and E. In addition, C has a one skip-bigram match with A ("*the Japanese*"); E also has one skip-bigram match with A ("*cured the*"). That is, ROUGE-S cannot deal with common words with less than a sufficient S skip distance.

3 IMPACT: Our Method

3.1 Outline of IMPACT

The intuitive common parts continuum using LCS is determined with IMPACT. Moreover, a new intuitive common parts continuum is determined recursively after the old intuitive common parts continuum is excluded. A penalty is assessed to the new intuitive common parts continuum as the number of repetitions. In the two sequences X = $x_1, x_2, ...x_m$ and Y = $y_1, y_2, ...y_n$, the IMPACT score (We call this **Im**proved **P**ractical: IP) is calculated as the following.

$$R_{IP} = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in CC} length(c)^\beta \right)}{m^\beta} \right)^{\frac{1}{\beta}} \quad (1)$$

$$P_{IP} = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in CC} length(c)^\beta \right)}{n^\beta} \right)^{\frac{1}{\beta}} \quad (2)$$

$$IP(X,Y) = \frac{(1+\gamma^2) R_{IP} P_{IP}}{R_{IP} + \gamma^2 P_{IP}} \quad (3)$$

In those equations, α is the weight based on the repetition number i of the recursive process of the intuitive common parts continuum (α is under 1.0). This parameter is extremely important to control the word order difference. That is, the IP score is independent of the word order when α is 1.0. It depends strongly on the word order when α is close to 0.0. The recursive process is performed as long as two sequences share common parts. In Eqs. (1) and (2), RN denotes the number of repetition processes. Each c is a common part in the intuitive common parts continuum (CC). The parameter β is the weight based on the length of each common part (β is over 1.0) according to GTM and ROUGE-W. Moreover, $\gamma = P_{IP}/R_{IP}$ in Eq. (3).

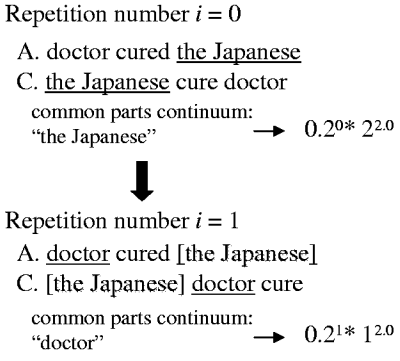

Figure 1: Recursive process between A and C.

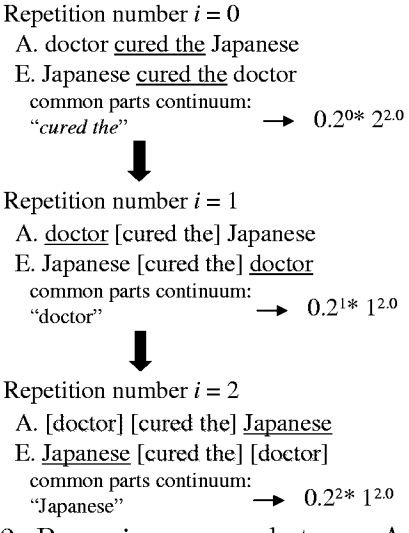

Figure 2: Recursive process between A and E.

Figure 1 shows the recursive process between A and C in section 2. When the repetition number i is 0, "*the Japanese*" is determined as the common part continuum using LCS. Next, "*doctor*" is determined recursively as the new common parts continuum after "*the Japanese*" is excluded from A and C. In addition, R_{IP} between A and C is 0.5123 ($\sqrt{(0.2^0 \times 2^{2.0} + 0.2^1 \times 1^{2.0})/4^{2.0}} = \sqrt{(4+0.2)/16} = \sqrt{0.2625}$) when α and β are, respectively, 0.2 and 2.0. The value of P_{IP} between A and C is also 0.5123. Therefore, the IP score between A and C is 0.5123. Figure 2 shows the recursive process between A and E. The common parts "*cured the*" are determined as the common parts continuum using LCS. Next, "*doctor*" is determined recursively as the new common parts continuum after "*cured the*" is excluded from A and E. In this case, the repetition number i is 1. Moreover, "*Japanese*" is determined as the common parts continuum after "*doctor*" is excluded from A and E. In this case, the repetition number i is 2. The R_{IP} and between A and E is 0.5148 ($\sqrt{(0.2^0 \times 2^{2.0} + 0.2^1 \times 1^{2.0} + 0.2^2 \times 1^{2.0})/4^{2.0}} = \sqrt{(4+0.2+0.04)/16} = \sqrt{0.265}$) when α and β are, respectively, 0.2 and 2.0. The P_{IP} between A and E is also 0.5148. Therefore, the IP score between A and E is 0.5148: IMPACT can distinguish C and E. Finally, the IP scores between A and others are as follows when α and β are, respectively, 0.2 and 2.0. This ranking is natural.

A. doctor cured the Japanese	
B. doctor cure the Japanese	0.5590
D. the Japanese doctor cured	0.5477
E. Japanese cured the doctor	0.5148
C. the Japanese cure doctor	0.5123

3.2 Multiple References

We use the maximum common parts continuum matches $\sum_{i=0}^{RN}(\alpha^i \sum_{c \in cc} length(c)^\beta)$ in the same way as ROUGE when multiple references are given. The IMPACT score between a candidate translation of n words and a set of u references of m_j words is calculated as the following.

$$R_{IP-multi} = \max_{j=1}^{u} \left(\left(\frac{\left(\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in CC} length(c)^\beta \right) \right)_j}{m_j^\beta} \right)^{\frac{1}{\beta}} \right) \quad (4)$$

$$P_{IP-multi} = \max_{j=1}^{u} \left(\left(\frac{\left(\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in CC} length(c)^\beta \right) \right)_j}{n^\beta} \right)^{\frac{1}{\beta}} \right) \quad (5)$$

$$IP_{multi}(X,Y) = \frac{(1+\gamma^2)\, R_{IP-multi} P_{IP-multi}}{R_{IP-multi} + \gamma^2 P_{IP-multi}} \quad (6)$$

In those equations, $\gamma = P_{IP-multi}/R_{IP-multi}$ in Eq. (6). In both Eqs. (3) and (6), the IP score is between 0.0 and 1.0.

3.3 Intuitive Common Parts Continuum

3.3.1 Concept

Determination of the common parts continuum between two sequences is very important for IMPACT because the recursive process can be performed by determining the common parts continuum. We previously proposed an approach to acquire linguistic knowledge by determining the common parts between two sequences from the perspective of learning (Echizen-ya et al., 1996; Echizen-ya et al., 2002; Echizen-ya et al., 2006). In this paper, we propose a method to acquire the most intuitive common parts continuum. In IMPACT, the position information of each common part is used to determine the most intuitive common parts continuum, not only the length information. Such a determination process is necessary because it is difficult to determine the common parts clearly when several common parts exist simultaneously in two sequences. Figure 3 shows an example in which several LCS routes exist in two English sentences.

In Fig. 3, the LCS is nine between the translation sentence and reference. Two routes have LCS of nine. The difference between Route No. 1 and No. 2 is the last common part "*the*" in the translation sentence. In such a case, IMPACT determines one LCS route using the length and position information of the common parts, which indicates the acquisition of the intuitive common parts continuum.

3.3.2 Acquisition of the Intuitive Common Parts Continuum

First, IMPACT determines the common parts using LCS. For this study, a common part is a word subsequence for which the word number is greater than 1. For example, in Fig. 3, the common words are "*the*", "*announcement*", "*to*", "*the*", "*exchange*", "*fire*", "*production*", "*and*" and "*the.*" On the other hand, the common parts are "*the*", "*announcement*", "*to the*", "*exchange*", "*fire*", "*production*", "*and*" and "*the.*" That is, "*to the*" is the common part because the common words "*to*" and "*the*" exist sequentially in both the translation sentence and reference.

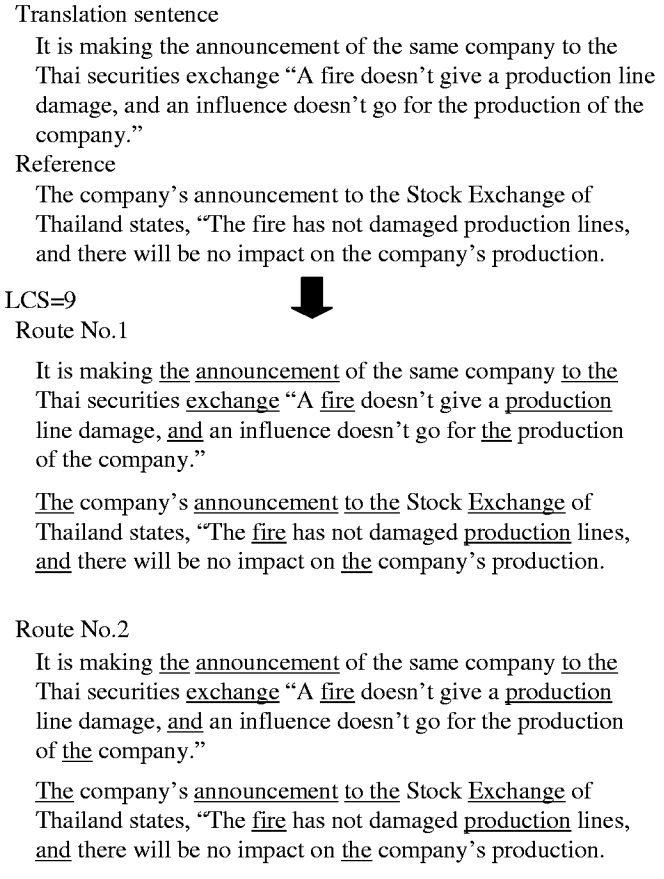

Figure 3: An example of two English sentences for which several LCS routes exist.

Next, the length and position of each common part are used to determine only one LCS route. That is, the difference of the position between two same common parts in two sequences is used as the penalty. In the following two sequences X and Y, the positions of common part "*B*" are, respectively 2 and 3 between X and Y. Therefore, the difference of position in the common part "*B*" is 1 by $|2-3|$. In the common parts for which the word number is more than 2, the position of common parts is the position of first words.

X. ABC
Y. DEB

The route score (RS) of each LCS route is calculated using the following Eqs. (7) and (8):

$$pos_w = \left(1.0 - \frac{|posX(c) - posY(c)|}{length}\right)^{\alpha} \quad (7)$$

$$RS = \left(\sum_{c \in LCS} \left(length(c)^{\beta} \times pos_w\right)\right)^{\frac{1}{\beta}} \quad (8)$$

where *length* in Eq. (7) is the length of sequence X or Y, and is determined as following.

$$length = \begin{cases} m & m \geq n \\ n & m < n \end{cases}$$

In Eq. (7), pos_w represents the weight based on the difference of the position of the common part c in sequences X and Y. Therein, posX(c) and posY(c) respectively represent the position numbers of c in sequences X and Y. Moreover, the pos_w is controlled using the parameter α ($\alpha > 0.0$). In Eq. (8), the pos_w incurs a large penalty when the difference between posX(c) and posY(c) is large. Moreover, the β is the weight based on the length of each common part ($\beta > 1.0$) following Eqs. (1) and (2). Furthermore, IMPACT selects only the one LCS route that has the maximum RS score. For example, RS of "*B*" is 0.6667 ($\sqrt{1^{2.0} \times (1.0 - \frac{|2-3|}{3})^{2.0}} = \sqrt{1 \times 0.4444}$) between X and Y above when α and β are, respectively, 2.0 and 2.0. The common parts continuum determined by this process become the intuitive common parts continuum; this process need not perform normalization using the length of sequence because the acquisition of only one LCS route is a relative process based on the same sequences.

In Fig. 3, IMPACT must determine whether "*the*" among "*the company's production*" in reference corresponds to the first "*the*" among "**the** *production of the company*" in the translation sentence or the second "*the*" among "*the production of* **the** *company*" in the translation sentence. In this case, "*the*" among "*the company's production*" corresponds to the first "*the*" among "*the production of the company*" because "*company's production*" corresponds to "*production of the company*", not "*company.*" Therefore, Route No. 1 must be selected. IMPACT can select Route No. 1 using Eqs. (7) and (8) in Fig. 3. Figure 4 presents an example

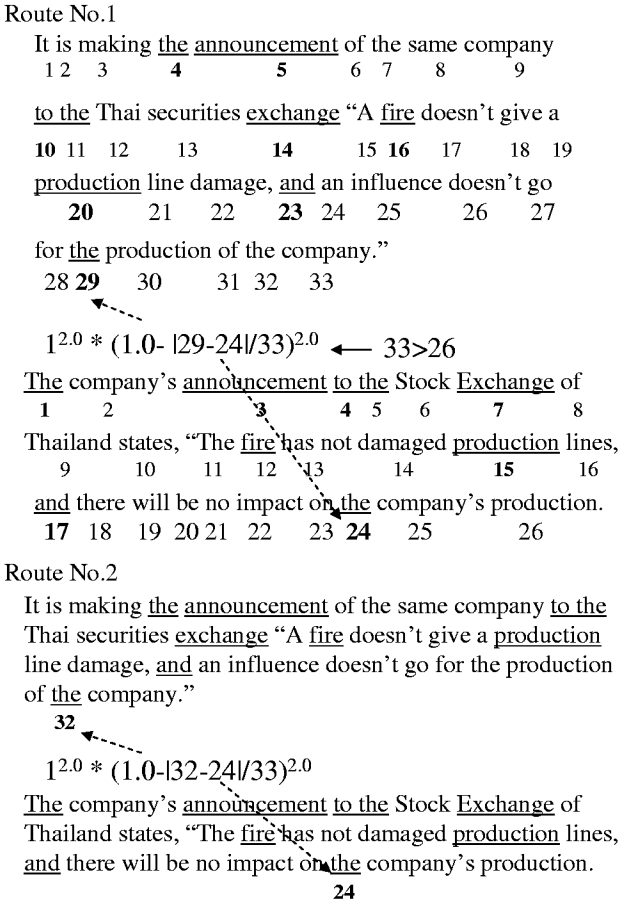

Figure 4: An example of use of position information in IMPACT.

of the use of the position information in Fig. 3. The score of the last common part "*the*" in the translation sentence of Route No. 1 is 0.7199 ($= 1^{2.0} \times (1.0 - \frac{|29-24|}{33})^{2.0}$). Moreover, the score of the last common part "*the*" in the translation sentence of Route No. 2 is 0.5739 ($= 1^{2.0} \times (1.0 - \frac{|32-24|}{33})^{2.0}$). The scores of other common parts in Route No. 1 are equal to that of Route No. 2. Therefore, IMPACT selects Route No. 1 because the score of the last common part "*the*" in Route No. 1 is larger than that of Route No. 2. Thereby, IMPACT determines the most intuitive common parts continuum based on this process.

4 Experiments

4.1 Experimental Procedure

We used 120 Japanese sentences in bilingual parallel corpora from Reuters articles(Utiyama

Table 1: Scale for a human judge.

Rank	Adequacy	Fluency
5	All Information	Flawless English
4	Most Information	Good English
3	Much Information	Non-native English
2	Little Information	Disfluent English
1	None	Incomprehensible

Table 2: Experimental results.

Adequacy				Fluency			
Method	arith-mean	pre-sys	pre-trans	Method	arith-mean	pre-sys	pre-trans
IMPACT	**1.0000**	**0.4552**	0.5065	IMPACT	0.9815	**0.5246**	**0.5914**
BLEU-4	0.9498	0.3395	0.4144	BLEU-4	0.9932	0.3408	0.4505
GTM2	0.9989	0.4265	0.4569	GTM2	0.9879	0.5181	0.5721
METEOR	0.9923	0.3962	0.4528	METEOR	0.9970	0.4071	0.4708
ROUGE-L	**1.0000**	0.4507	**0.5134**	ROUGE-L	0.9788	0.4964	0.5736
ROUGE-W	0.9999	0.4535	0.5038	ROUGE-W	0.9819	0.5215	0.5881
ROUGE-S3	0.9910	0.3436	0.4055	ROUGE-S3	**0.9977**	0.4219	0.5230

and Isahara, 2003). Three commercial machine translation systems translated these Japanese sentences into English sentences. The number of references is three. One was obtained from a parallel corpus. The others were obtained from two bilingual people. All 360 ($= 120 \times 3$) English sentences translated by the three MT systems were evaluated by a human judge. In that case, the human judge scored all translated English sentences from the perspective of adequacy and fluency on a scale of 1 to 5. The details are presented in Table 1.

We used six evaluation methods for comparison with IMPACT. The six evaluation methods are BLEU, GTM, METEOR, ROUGE-L, ROUGE-W and ROUGE-S. Moreover, we calculated the Pearson's R correlation value, which is a correlation analysis based on two different correlation statistics. In that case, we used the following three evaluation values.

arith-mean: Each vector element of human judgment and evaluation measure corresponds to the arithmetic mean of score for each MT system. Therefore, in these experiments, the number of vector elements is three because three MT systems are used.

pre-sys: Each vector element of human judgment and evaluation measure corresponds to the score for each translated sentence for each MT system. Therefore, in these experiments, the number of vector elements is 120 because each MT system translated 120 Japanese sentences into 120 English sentences. Moreover, we calculated the average of three of the correlation values as the final correlation values.

pre-trans: Each vector element of human judgment and evaluation measure corresponds to the score for all translated sentences for all MT systems. Therefore, in these experiments, the number of vector elements is 360 ($= 120 \times 3$) because 360 English sentences were obtained using three MT systems.

4.2 Experimental Results

Table 2 shows experimental results obtained using IMPACT and other methods. In these IMPACT experiments, 0.4 and 1.2 were used as the α value and the β value, respectively, in Eqs. (4) and (5). Moreover, 1.5 and 1.2 were used as the α value and the β value, respectively, in Eqs. (7) and (8). These values indicated the highest Pearson's R correlation values. In BLEU, BLEU-4 indicated the highest correlation value among BLEU-1 to BLEU-7. In

Table 3: Experimental results using Japanese sentences.

Adequacy				Fluency			
Method	arith-mean	pre-sys	pre-trans	Method	arith-mean	pre-sys	pre-trans
IMPACT	0.9971	**0.4667**	0.4673	IMPACT	0.9855	0.5988	0.5860
GTM2	0.9845	0.3348	0.3195	GTM2	0.9637	0.5418	0.5067
ROUGE-L	**0.9984**	0.4664	**0.4810**	ROUGE-L	**0.9886**	0.5695	0.5724
ROUGE-W	0.9961	0.4659	0.4668	ROUGE-W	0.9833	**0.5989**	**0.5863**

GTM, GTM2 indicated the highest correlation value among GTM1($e = 1$: e means exponent), GTM2($e = 2$) and GTM3($e = 3$). Moreover, in METEOR, the exact METEOR based on matching using the surface form was used. In ROUGE-W, the weight for the length of consecutive matches is 1.2. In ROUGE-S, ROUGE-S3 indicated the highest correlation value among ROUGE-S1 to ROUGE-S4. In Table 2, IMPACT shows the highest correlation with human judgment except pre-trans of adequacy and arith-mean of fluency.

4.3 Discussion

These experimental results confirmed the effectiveness of IMPACT. In Table 2, the correlation values of IMPACT are the closest to those of ROUGE-W. The reason is that English sentences do not require a recursive process in IMPACT. English is a language for which the word order limitation is comparatively strict. In these experiments, many English sentences that were used did not allow changing of the word order between the translated sentences and the references. Results showed that IMPACT did not improve the result drastically compared to ROUGE-W. However, from the perspective of the word order, various sentences might appear in the languages for which the limitation of the word order is not strict, compared to that of English. In such cases, ROUGE-L and ROUGE-W are insufficient because they can not process all common parts between two sequences. Although GTM can deal with all common parts between two sequences, it does not use word-order information. Therefore, it is insufficient for the languages for which the word order limitation is strict. The correlation value of GTM2 in adequacy is lower than that of IMPACT in Table 2.

Moreover, we applied IMPACT to Japanese sentences. In the experiments, we used Japanese sentences in which three MT systems translated 60 English sentences. All English sentences are included in the bilingual parallel corpus from the article of Reuters(Utiyama and Isahara, 2003). The number of references is three. One was obtained from the parallel corpus. The others were obtained from two bilingual people. All 180 ($= 60 \times 3$) Japanese sentences translated by three MT systems were evaluated by a human judge. In addition, they were inserted after each morpheme using the Japanese morphological analysis system "ChaSen"(Matsumoto et al., 2000). Table 3 shows the experimental results using Japanese sentences in IMPACT, GTM2, ROUGE-L and ROUGE-W. In IMPACT, 0.01 and 1.1 were used as the α value and the β value, respectively, in Eqs. (4) and (5). Moreover, 1.5 and 1.1 were used as the α value and the β value, respectively, in Eqs. (7) and (8). In ROUGE-W, the weight for the length of consecutive matches is 1.1.

In Table 3, IMPACT results are ranked top or second in all evaluation values. In IMPACT, the low α value (α=0.01) in Eqs. (4) and (5) was effective. In Japanese sentences that were used for these evaluation experiments, the limitation of the word order was strict although the limitation of the word order in Japanese is not stricter, than that of English. These experimental results in Tables 2 and 3 indicate that IMPACT can constantly obtain the high correlation with human judgment among the some methods for automatic evaluation.

In addition, we investigated the effectiveness of the position information. Table 4 shows a comparison with IMPACT using the position information and IMPACT without the position information in the experiments using English sentences. IMPACT without the position in-

Table 4: Comparison with IMPACT using the position information (PI) and without the position information.

Adequacy				Fluency			
Method	arith-mean	pre-sys	pre-trans	Method	arith-mean	pre-sys	pre-trans
PI	1.0000	0.4552	0.5065	PI	0.9815	0.5246	0.5914
Non-PI	0.9998	0.4446	0.4993	Non-PI	0.9760	0.5154	0.5827

formation means the case that 1.0 was used as the α value of Eqs. (4) and (5). That is, the penalty is not assessed. Moreover, Eq. (8) is replaced as the following Eq. (9).

$$RS = \left(\sum_{c \in LCS} \left(length(c)^{\beta} \right) \right)^{\frac{1}{\beta}} \qquad (9)$$

In Table 4, the results of IMPACT using the position information are the same as the results of IMPACT in Table 2. We confirmed the effectiveness of the use of the position information in IMPACT. All correlation values of IMPACT without the position information in Table 4 are lower than that of ROUGE-W in Table 2. Therefore, it is very important for IMPACT to use position information.

5 Conclusion

In this paper, we proposed a new automatic evaluation method for machine translation evaluation. We call this method Recursive Acquisition of **I**ntuitive Com**M**on **PA**rts **C**on**T**inuum (IMPACT). It uses both the length information and the position information of the common parts between two sequences, not only the length information. Experimental results indicated that IMPACT is effective to obtain the correlation with human judgment. Moreover, IMPACT can deal with various languages by changing the parameters.

In the future, we will perform the experiments by various languages, and improve IMPACT to obtain higher correlation values. Moreover, we plan to release the IMPACT software at an early date by "http://www.eli.hokkai-s-u.ac.jp/~echi/impact.html".

References

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

Echizen-ya, Hiroshi, Kenji Araki, Yoshio Momouchi, and Koji Tochinai. 1996. Machine Translation Method Using Inductive Learning with Genetic Algorithms. pp.1020–1023. *In Proc. of the Coling'96.*

Echizen-ya, Hiroshi, Kenji Araki, Yoshio Momouchi, and Koji Tochinai. 2002. Study of Practical Effectiveness for Machine Translation using Recursive Chain-link-type Learning. pp.246–252. *In Proc. of the Coling'02.*

Echizen-ya, Hiroshi, Kenji Araki and Yoshio Momouchi. 2006. Automatic Extraction of Bilingual Word Pairs Using Inductive Chain Learning in Various Languages. pp.1294-1315. *Information Processing and Management, Elsevier*, vol.42, no.5.

Lin, Chin-Yew, and Franz Josef Och. 2004. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. pp.606–613. *In Proc. of the ACL'04.*

Matsumoto, Y., A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka and M. Asahara. 2000. Japanese Morphological Analysis System ChaSen version 2.2.1 manual. *Nara Institute of Science and Technology.*

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. pp.311–318. *In Proc. of the ACL'02.*

Turian, Joseph P., Luke Shen and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. pp.386–393. *In Proc. of the MT Summit IX.*

Utiyama, Masao, and Hitoshi Isahara. 2003. Reliable Measures for Aligning Japanese-English News Articles and Sentences. pp.72–79. *In Proc. of the ACL'03.*