Automatic Evaluation Method for Machine Translation using Noun-Phrase Chunking

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

Kenji Araki
Hokkaido University
N 14-Jo, W 9-Chome, Kita-ku,
Sapporo, 060-0814 Japan
araki@media.eng.hokudai.ac.jp

Abstract

As described in this paper, we propose a new automatic evaluation method for machine translation using noun-phrase chunking. Our method correctly determines the matching words between two sentences using corresponding noun phrases. Moreover, our method determines the similarity between two sentences in terms of the noun-phrase order of appearance. Evaluation experiments were conducted to calculate the correlation among human judgments, along with the scores produced using automatic evaluation methods for MT outputs obtained from the 12 machine translation systems in NTCIR-7. Experimental results show that our method obtained the highest correlations among the methods in both sentence-level adequacy and fluency.

1 Introduction

High-quality automatic evaluation has become increasingly important as various machine translation systems have developed. The scores of some automatic evaluation methods can obtain high correlation with human judgment in document-level automatic evaluation(Coughlin, 2007). However, sentence-level automatic evaluation is insufficient. A great gap exists between language processing of automatic evaluation and the processing by humans. Therefore, in recent years, various automatic evaluation methods particularly addressing sentence-level automatic evaluations have been proposed. Methods based on word strings (*e.g.*, BLEU(Papineni et al., 2002), NIST(NIST, 2002), METEOR(Banerjee and Lavie., 2005), ROUGE-L(Lin and Och, 2004), and IMPACT(Echizen-ya and Araki, 2007)) calculate matching scores using only common words between MT outputs and references from bilingual humans. However, these methods cannot determine the correct word correspondences sufficiently because they fail to focus solely on phrase correspondences. Moreover, various methods using syntactic analytical tools(Pozar and Charniak, 2006; Mutton et al., 2007; Mehay and Brew, 2007) are proposed to address the sentence structure. Nevertheless, those methods depend strongly on the quality of the syntactic analytical tools.

As described herein, for use with MT systems, we propose a new automatic evaluation method using noun-phrase chunking to obtain higher sentence-level correlations. Using noun phrases produced by chunking, our method yields the correct word correspondences and determines the similarity between two sentences in terms of the noun phrase order of appearance. Evaluation experiments using MT outputs obtained by 12 machine translation systems in NTCIR-7(Fujii et al., 2008) demonstrate that the scores obtained using our system yield the highest correlation with the human judgments among the automatic evaluation methods in both sentence-level adequacy and fluency. Moreover, the differences between correlation coefficients obtained using our method and other methods are statistically significant at the 5% or lower significance level for adequacy. Results confirmed that our method using noun-phrase chunking is effective for automatic evaluation for machine translation.

2 Automatic Evaluation Method using Noun-Phrase Chunking

The system based on our method has four processes. First, the system determines the corre-

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 108–117,
Uppsala, Sweden, 11-16 July 2010.

spondences of noun phrases between MT outputs and references using chunking. Secondly, the system calculates word-level scores based on the correct matched words using the determined correspondences of noun phrases. Next, the system calculates phrase-level scores based on the noun-phrase order of appearance. The system calculates the final scores combining word-level scores and phrase-level scores.

2.1 Correspondence of Noun Phrases by Chunking

The system obtains the noun phrases from each sentence by chunking. It then determines corresponding noun phrases between MT outputs and references calculating the similarity for two noun phrases by the PER score(Su et al., 1992). In that case, PER scores of two kinds are calculated. One is the ratio of the number of match words between an MT output and reference for the number of all words of the MT output. The other is the ratio of the number of match words between the MT output and reference for the number of all words of the reference. The similarity is obtained as an F-measure between two PER scores. The high score represents that the similarity between two noun phrases is high. Figure 1 presents an example of the determination of the corresponding noun phrases.

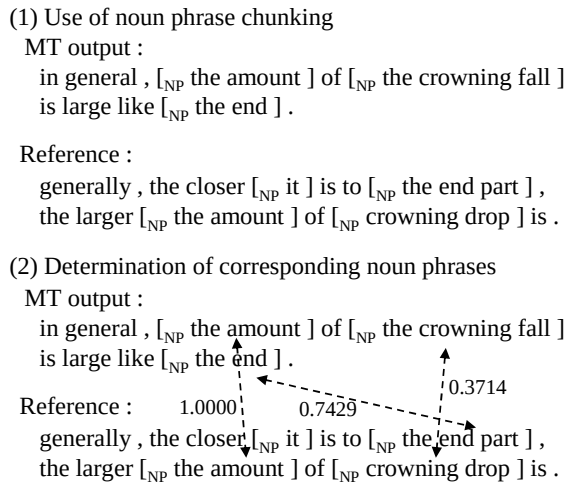

Figure 1: Example of determination of corresponding noun phrases.

In Fig. 1, "the amount", "the crowning fall" and "the end" are obtained as noun phrases in MT output by chunking, and "it", "the end part", "the amount" and "crowning drop" are obtained in the reference by chunking. Next, the system determines the corresponding noun phrases from these noun phrases between the MT output and reference. The score between "the end" and "the end part" is the highest among the scores between "the end" in the MT output and "it", "the end part", "the amount", and "crowning drop" in the reference. Moreover, the score between "the end part" and "the end" is the highest among the scores between "the end part" in reference and "the amount", "the crowning fall", "the end" in the MT output. Consequently, "the end" and "the end part" are selected as noun phrases with the highest mutual scores: "the end" and "the end part" are determined as one corresponding noun phrase. In Fig. 1, "the amount" in the MT output and "the amount" in reference, and "the crowning fall" in the MT output and "crowning drop" in the reference also are determined as the respective corresponding noun phrases. The noun phrase for which the score between it and other noun phrases is 0.0 (*e.g.*, "it" in reference) has no corresponding noun phrase. The use of the noun phrases is effective because the frequency of the noun phrases is higher than those of other phrases. The verb phrases are not used for this study, but they can also be generated by chunking. It is difficult to determine the corresponding verb phrases correctly because the words in each verb phrase are often fewer than the noun phrases.

2.2 Word-level Score

The system calculates the word-level scores between MT output and reference using the corresponding noun phrases. First, the system determines the common words based on Longest Common Subsequence (LCS). The system selects only one LCS route when several LCS routes exist. In such cases, the system calculates the Route Score (RS) using the following Eqs. (1) and (2):

$$RS = \sum_{c \in LCS} \left(\sum_{w \in c} weight(w) \right)^{\beta} \qquad (1)$$

$$weight(w) = \begin{cases} 2 & \text{words in corresponding noun phrase} \\ 1 & \text{words in non-corresponding noun phrase} \end{cases} \quad (2)$$

In Eq. (1), β is a parameter for length weighting of common parts; it is greater than 1.0. Figure 2 portrays an example of determination of the common parts. In the first process of Fig. 2, LCS is 7. In this example, several LCS routes exist. The system selects the LCS route which has ",", "the amount of", "crowning", "is", and "." as the common parts. The common part is the part for which the common words appear continuously. In contrast, IMPACT selects a different LCS route that includes ", the", "amount of", "crowning", "is", and "." as the common parts. In IMPACT, using no analytical knowledge, the LCS route is determined using the information of the number of words in the common parts and the position of the common parts. The RS for LCS route selected using our method is 32 $(= 1^{2.0} + (2 + 2 + 1)^{2.0} + 2^{2.0} + 1^{2.0} + 1^{2.0})$ when β is 2.0. The RS for LCS route selected by IMPACT is 19 $(= (1 + 1)^{2.0} + (2 + 1)^{2.0} + 2^{2.0} + 1^{2.0} + 1^{2.0})$. In the LCS route selected by IMPACT, the weight of "the" in the common part ", the" is 1 because "the" in the reference is not included in the corresponding noun phrase. In the LCS route selected using our method, the weight of "the" in "the amount of" is 2 because "the" in MT output and "the" in the reference are included in the corresponding noun phrase "NP1". Therefore, the system based on our method can select the correct LCS route.

Moreover, the word-level score is calculated using the common parts in the selected LCS route as the following Eqs. (3), (4), and (5).

$$R_{wd} = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in LCS} length(c)^\beta \right)}{m^\beta} \right)^{\frac{1}{\beta}} \quad (3)$$

$$P_{wd} = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in LCS} length(c)^\beta \right)}{n^\beta} \right)^{\frac{1}{\beta}} \quad (4)$$

(1) First process for determination of common parts :
LCS = 7
Our method
MT output :
in general **,** [$_{NP1}$ **the amount] of** [$_{NP2}$ the **crowning** fall] **is** large like [$_{NP3}$ the end] **.**
Reference : $1^{2.0}$ $(2+2+1)^{2.0}$ $1^{2.0}$ $1^{2.0}$ $2^{2.0}$
generally **,** the closer [$_{NP}$ it] is to [$_{NP3}$ the end part] , the larger [$_{NP1}$ **the amount] of** [$_{NP2}$ **crowning** drop] **is .**

IMPACT
MT output :
in general **,** [$_{NP1}$ **the amount] of** [$_{NP2}$ the **crowning** fall] **is** large like [$_{NP3}$ the end] **.**
Reference : $(1+1)^{2.0}$ $(2+1)^{2.0}$ $2^{2.0}$ $1^{2.0}$ $1^{2.0}$
generally **, the** closer [$_{NP}$ it] is to [$_{NP3}$ the end part] , the larger [$_{NP1}$ the **amount] of** [$_{NP2}$ **crowning** drop] **is .**

(2) Second process for determination of common parts :
LCS=3
Our method
MT output :
in general , [$_{NP1}$ *the amount*] *of* [$_{NP2}$ **the** *crowning* fall] *is* large like [$_{NP3}$ **the end**] .
Reference :
generally , **the** closer [$_{NP}$ it] is to [$_{NP3}$ **the end** part] , the larger [$_{NP1}$ *the amount*] *of* [$_{NP2}$ *crowning* drop] *is* .

Figure 2: Example of common-part determination.

$$score_{wd} = \frac{(1 + \gamma^2) R_{wd} P_{wd}}{R_{wd} + \gamma^2 P_{wd}} \quad (5)$$

Equation (3) represents recall and Eq. (4) represents precision. Therein, m signifies the word number of the reference in Eq. (3), and n stands for the word number of the MT output in Eq. (4). Here, RN denotes the repetition number of the determination process of the LCS route, and i, which has initial value 0, is the counter for RN. In Eqs. (3) and (4), α is a parameter for the repetition process of the determination of LCS route, and is less than 1.0. Therefore, R_{wd} and P_{wd} becomes small as the appearance order of the common parts between MT output and reference is different. Moreover, $length(c)$ represents the number of words in each common part; β is a parameter related to the length weight of common parts, as in Eq. (1). In this case, the weight of each common word in the common part is 1. The system calculates $score_{wd}$ as the word-level score in Eq. (5). In Eq. (5), γ is determined as P_{wd}/R_{wd}. The $score_{wd}$ is between 0.0 and 1.0.

In the first process of Fig. 2, $\alpha^i \sum_{c \in LCS} length(c)^\beta$ is 13.0 ($=0.5^0 \times (1^{2.0} + 3^{2.0} + 1^{2.0} + 1^{2.0} + 1^{2.0})$) when α and β are 0.5 and 2.0, respectively. In this case, the counter i is 0. Moreover, in the second process of Fig. 2, $\alpha^i \sum_{c \in LCS} length(c)^\beta$ is 2.5 ($=0.5^1 \times (1^{2.0} + 2^{2.0})$) using two common parts "the" and "the end", except the common parts determined using the first process. In Fig. 2, RN is 1 because the system finishes calculating $\alpha^i \sum_{c \in LCS} length(c)^\beta$ when counter i became 1: this means that all common parts were processed until the second process. As a result, R_{wd} is 0.1969 ($=\sqrt{(13.0 + 2.5)/20^{2.0}} = \sqrt{0.0388}$), and P_{wd} is 0.2625 ($=\sqrt{(13.0 + 2.5)/15^{2.0}} = \sqrt{0.0689}$). Consequently, $score_{wd}$ is 0.2164 ($=\frac{(1+1.3332^2) \times 0.1969 \times 0.2625}{0.1969 + 1.3332^2 \times 0.2625}$). In this case, γ becomes 1.3332 ($=\frac{0.2625}{0.1969}$). The system can determine the matching words correctly using the corresponding noun phrases between the MT output and the reference.

The system calculates $score_{wd\text{-}multi}$ using $R_{wd\text{-}multi}$ and $P_{wd\text{-}multi}$ which are, respectively, maximum R_{wd} and P_{wd} when multiple references are used as the following Eqs. (6), (7) and (8). In Eq. (8), γ is determined as $P_{wd\text{-}multi}/R_{wd\text{-}multi}$. The $score_{wd\text{-}multi}$ is between 0.0 and 1.0.

$$R_{wd\text{-}multi} = \max_{j=1}^{u} \left(\left(\frac{\left(\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in LCS} length(c)^\beta \right) \right)_j}{m_j^\beta} \right)^{\frac{1}{\beta}} \right) \quad (6)$$

$$P_{wd\text{-}multi} = \max_{j=1}^{u} \left(\left(\frac{\left(\sum_{i=0}^{RN} \left(\alpha^i \sum_{c \in LCS} length(c)^\beta \right) \right)_j}{n_j^\beta} \right)^{\frac{1}{\beta}} \right) \quad (7)$$

$$score_{wd\text{-}multi} = \frac{(1 + \gamma^2 R_{wd\text{-}multi}) P_{wd\text{-}multi}}{R_{wd\text{-}multi} + \gamma^2 P_{wd\text{-}multi}} \quad (8)$$

2.3 Phrase-level Score

The system calculates the phrase-level score using the noun phrases obtained by chunking. First, the system extracts only noun phrases from sentences. Then it generalizes each noun phrase as each word. Figure 3 presents examples of generalization by noun phrases.

(1) Corresponding noun phrases
MT output :
in general , [$_{NP1}$ the amount] of [$_{NP2}$ the crowning fall] is large like [$_{NP3}$ the end] .
Reference :
generally , the closer [$_{NP}$ it] is to [$_{NP3}$ the end part] , the larger [$_{NP1}$ the amount] of [$_{NP2}$ crowning drop] is .
(2) Generalization by noun phrases
MT output :
NP1 NP2 NP3
Reference :
NP NP3 NP1 NP2

Figure 3: Example of generalization by noun phrases.

Figure 3 presents three corresponding noun phrases between the MT output and the reference. The noun phrase "it", which has no corresponding noun phrase, is expressed as "NP" in the reference. Consequently, the MT output is generalized as "NP1 NP2 NP3"; the reference is generalized as "NP NP3 NP1 NP2". Subsequently, the system obtains the phrase-level score between the generalized MT output and reference as the following Eqs. (9), (10), and (11).

$$R_{np} = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{cnpp \in LCS} length(cnpp)^\beta \right)}{\left(m_{cnp} \times \sqrt{m_{no\text{-}cnp}} \right)^\beta} \right)^{\frac{1}{\beta}} \quad (9)$$

$$P_{np} = \left(\frac{\sum_{i=0}^{RN} \left(\alpha^i \sum_{cnpp \in LCS} length(cnpp)^\beta \right)}{\left(n_{cnp} \times \sqrt{n_{no\text{-}cnp}} \right)^\beta} \right)^{\frac{1}{\beta}} \quad (10)$$

Table 1: Machine translation system types.

	System No. 1	System No. 2	System No. 3	System No. 4	System No. 5	System No. 6
Type	SMT	SMT	RBMT	SMT	SMT	SMT
	System No. 7	System No. 8	System No. 9	System No. 10	System No. 11	System No. 12
Type	SMT	SMT	EBMT	SMT	SMT	RBMT

$$score_{np} = \frac{(1+\gamma^2)R_{np}P_{np}}{R_{np}+\gamma^2 P_{np}} \quad (11)$$

In Eqs. (9) and (10), *cnpp* denotes the common noun phrase parts; m_{cnp} and n_{cnp} respectively signify the quantities of common noun phrases in the reference and MT output. Moreover, $m_{no\text{-}cnp}$ and $n_{no\text{-}cnp}$ are the quantities of noun phrases except the common noun phrases in the reference and MT output. The values of $m_{no\text{-}cnp}$ and $n_{no\text{-}cnp}$ are processed as 1 when no non-corresponding noun phrases exist. The square root used for $m_{no\text{-}cnp}$ and $n_{no\text{-}cnp}$ is to decrease the weight of the non-corresponding noun phrases. In Eq. (11), γ is determined as P_{np}/R_{np}. In Fig. 3, R_{np} and P_{np} are 0.7071 ($=\sqrt{\frac{1\times 2^{2.0}+0.5\times 1^{2.0}}{(3\times 1)^{2.0}}}$) when α is 0.5 and β is 2.0. Therefore, $score_{np}$ is 0.7071.

The system obtains $score_{np\text{-}multi}$ calculating the average of $score_{np}$ when multiple references are used as the following Eq. (12).

$$score_{np\text{-}multi} = \frac{\sum_{j=0}^{u}\left(score_{np}\right)_j}{u} \quad (12)$$

2.4 Final Score

The system calculates the final score by combining the word-level score and the phrase-level score as shown in the following Eq. (13).

$$score = \frac{score_{wd} + \delta \times score_{np}}{1+\delta} \quad (13)$$

Therein, δ represents a parameter for the weight of $score_{np}$: it is between 0.0 and 1.0. The ratio of $score_{wd}$ to $score_{np}$ is 1:1 when δ is 1.0. Moreover, $score_{wd\text{-}multi}$ and $score_{np\text{-}multi}$ are used for Eq. (13) in multiple references. In Figs. 2 and 3, the final score between the MT output and the reference is 0.4185 ($=\frac{0.2164+0.7\times 0.7071}{1+0.7}$) when δ is 0.7. The system can realize high-quality automatic evaluation using both word-level information and phrase-level information.

3 Experiments

3.1 Experimental Procedure

We calculated the correlation between the scores obtained using our method and scores produced by human judgment. The system based on our method obtained the evaluation scores for 1,200 English output sentences related to the patent sentences. These English output sentences are sentences that 12 machine translation systems in NTCIR-7 translated from 100 Japanese sentences. Moreover, the number of references to each English sentence in 100 English sentences is four. These references were obtained from four bilingual humans. Table 1 presents types of the 12 machine translation systems.

Moreover, three human judges evaluated 1,200 English output sentences from the perspective of adequacy and fluency on a scale of 1–5. We used the median value in the evaluation results of three human judges as the final scores of 1–5. We calculated Pearson's correlation efficient and Spearman's rank correlation efficient between the scores obtained using our method and the scores by human judgments in terms of sentence-level adequacy and fluency.

Additionally, we calculated the correlations between the scores using seven other methods and the scores by human judgments to compare our method with other automatic evaluation methods. The other seven methods were IMPACT, ROUGE-L, BLEU[1], NIST, NMG-WN(Ehara, 2007; Echizen-ya et al., 2009), METEOR[2], and WER(Leusch et al., 2003). Using our method, 0.1 was used as the value of the parameter α in Eqs. (3)-(10) and 1.1 was used as the value of the parameter β in Eqs. (1)–(10). Moreover, 0.3 was used as the value of the parameter δ in Eq. (13). These val-

[1]BLEU was improved to perform sentence-level evaluation: the maximum N value between MT output and reference is used(Echizen-ya et al., 2009).

[2]The matching modules of METEOR are the exact and stemmed matching module, and a WordNet-based synonym-matching module.

Table 2: Pearson's correlation coefficient for sentence-level adequacy.

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7
Our method	**0.7862**	**0.4989**	0.5970	**0.5713**	**0.6581**	**0.6779**	**0.7682**
IMPACT	0.7639	0.4487	0.5980	0.5371	0.6371	0.6255	0.7249
ROUGE-L	0.7597	0.4264	**0.6111**	0.5229	0.6183	0.5927	0.7079
BLEU	0.6473	0.2463	0.4230	0.4336	0.3727	0.4124	0.5340
NIST	0.5135	0.2756	0.4142	0.3086	0.2553	0.2300	0.3628
NMG-WN	0.7010	0.3432	0.6067	0.4719	0.5441	0.5885	0.5906
METEOR	0.4509	0.0892	0.3907	0.2781	0.3120	0.2744	0.3937
WER	0.7464	0.4114	0.5519	0.5185	0.5461	0.5970	0.6902
Our method II	0.7870	0.5066	0.5967	0.5191	0.6529	0.6635	0.7698
BLEU with our method	0.7244	0.3935	0.5148	0.5231	0.4882	0.5554	0.6459
	No. 8	No. 9	No. 10	No. 11	No. 12	Avg.	All
Our method	**0.7664**	**0.7208**	**0.6355**	**0.7781**	0.5707	**0.6691**	**0.6846**
IMPACT	0.7007	0.7125	0.5981	0.7621	0.5345	0.6369	0.6574
ROUGE-L	0.6834	0.7042	0.5691	0.7480	0.5293	0.6228	0.6529
BLEU	0.5188	0.5884	0.3697	0.5459	0.4357	0.4607	0.4722
NIST	0.4218	0.4092	0.1721	0.3521	0.4769	0.3493	0.3326
NMG-WN	0.6658	0.6068	0.6116	0.6770	**0.5740**	0.5818	0.5669
METEOR	0.3881	0.4947	0.3127	0.2987	0.4162	0.3416	0.2958
WER	0.6656	0.6570	0.5740	0.7491	0.5301	0.6031	0.5205
Our method II	0.7676	0.7217	0.6343	0.7917	0.5474	0.6632	0.6774
BLEU with our method	0.6395	0.6696	0.5139	0.6611	0.5079	0.5698	0.5790

ues of the parameter are determined using English sentences from Reuters articles(Utiyama and Isahara, 2003). Moreover, we obtained the noun phrases using a shallow parser(Sha and Pereira, 2003) as the chunking tool. We revised some erroneous results that were obtained using the chunking tool.

3.2 Experimental Results

As described in this paper, we performed comparison experiments using our method and seven other methods. Tables 2 and 3 respectively show Pearson's correlation coefficient for sentence-level adequacy and fluency. Tables 4 and 5 respectively show Spearman's rank correlation coefficient for sentence-level adequacy and fluency. In Tables 2–5, bold typeface signifies the maximum correlation coefficients among eight automatic evaluation methods. Underlining in our method signifies that the differences between correlation coefficients obtained using our method and IMPACT are statistically significant at the 5% significance level. Moreover, "Avg." signifies the average of the correlation coefficients obtained by 12 machine translation systems in respective automatic evaluation methods, and "All" are the correlation coefficients using the scores of 1,200 output sentences obtained using the 12 machine translation systems.

3.3 Discussion

In Tables 2–5, the "Avg." score of our method is shown to be higher than those of other methods. Especially in terms of the sentence-level adequacy shown in Tables 2 and 4, "Avg." of our method is about 0.03 higher than that of IMPACT. Moreover, in system No. 8 and "All" of Tables 2 and 4, the differences between correlation coefficients obtained using our method and IMPACT are statistically significant at the 5% significance level.

Moreover, we investigated the correlation of machine translation systems of every type. Table 6 shows "All" of Pearson's correlation coefficient and Spearman's rank correlation coefficient in SMT (*i.e.*, system Nos. 1–2, system Nos. 4–8 and system Nos. 10–11) and RBMT (*i.e.*, system Nos. 3 and 12). The scores of 900 output sentences obtained by 9 machine

Table 3: Pearson's correlation coefficient for sentence-level fluency.

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7
Our method	**0.5853**	**0.3782**	0.5689	0.4673	0.5739	**0.5344**	**0.7193**
IMPACT	0.5581	0.3407	0.5821	0.4586	**0.5768**	0.4852	0.6896
ROUGE-L	0.5551	0.3056	**0.5925**	0.4391	0.5666	0.4475	0.6756
BLEU	0.4793	0.0963	0.4488	0.3033	0.4690	0.3602	0.5272
NIST	0.4139	0.0257	0.4987	0.1682	0.3923	0.2236	0.3749
NMG-WN	0.5782	0.3090	0.5434	**0.4680**	0.5070	0.5234	0.5363
METEOR	0.4050	0.1405	0.4420	0.1825	0.4259	0.2336	0.4873
WER	0.5143	0.3031	0.5220	0.4262	0.4936	0.4405	0.6351
Our method II	0.5831	0.3689	0.5753	0.3991	0.5610	0.5445	0.7186
BLEU with our method	0.5425	0.2304	0.5115	0.3770	0.5358	0.4741	0.6142
	No. 8	No. 9	No. 10	No. 11	No. 12	Avg.	All
Our method	**0.5796**	**0.6424**	0.3241	0.5920	0.4321	**0.5331**	**0.5574**
IMPACT	0.5612	0.6320	0.3492	0.6034	0.4166	0.5211	0.5469
ROUGE-L	0.5414	0.6347	0.3231	0.5889	0.4127	0.5069	0.5387
BLEU	0.5040	0.5521	0.2134	0.4783	0.4078	0.4033	0.4278
NIST	0.3682	0.3811	0.1682	0.3116	**0.4484**	0.3146	0.3142
NMG-WN	0.5526	0.5799	**0.4509**	**0.6308**	0.4124	0.5007	0.5074
METEOR	0.2511	0.4153	0.1376	0.3351	0.2902	0.3122	0.2933
WER	0.5492	0.6421	0.3962	0.6228	0.4063	0.4960	0.4478
Our method II	0.5774	0.6486	0.3428	0.5975	0.4197	0.5280	0.5519
BLEU with our method	0.5660	0.6247	0.2536	0.5495	0.4550	0.4770	0.5014

translation systems in SMT and the scores of 200 output sentences obtained by 2 machine translation systems in RBMT are used respectively. However, EBMT is not included in Table 6 because EBMT is only system No. 9. In Table 6, our method obtained the highest correlation among the eight methods, except in terms of the adequacy of RBMT in Pearson's correlation coefficient. The differences between correlation coefficients obtained using our method and IMPACT are statistically significant at the 5% significance level for adequacy of SMT.

To confirm the effectiveness of noun-phrase chunking, we performed the experiment using a system combining BLEU with our method. In this case, BLEU scores were used as $score_{wd}$ in Eq. (13). This experimental result is shown as "BLEU with our method" in Tables 2–5. In the results of "BLEU with our method" in Tables 2–5, underlining signifies that the differences between correlation coefficients obtained using BLEU with our method and BLEU alone are statistically significant at the 5% significance level. The coefficients of correlation for BLEU with our method are higher than those of BLEU in any machine translation system, "Avg." and "All" in Tables 2–5. Moreover, for sentence-level adequacy, BLEU with our method is significantly better than BLEU in almost all machine translation systems and "All" in Tables 2 and 4. These results indicate that our method using noun-phrase chunking is effective for some methods and that it is statistically significant in each machine translation system, not only "All", which has large sentences.

Subsequently, we investigated the precision of the determination process of the corresponding noun phrases described in section 2.1: in the results of system No. 1, we calculated the precision as the ratio of the number of the correct corresponding noun phrases for the number of all noun-phrase correspondences obtained using the system based on our method. Results show that the precision was 93.4%, demonstrating that our method can determine the corresponding noun phrases correctly.

Moreover, we investigated the relation be-

Table 4: Spearman's rank correlation coefficient for sentence-level adequacy.

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7
Our method	0.7456	**0.5049**	0.5837	**0.5146**	**0.6514**	**0.6557**	**0.6746**
IMPACT	0.7336	0.4881	0.5992	0.4741	0.6382	0.5841	0.6409
ROUGE-L	0.7304	0.4822	**0.6092**	0.4572	0.6135	0.5365	0.6368
BLEU	0.5525	0.2206	0.4327	0.3449	0.3230	0.2805	0.4375
NIST	0.5032	0.2438	0.4218	0.2489	0.2342	0.1534	0.3529
NMG-WN	**0.7541**	0.3829	0.5579	0.4472	0.5560	0.5828	0.6263
METEOR	0.4409	0.1509	0.4018	0.2580	0.3085	0.1991	0.4115
WER	0.6566	0.4147	0.5478	0.4272	0.5524	0.4884	0.5539
Our method II	0.7478	0.4972	0.5817	0.4892	0.6437	0.6428	0.6707
BLEU with our method	0.6644	0.3926	0.5065	0.4522	0.4639	0.4715	0.5460
	No. 8	No. 9	No. 10	No. 11	No. 12	Avg.	All
Our method	**0.7298**	**0.7258**	0.5961	**0.7633**	**0.6078**	**0.6461**	**0.6763**
IMPACT	0.6703	0.7067	0.5617	0.7411	0.5583	0.6164	0.6515
ROUGE-L	0.6603	0.6983	0.5340	0.7280	0.5281	0.6012	0.6435
BLEU	0.4571	0.5827	0.3220	0.4987	0.4302	0.4069	0.4227
NIST	0.4255	0.4424	0.1313	0.2950	0.4785	0.3276	0.3062
NMG-WN	0.6863	0.6524	**0.6412**	0.7015	0.5728	0.5968	0.5836
METEOR	0.4242	0.4776	0.3335	0.2861	0.4455	0.3448	0.2887
WER	0.6234	0.6480	0.5463	0.7131	0.5684	0.5617	0.4797
Our method II	0.7287	0.7255	0.5936	0.7761	0.5798	0.6397	0.6699
BLEU with our method	0.5850	0.6757	0.4596	0.6272	0.5452	0.5325	0.5474

tween the correlation obtained by our method and the quality of chunking. In "Our method" shown in Tables 2–5, noun phrases for which some erroneous results obtained using the chunking tool were revised. "Our method II" of Tables 2–5 used noun phrases that were given as results obtained using the chunking tool. Underlining in "Our method II" of Tables 2–5 signifies that the differences between correlation coefficients obtained using our method II and IMPACT are statistically significant at the 5% significance level. Fundamentally, in both "Avg." and "All" of Tables 2–5, the correlation coefficients of our method II without the revised noun phrases are lower than those of our method using the revised noun phrases. However, the difference between our method and our method II in "Avg." and "All" of Tables 2–5 is not large. The performance of the chunking tool has no great influence on the results of our method because $score_{wd}$ in Eqs. (3), (4), and (5) do not depend strongly on the performance of the chunking tool. For example, in sentences shown in Fig. 2, all common parts are the same as the common parts of Fig. 2 when "the crowning fall" in the MT output and "crowning drop" in the reference are not determined as the noun phrases. Other common parts are determined correctly because the weight of the common part "the amount of" is higher than those of other common parts by Eqs. (1) and (2). Consequently, the determination of the common parts except "the amount of" is not difficult.

In other language sentences, we already performed the experiments using Japanese sentences from Reuters articles(Oyamada et al., 2010). Results show that the correlation coefficients of IMPACT with our method, for which IMPACT scores were used as $score_{wd}$ in Eq. (13), were highest among some methods. Therefore, our method might not be language-dependent. Nevertheless, experiments using various language data are necessary to elucidate this point.

4 Conclusion

As described herein, we proposed a new automatic evaluation method for machine transla-

Table 5: Spearman's rank correlation coefficient for sentence-level fluency.

	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7
Our method	**0.5697**	0.3299	0.5446	0.4199	0.5733	0.5060	**0.6459**
IMPACT	0.5481	0.3285	0.5572	0.3976	**0.5960**	0.4317	0.6334
ROUGE-L	0.5470	0.3041	**0.5646**	0.3661	0.5638	0.3879	0.6255
BLEU	0.4157	0.0559	0.4286	0.2018	0.4475	0.2569	0.4909
NIST	0.4209	0.0185	0.4559	0.1093	0.3186	0.1898	0.3634
NMG-WN	0.5569	**0.3461**	0.5381	**0.4300**	0.5052	**0.5264**	0.5328
METEOR	0.4608	0.1429	0.4438	0.1783	0.4073	0.1596	0.4821
WER	0.4469	0.2395	0.5087	0.3292	0.4995	0.3482	0.5637
Our method II	0.5659	0.3216	0.5484	0.3773	0.5638	0.5211	0.6343
BLEU with our method	0.5188	0.1534	0.4793	0.3005	0.5255	0.3942	0.5676
	No. 8	No. 9	No. 10	No. 11	No. 12	Avg.	All
Our method	0.5646	**0.6617**	0.3319	0.6256	0.4485	**0.5185**	**0.5556**
IMPACT	0.5471	0.6454	0.3222	0.6319	0.4358	0.5062	0.5489
ROUGE-L	0.5246	0.6428	0.2949	0.6159	0.3928	0.4858	0.5359
BLEU	0.4882	0.5419	0.1407	0.4740	0.4176	0.3633	0.3971
NIST	0.4150	0.4193	0.0889	0.3006	**0.4752**	0.2980	0.2994
NMG-WN	**0.5684**	0.5850	**0.4451**	**0.6502**	0.4387	0.5102	0.5156
METEOR	0.2911	0.4267	0.1735	0.3264	0.3512	0.3158	0.2886
WER	0.5320	0.6505	0.3828	0.6501	0.4003	0.4626	0.4193
Our method II	0.5609	0.6687	0.3629	0.6223	0.4384	0.5155	0.5531
BLEU with our method	0.5470	0.6213	0.2184	0.5808	0.4870	0.4495	0.4825

Table 6: Correlation coefficient for SMT and RBMT.

	Pearson's correlation coefficient				Spearman's rank correlation coefficient			
	Adequacy		Fluency		Adequacy		Fluency	
	SMT	RBMT	SMT	RBMT	SMT	RBMT	SMT	RBMT
Our method	**0.7054**	0.5840	**0.5477**	**0.5016**	**0.6710**	**0.5961**	**0.5254**	**0.5003**
IMPACT	0.6721	0.5650	0.5364	0.4960	0.6397	0.5811	0.5162	0.4951
ROUGE-L	0.6560	0.5691	0.5179	0.4988	0.6225	0.5701	0.4942	0.4783
NMG-WN	0.5958	**0.5850**	0.5201	0.4732	0.6129	0.5755	0.5238	0.4959

tion. Our method calculates the scores for MT outputs using noun-phrase chunking. Consequently, the system obtains scores using the correctly matched words and phrase-level information based on the corresponding noun phrases. Experimental results demonstrate that our method yields the highest correlation among eight methods in terms of sentence-level adequacy and fluency.

Future studies will improve our method, enabling it to achieve high correlation in sentence-level fluency. Future studies will also include experiments using data of various languages.

Acknowledgements

This work was done as research under the AAMT/JAPIO Special Interest Group on Patent Translation. The Japan Patent Information Organization (JAPIO) and the National Institute of Informatics (NII) provided corpora used in this work. The author gratefully acknowledges JAPIO and NII for their support. Moreover, this work was partially supported by Grants from the High-Tech Research Center of Hokkai-Gakuen University and the Kayamori Foundation of Informational Science Advancement.

References

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

Deborah Coughlin. 2003. Correlating Automated and Human Assessments of Machine Translation Quality. *In Proc. of MT Summit IX*, 63–70.

Hiroshi Echizen-ya and Kenji Araki. 2007. Automatic Evaluation of Machine Translation based on Recursive Acquisition of an Intuitive Common Parts Continuum. *In Proc. of MT Summit XII*, 151–158.

Hiroshi Echizen-ya, Terumasa Ehara, Sayori Shimohata, Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro and Noriko Kando. 2009. Meta-Evaluation of Automatic Evaluation Methods for Machine Translation using Patent Translation Data in NTCIR-7. *In Proc. of the 3rd Workshop on Patent Translation*, 9–16.

Terumasa Ehara. 2007. Rule Based Machine Translation Combined with Statistical Post Editor for Japanese to English Patent Translation. *In Proc. of MT Summit XII Workshop on Patent Translation*, 13–18.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro. 2008. Overview of the Patent Translation Task at the NTCIR-7 Workshop. *In Proc. of 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access*, 389–400.

Gregor Leusch, Nicola Ueffing and Hermann Ney. 2003. A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation. *In Proc. of MT Summit IX*, 240–247.

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

Dennis N. Mehay and Chris Brew. 2007. BLEUÂTRE: Flattening Syntactic Dependencies for MT Evaluation. *In Proc. of MT Summit XII*, 122–131.

Andrew Mutton, Mark Dras, Stephen Wan and Robert Dale. 2007. GLEU: Automatic Evaluation of Sentence-Level Fluency. *In Proc. of ACL'07*, 344–351.

NIST. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf*.

Takashi Oyamada, Hiroshi Echizen-ya and Kenji Araki. 2010. Automatic Evaluation of Machine Translation Using both Words Information and Comprehensive Phrases Information. *In IPSJ SIG Technical Report, Vol.2010-NL-195, No. 3 (in Japanese)*.

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

Michael Pozar and Eugene Charniak. 2006. Bllip: An Improved Evaluation Metric for Machine Translation. *Brown University Master Thesis*.

Fei Sha and Fernando Pereira. 2003. Shallow Parsing with Conditional Random Fields. *In Proc. of HLT-NAACL 2003*, 134–141.

Keh-Yih Su, Ming-Wen Wu and Jing-Shin Chang. 1992. A New Quantitative Quality Measure for Machine Translation Systems. *In Proc. of GOLING'92*, 433–439.

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