Automatic Evaluation of Translation Quality for Distant Language Pairs

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Katsuhito Sudoh, Hajime Tsukada
NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0237, Japan
{isozaki,hirao,kevinduh,sudoh,tsukada}@cslab.kecl.ntt.co.jp

Abstract

Automatic evaluation of Machine Translation (MT) quality is essential to developing high-quality MT systems. Various evaluation metrics have been proposed, and BLEU is now used as the de facto standard metric. However, when we consider translation between distant language pairs such as Japanese and English, most popular metrics (e.g., BLEU, NIST, PER, and TER) do not work well. It is well known that Japanese and English have completely different word orders, and special care must be paid to word order in translation. Otherwise, translations with wrong word order often lead to misunderstanding and incomprehensibility. For instance, SMT-based Japanese-to-English translators tend to translate 'A because B' as 'B because A.' Thus, word order is the most important problem for distant language translation. However, conventional evaluation metrics do not significantly penalize such word order mistakes. Therefore, locally optimizing these metrics leads to inadequate translations. In this paper, we propose an automatic evaluation metric based on rank correlation coefficients modified with precision. Our meta-evaluation of the NTCIR-7 PATMT JE task data shows that this metric outperforms conventional metrics.

1 Introduction

Automatic evaluation of machine translation (MT) quality is essential to developing high-quality machine translation systems because human evaluation is time consuming, expensive, and irreproducible. If we have a perfect automatic evaluation metric, we can tune our translation system for the metric.

BLEU (Papineni et al., 2002b; Papineni et al., 2002a) showed high correlation with human judgments and is still used as the de facto standard automatic evaluation metric. However, Callison-Burch et al. (2006) argued that the MT community is overly reliant on BLEU by showing examples of poor performance. For Japanese-to-English (JE) translation, Echizen-ya et al. (2009) showed that the popular BLEU and NIST do not work well by using the system outputs of the NTCIR-7 PATMT (patent translation) JE task (Fujii et al., 2008). On the other hand, ROUGE-L (Lin and Hovy, 2003), Word Error Rate (WER), and IMPACT (Echizen-ya and Araki, 2007) worked better.

In these studies, Pearson's correlation coefficient and Spearman's rank correlation ρ with human evaluation scores are used to measure how closely an automatic evaluation method correlates with human evaluation. This evaluation of automatic evaluation methods is called **meta-evaluation**. In human evaluation, people judge the adequacy and the fluency of each translation.

Denoual and Lepage (2005) pointed out that BLEU assumes word boundaries, which is ambiguous in Japanese and Chinese. Here, we assume the word boundaries given by ChaSen, one of the standard morphological analyzers (`http://chasen-legacy.sourceforge.jp/`) following Fujii et al. (2008)

In JE translation, most Statistical Machine Translation (SMT) systems translate the Japanese sentence

(J0) *kare wa sono hon wo yonda node sekaishi ni kyoumi ga atta*

which means

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 944–952,
MIT, Massachusetts, USA, 9-11 October 2010.

(R0) he was interested in world history because he read the book

into an English sentence such as

(H0) he read the book because he was interested in world history

in which the cause and the effect are swapped. Why does this happen? The former half of (J0) means "He read the book," and the latter half means "(he) was interested in world history." The middle word "node" between them corresponds to "because." Therefore, SMT systems output sentences like (H0). On the other hand, Rule-based Machine Translation (RBMT) systems correctly give (R0).

In order to find (R0), SMT systems have to search a very large space because we cannot restrict its search space with a small distortion limit. Most SMT systems thus fail to find (R0).

Consequently, the global word order is essential for translation between distant language pairs, and wrong word order can easily lead to misunderstanding or incomprehensibility. Perhaps, some readers do not understand why we emphasize word order from this example alone. A few more examples will clarify what happens when SMT is applied to Japanese-to-English translation. Even the most famous SMT service available on the web failed to translate the following very simple sentence at the time of writing this paper.

Japanese: *meari wa jon wo koroshita.*
Reference: Mary killed John.
SMT output: John killed Mary.

Since it cannot translate such a simple sentence, it obviously cannot translate more complex sentences correctly.

Japanese: *bobu ga katta hon wo jon wa yonda.*
Reference: John read a book that Bob bought.
SMT output: Bob read the book John bought.

Another example is:

Japanese: *bobu wa meari ni yubiwa wo kau tameni, jon no mise ni itta.*
Reference: Bob went to John's store to buy a ring for Mary.
SMT output: Bob Mary to buy the ring, John went to the store.

In this way, this SMT service usually gives incomprehensible or misleading translations, and thus people prefer RBMT services. Other SMT systems also tend to make similar word order mistakes, and special care should be paid to the translation between distant language pairs such as Japanese and English.

Even Japanese people cannot solve this word order problem easily: It is well known that Japanese people are not good at speaking English.

From this point of view, conventional automatic evaluation metrics of translation quality disregard word order mistakes too much. Single-reference BLEU is defined by a geometrical mean of n-gram precisions p_n and is modified by Brevity Penalty (BP) $\min(1, \exp(1 - r/h))$, where r is the length of the reference and h is the length of the hypothesis.

$$\mathrm{BLEU} = \mathrm{BP} \times (p_1 p_2 p_3 p_4)^{1/4}.$$

Its range is [0, 1]. The BLEU score of (H0) with reference (R0) is $1.0 \times (11/11 \times 9/10 \times 6/9 \times 4/8)^{1/4} = 0.740$. Therefore, BLEU gives a very good score to this inadequate translation because it checks only n-grams and does not regard global word order.

Since (R0) and (H0) look similar in terms of fluency, **adequacy** is more important than fluency in the translation between distant language pairs.

Similarly, other popular scores such as NIST, PER, and TER (Snover et al., 2006) also give relatively good scores to this translation. NIST also considers only local word orders (n-grams). PER (Position-Independent Word Error Rate) was designed to disregard word order completely. TER (Snover et al., 2006) was designed to allow phrase movements without large penalties. Therefore, these standard metrics are not optimal for evaluating translation between distant language pairs.

In this paper, we propose an alternative automatic evaluation metric appropriate for distant language pairs. Our method is based on **rank correlation coefficients**. We use them to compare the word ranks in the reference with those in the hypothesis.

There are two popular rank correlation coefficients: Spearman's ρ and Kendall's τ (Kendall, 1975). In Isozaki et al. (2010), we used Kendall's τ to measure the effectiveness of our **Head Finalization** rule as a preprocessor for English-to-Japanese translation, but we measured the quality of translation by using conventional metrics.

It is not clear how well τ works as an automatic evaluation metric of translation quality. Moreover, Spearman's ρ might work better than Kendall's τ. As we discuss later, τ considers only the direction of the rank change, whereas ρ considers the distance of the change.

The first objective of this paper is to examine which is the better metric for distant language pairs. The second objective is to find improvements of these rank correlation-metrics.

Spearman's ρ is based on Pearson's correlation coefficients. Suppose we have two lists of numbers

$$\begin{aligned} \boldsymbol{x} &= [0.1, 0.4, 0.2, 0.6], \\ \boldsymbol{y} &= [0.9, 0.6, 0.2, 0.7]. \end{aligned}$$

To obtain Pearson's coefficients between $\boldsymbol{x}$ and $\boldsymbol{y}$, we use the raw values in these lists. If we substitute their ranks for their raw values, we get

$$\boldsymbol{x'} = [1, 3, 2, 4] \text{ and } \boldsymbol{y'} = [4, 2, 1, 3].$$

Then, Spearman's ρ between $\boldsymbol{x}$ and $\boldsymbol{y}$ is given by Pearson's coefficients between $\boldsymbol{x'}$ and $\boldsymbol{y'}$. This ρ can be rewritten as follows when there is no tie:

$$\rho = 1 - \frac{\sum_i d_i^2}{{}_{n+1}C_3}.$$

Here, d_i indicates the difference in the ranks of the i-th element. Rank distances are **squared** in this formula. Because of this square, we expect that ρ decreases drastically when there is an element that significantly changes in rank. But we are also afraid that ρ may be too severe for alternative good translations.

Since Pearson's correlation metric assumes linearity, nonlinear monotonic functions can change its score. On the other hand, Spearman's ρ and Kendall's τ uses ranks instead of raw evaluation scores, and simple application of monotonic functions cannot change them (use of other operations such as averaging sentence scores can change them).

2 Methodology

2.1 Word alignment for rank correlations

We have to determine word ranks to obtain rank correlation coefficients. Suppose we have:

(R1) `John hit Bob yesterday`
(H1) `Bob hit John yesterday`

The 1st word "John" in R1 becomes the **3**rd word in H1. The 2nd word "hit" in R1 becomes the **2**nd word in H1. The 3rd word "Bob" in R1 becomes the **1**st word in H1. The 4th word "yesterday" in R1 becomes the **4**th word in H1. Thus, we get H1's word order list [3, 2, 1, 4]. The number of all pairs of integers in this list is ${}_4C_2 = 6$. It has three increasing pairs: (3,4), (2,4), and (1,4). Since Kendall's τ is given by:

$$\tau = 2 \times \frac{\text{the number of increasing pairs}}{\text{the number of all pairs}} - 1,$$

H1's τ is $2 \times 3/6 - 1 = 0.0$.

In this case, we can obtain Spearman's ρ as follows: "John" moved by $d_1 = 2$ words, "hit" moved by $d_2 = 0$ words, "Bob" moved by $d_3 = 2$ words, and "yesterday" moved by $d_4 = 0$ words. Therefore, H1's ρ is $1 - (2^2 + 0^2 + 2^2 + 0^2)/{}_5C_3 = 0.2$.

Thus, τ considers only the direction of the movement, whereas ρ considers the distance of the movement. Both ρ and τ have the same range $[-1, 1]$. The main objective of this paper is to clarify which rank correlation is closer to human evaluation scores.

We have to consider the limitation of the rank correlation metrics. They are defined only when there is **one-to-one correspondence**. However, a reference sentence and a hypothesis sentence may have different numbers of words. They may have two or more occurrences of the same word in one sentence. Sometimes, a word in the reference does not appear in the hypothesis, or a word in the hypothesis does not appear in the reference. Therefore, we cannot calculate τ and ρ following the above definitions in general.

Here, we determine the correspondence of words between hypotheses and references as follows. First, we find one-to-one corresponding words. That is, we find words that appear in both sentences and only once in each sentence. Suppose we have:

(R2) `the boy read the book`
(H2) `the book was read by the boy`

By removing non-aligned words by one-to-one correspondence, we get:

(R3) `boy read book`
(H3) `book read boy`

Thus, we lost "`the`." We relax this one-to-one correspondence constraint by using one-to-one corresponding **bigrams**. (R2) and (H2) share "`the boy`" and "`the book`," and we can align these instances of "the" correctly.

(R4) $\texttt{the}_1\ \texttt{boy}_2\ \texttt{read}_3\ \texttt{the}_4\ \texttt{book}_5$
(H4) $\texttt{the}_4\ \texttt{book}_5\ \texttt{read}_3\ \texttt{the}_1\ \texttt{boy}_2$

Now, we have five aligned words, and H4's word order is represented by [4, 5, 3, 1, 2].

In returning to H0 and R0, we find that each of these sentences has eleven words. Almost all words are aligned by one-to-one correspondence but "he" is not aligned because it appears twice in each sentence. By considering one-to-one corresponding bigrams ("he was" and "he read"), "he" is aligned as follows.

(R5) $\texttt{he}_1\ \texttt{was}_2\ \texttt{interested}_3\ \texttt{in}_4\ \texttt{world}_5\ \texttt{history}_6\ \texttt{because}_7\ \texttt{he}_8\ \texttt{read}_9\ \texttt{the}_{10}\ \texttt{book}_{11}$
(H5) $\texttt{he}_8\ \texttt{read}_9\ \texttt{the}_{10}\ \texttt{book}_{11}\ \texttt{because}_7\ \texttt{he}_1\ \texttt{was}_2\ \texttt{interested}_3\ \texttt{in}_4\ \texttt{world}_5\ \texttt{history}_6$

H5's word order is [8, 9, 10, 11, 7, 1, 2, 3, 4, 5, 6]. The number of increasing pairs is: ${}_4C_2 = 6$ pairs in [8, 9, 10, 11] and ${}_6C_2 = 15$ pairs in [1, 2, 3, 4, 5, 6]. Then we obtain $\tau = 2 \times (6 + 15)/{}_{11}C_2 - 1 = -0.236$. On the other hand, $\sum_i d_i^2 = 5^2 \times 6 + 2^2 + 7^2 \times 4 = 350$, and we obtain $\rho = 1 - 350/{}_{12}C_3 = -0.591$.

Therefore, both Spearman's ρ and Kendall's τ give very bad scores to the misleading translation H0. This fact implies they are much better metrics than BLEU, which gave a good score to it. ρ is much lower than τ as we expected.

In general, we can use higher-order n-grams for this alignment, but here we use only unigrams and bigrams for simplicity. This algnment algorithm is given in Figure 1. Since some hypothesis words do not have corresponding reference words, the output integer list `worder` is sometimes shorter than the evaluated sentence. Therefore, we should not use $\texttt{worder}[i] - i$ as d_i directly. We have to renumber the list by rank as we did in Section 1.

Read a hypothesis sentence $h = h_1 h_2 \ldots h_m$
and its reference sentence $r = r_1 r_2 \ldots r_n$.

Initialize `worder` with an empty list.

For each word h_i in h:

- If h_i appears only once each in h and r, append j s.t. $r_j = h_i$ to `worder`.
- Otherwise, if the bigram $h_i h_{i+1}$ appears only once each in h and r, append j s.t. $r_j r_{j+1} = h_i h_{i+1}$ to `worder`.
- Otherwise, if the bigram $h_{i-1} h_i$ appears only once each in h and r, append j s.t. $r_{j-1} r_j = h_{i-1} h_i$ to `worder`.

Return `worder`.

Figure 1: Word alignment algorithm for rank correlation

2.2 Word order metrics and meta-evaluation metrics

These rank correlation metrics sometimes have negative values. In order to make them just like other automatic evaluation metrics, we normalize them as follows.

- Normalized Kendall's τ: NKT = $(\tau + 1)/2$.
- Normalized Spearman's ρ: NSR = $(\rho + 1)/2$.

Accordingly, NKT is 0.382 and NSR is 0.205.

These metrics are defined only when the number of aligned words is two or more. We define both NKT and NSR as zero when the number is one or less. Consequently, these normalized metrics have the same range [0, 1].

In order to avoid confusion, we use these abbreviations (NKT and NSR) when we use rank correlations as **word order metrics**, because these correlation metrics are also used in the machine translation community for **meta-evaluation**. For meta-evaluation, we use Spearman's ρ and Pearson's correlation coefficient and call them "Spearman" and "Pearson," respectively.

2.3 Overestimation problem

Since we measure the rank correlation of only corresponding words, these metrics will overestimate the correlation. For instance, a hypothesis sentence might have only two corresponding words among

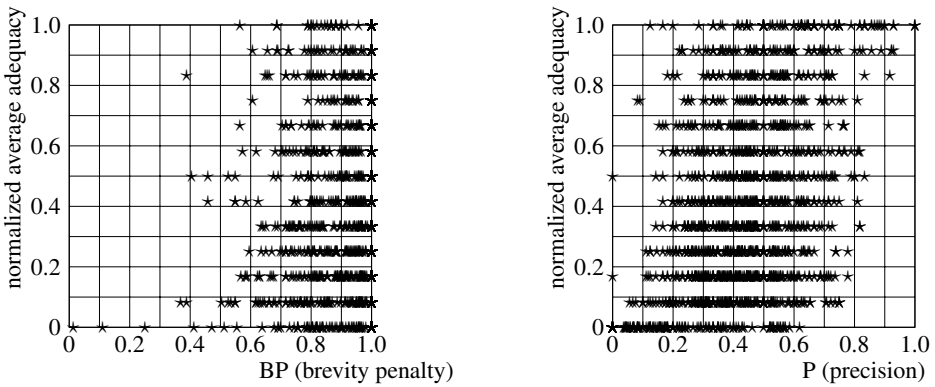

Figure 2: Scatter plots of normalized average adequacy with brevity penalty (left) and precision (right). (Each $\star$ corresponds to one sentence generated by one MT system)

dozens of words. In this case, these two words determine the score of the whole sentence. If the two words appear in their order in the reference, the whole sentence obtains the best score, NSR = NKT = 1.0, in spite of the fact that only two words matched.

Solving this overestimation problem is the second objective of this paper. BLEU uses "Brevity Penalty (BP)" (Section 1) to reduce the scores of too-short sentences. We can combine the above word order metrics with BP, e.g., $\mathrm{NKT} \times \mathrm{BP}$ and $\mathrm{NSR} \times \mathrm{BP}$.

However, we cannot very much expect from this solution because BP scores do not correlate with human judgments well. The left graph of Figure 2 shows a scatter plot of BP and "normalized average adequacy." This graph has 15 (systems) $\times$ 100 (sentences) dots. Each dot ($\star$) corresponds to one sentence from one translation system.

In the NTCIR-7 data, three human judges gave five-point scores (1, 2, 3, 4, 5) for "adequacy" and "fluency" of each translated sentence. Although each system translated 1,381 sentences, only 100 sentences were evaluated by the judges.

For each translated sentence, we averaged three judges' adequacy scores and normalized this average x by $(x-1)/4$. This is our "normalized average adequacy," and the dots appears only at multiples of $1/3 \times 1/4$.

This graph shows that BP has very little correlation with adequacy, and we cannot expect BP to improve the meta-evaluation performance very much. Perhaps, BP's poor performance was caused by the fact that most MT systems output almost the same number of words, and if the number exceeds the length of the reference, BP=1.0 holds.

Therefore, we have to consider other modifiers for this overestimation problem. We can use other common metrics such as precision, recall, and F-measure to reduce the overestimation of NSR and NKT.

- Precision: $P = c/|h|$, where c is the number of corresponding words and $|h|$ is the number of words in the hypothesis sentence h.
- Recall: $R = c/r$, where $|r|$ is the number of words in the reference sentence r.
- F-measure: $F_\beta = (1+\beta^2)PR/(\beta^2 P + R)$, where β is a parameter.

In (R2)&(H2)'s case, precision is 5/7 = 0.714 and recall is 5/5 = 1.000.

Which metric should we use? Our preliminary experiments with NTCIR-7 data showed that **precision correlated best with adequacy** among these three metrics (P, R, and $F_{\beta=1}$). In addition, BLEU is essentially made for precision. Therefore, precision seems the most promising modifier.

The right graph of Figure 2 shows a scatter plot of precision and normalized average adequacy. The graph shows that precision has more correlation with adequacy than BP. We can observe that sentences with very small P values usually obtain very low adequacy scores but those with mediocre P values often obtain good adequacy scores.

If we multiply P directly by NSR or NKT, those sentences with mediocre P values will lose too much of their scores. The use of $\sqrt{x}$ will mitigate this problem. Since $\sqrt{P}$ is closer to 1.0 than P itself, multiplication of $\sqrt{P}$ instead of P itself will save these sentences. If we apply $\sqrt{x}$ twice ($\sqrt{\sqrt{P}} = \sqrt[4]{P}$), it will further save them. Therefore, we expect $\times\sqrt{P}$ and $\times\sqrt[4]{P}$ to work better than $\times P$. Now, we propose two new metrics:

$$\text{NSR}P^{\alpha} \quad \text{and} \quad \text{NKT}P^{\alpha},$$

where α is a parameter ($0 \leq \alpha \leq 1$).

3 Experiments

3.1 Meta-evaluation with NTCIR-7 data

In order to compare automatic translation evaluation methods, we use submissions to the NTCIR-7 Patent Translation (PATMT) task (Fujii et al., 2008). Fourteen MT systems participated in the Japanese-English intrinsic evaluation. There were two Rule-Based MT (RMBT) systems and one Example-based MT (EBMT) system. All other systems were Statistical MT (SMT) systems. The task organizers provided a baseline SMT system. These 15 systems translated 1,381 Japanese sentences into English. The organizers evaluated these translations by using BLEU and human judgments. In the human judgements, three experts independently evaluated 100 selected sentences in terms of '*adequacy*' and '*fluency*.'

For automatic evaluation, we used a single reference sentence for each of these 100 manually evaluated sentences. Echizen-ya et al. (2009) used multi-reference data, but their data is not publicly available yet.

For this meta-evaluation, we measured the *corpus-level* correlation between the human evaluation scores and the automatic evaluation scores. We simply averaged scores of 100 sentences for the proposed metrics. For existing metrics such as BLEU, we followed their definitions for corpus-level evaluation instead of simple averages of sentence-level scores. We used default settings for conventional metrics, but we tuned GTM (Melamed et al., 2007) with `-e` option. This option controls preferences on longer word runs. We also used the paraphrase database TERp (`http://www.umiacs.umd.edu/~snover/terp`) for METEOR (Banerjee and Lavie, 2005).

3.2 Meta-evaluation with WMT-07 data

We developed our metric mainly for automatic evaluation of translation quality for distant language pairs such as Japanese-English, but we also want to know how well the metric works for similar language pairs. Therefore, we also use the WMT-07 data (Callison-Burch et al., 2007) that covers only European language pairs. Callison-Burch et al. (2007) tried different human evaluation methods and showed detailed evaluation scores. The Europarl test set has 2,000 sentences, and The News Commentary test set has 2,007 sentences.

This data has different language pairs: Spanish, French, German $\Rightarrow$ English. We exclude Czech-English because there were so few systems (See the footnote of p. 146 in their paper).

4 Results

4.1 Meta-evaluation with NTCIR-7 data

Table 1 shows the main results of this paper. The left part has corpus-level meta-evaluation with **adequacy**. Error metrics, WER, PER, and TER, have negative correlation coefficients, but we did not show their minus signs here.

Both NSR-based metrics and NKT-based metrics perform better than conventional metrics for this NTCIR PATMT JE translation data. As we expected, $\times$BP and $\times P^{(1/1)}$ performed badly. Spearman of BP itself is zero.

NKT performed slightly better than NSR. Perhaps, NSR penalized alternative good translations too much. However, one of the NSR-based metrics, $\text{NSR}P^{1/4}$, gave the **best Spearman score of 0.947**, and the difference between $\text{NSR}P^{\alpha}$ and $\text{NKT}P^{\alpha}$ was small. Modification with P led to this improvement.

NKT gave the best Pearson score of 0.922. However, Pearson measures linearity and we can change its score through a nonlinear monotonic function without changing Spearman very much. For instance, $(\text{NSR}P^{1/4})^{1.5}$ also has **Spearman of 0.947** but its **Pearson is 0.931**, which is better than NKT's 0.922. Thus, we think Spearman is a better **meta**-evaluation metric than Pearson.

Table 1: NTCIR-7 Meta-evaluation: correlation with human judgments (Spm = Spearman, Prs = Pearson)

human judge	**Adequacy**		Fluency	
eval\ meta-eval	Spm	Prs	Spm	Prs
P	0.615	0.704	0.672	0.876
R	0.436	0.669	0.461	0.854
$F_{\beta=1}$	0.525	0.692	0.543	0.871
BP	0.000	0.515	-0.007	0.742
NSR	0.904	0.906	0.869	0.910
NSR$P^{1/8}$	0.937	0.905	0.890	0.934
NSR$P^{1/4}$	**0.947**	0.900	0.901	0.944
NSR$P^{1/2}$	0.937	0.890	0.926	0.949
NSR$P^{1/1}$	0.883	0.872	0.883	0.939
NSR $\times$ BP	0.851	0.874	0.769	0.910
NKT	0.940	0.922	0.887	0.931
NKT$P^{1/8}$	0.940	0.913	0.908	0.944
NKT$P^{1/4}$	0.940	0.904	0.908	0.949
NKT$P^{1/2}$	0.929	0.890	0.897	0.949
NKT$P^{1/1}$	0.897	0.869	0.879	0.936
NKT $\times$ BP	0.829	0.878	0.726	0.918
ROUGE-L	0.903	0.874	0.889	0.932
ROUGE-S(4)	0.593	0.757	0.640	0.869
IMPACT	0.797	0.813	0.751	0.932
WER	0.894	0.822	0.836	0.926
TER	0.854	0.806	0.372	0.856
PER	0.375	0.642	0.393	0.842
METEOR(TERp)	0.490	0.708	0.508	0.878
GTM(-e 12)	0.618	0.723	0.601	0.850
NIST	0.343	0.661	0.372	0.856
BLEU	0.515	0.653	0.500	0.795

The right part of Table 1 shows correlation with fluency, but **adequacy is more important**, because our motivation is to provide a metric that is useful to reduce incomprehensible or misunderstanding outputs of MT systems. Again, the correlation-based metrics gave better scores than conventional metrics, and BP performed badly. NSR-based metrics proved to be as good as NKT-based metrics.

Meta-evaluation scores of the de facto standard BLEU is much lower than those of other metrics. Echizen-ya et al. (2009) reported that IMPACT performed very well for *sentence-level* evaluation of NTCIR-7 PATMT JE data. This *corpus-level* result also shows that IMPACT works better than BLEU, but ROUGE-L, WER, and our methods give better scores than IMPACT.

Table 2: WMT-07 meta-evaluation: Each source language has two columns: the left one is News Corpus and the right one is Europarl.

Spearman's ρ with human "rank"			
source	French	Spanish	German
NSR	0.775 0.837	0.523 0.766	0.700 0.593
NSR$P^{1/8}$	0.821 0.857	0.786 0.595	0.400 0.685
NSR$P^{1/4}$	0.821 0.857	0.786 0.455	0.400 0.714
NSR$P^{1/2}$	0.821 0.857	0.786 0.347	0.400 0.714
NKT	0.845 0.857	0.607 0.838	0.700 0.630
NKT$P^{1/8}$	0.793 0.857	0.786 0.595	0.400 0.714
NKT$P^{1/4}$	0.793 0.857	0.786 0.524	0.400 0.714
NKT$P^{1/2}$	0.793 0.857	0.786 0.347	0.400 0.714
BLEU	0.786 0.679	0.750 0.595	0.400 0.821
WER	0.607 0.857	0.750 0.429	0.000 0.500
ROUGEL	0.893 0.739	0.786 0.707	0.700 0.857
ROUGES	0.883 0.679	0.786 0.690	0.400 0.929

4.2 Meta-evaluation with WMT-07 data

Callison-Burch et al. (2007) have performed different human evaluation methods for different language pairs and different corpora. Their Table 5 shows inter-annotator agreements for the human evaluation methods. According to their table, the "sentence ranking" (or "rank") method obtained better agreement than "adequacy." Therefore, we show Spearman's ρ for "rank." We used the scores given in their Tables 9, 10, and 11. (The "constituent" methods obtained the best inter-annotator agreement, but these methods focus on local translation quality and have nothing to do with global word order, which we are discussing here.)

Table 2 shows that our metrics designed for distant language pairs are comparable to conventional methods even for similar language pairs, but ROUGE-L and ROUGE-S performed better than ours for French News Corpus and German Europarl. BLEU scores in this table agree with those in Table 17 of Callison-Burch et al. (2007) within rounding errors.

After some experiments, we noticed that the use of R instead of P often gives better scores for WMT-07, but it degrades NTCIR-7 scores. We can extend our metric by F_β, weighted harmonic mean of P and R, or any other interpolation, but the introduction of new parameters into our metric makes it difficult

to control. Improvement without new parameters is beyond the scope of this paper.

5 Discussion

It has come to our attention that Birch et al. (2010) has independently proposed an automatic evaluation method based on Kendall's τ. First, they started with Kendall's τ distance, which can be written as "$1 - \mathrm{NKT}$" in our terminology, and then subtracted it from one. Thus, their metric is nothing but NKT.

Then, they proposed application of the square root to get better Pearson by improving "the sensitivity to small reorderings." Since they used "Kendall's τ" and "Kendall's τ distance" interchangeably, it is not clear what they mean by "$\sqrt{\text{Kendall's } \tau}$," but perhaps they mean $1 - \sqrt{1 - \mathrm{NKT}}$ because $\sqrt{\mathrm{NKT}}$ is more insensitive to small reorderings. Table 3 shows the performance of these metrics for NTCIR-7 data. Pearson's correlation coefficient with adequacy was improved by $1 - \sqrt{1 - \mathrm{NKT}}$, but other scores were degraded in this experiment.

The difference between our method and Birch et al. (2010)'s method comes from the fact that we used Japanese-English translation data and Spearman's correlation for meta-evaluation, whereas they used Chinese-English translation data and only Pearson's correlation for meta-evaluation. Chinese word order is different from English, but Chinese is a Subject-Verb-Object (SVO) language and thus is much closer to English word order than Japanese, a typical SOV language.

We preferred NSR because it penalizes global word order mistakes much more than does NKT, and as discussed above, global word order mistakes often lead to incomprehensibility and misunderstanding.

On the other hand, they also tried Hamming distance, and summarized their experiments as follows:

> However, the Hamming distance seems to be more informative than Kendall's tau for small amounts of reordering.

This sentence and the introduction of the square root to NKT imply that Chinese word order is close to that of English, and they have to measure subtle word order mistakes.

Table 3: NTCIR-7 meta-evaluation: Effects of square root ($b(x) = 1 - \sqrt{1 - x}$)

	NKT	$\sqrt{\mathrm{NKT}}$	b(NKT)
Spearman w/ adequacy	0.940	0.940	0.922
Pearson w/ adequacy	0.922	0.817	0.941
Spearman w/ fluency	0.887	0.865	0.858
Pearson w/ fluency	0.931	0.917	0.833

In spite of these differences, the two groups independently recognized the usefulness of rank correlations for automatic evaluation of translation quality for distant language pairs.

In their WMT-2010 paper (Birch and Osborne, 2010), they multiplied NKT with the brevity penalty and interpolated it with BLEU for the WMT-2010 shared task. This fact implies that incomprehensible or misleading word order mistakes are rare in translation among European languages.

6 Conclusions

When Statistical Machine Translation is applied to distant language pairs such as Japanese and English, word order becomes an important problem. SMT systems often fail to find an appropriate translation because of a large search space. Therefore, they often output misleading or incomprehensible sentences such as "A because B" vs. "B because A." To penalize such inadequate translations, we presented an automatic evaluation method based on rank correlation. There were two questions for this approach. First, which correlation coefficient should we use: Spearman's ρ or Kendall's τ? Second, how should we solve the overestimation problem caused by the nature of one-to-one correspondence?

We answered these questions through our experiments using the NTCIR-7 PATMT JE translation data. For the first question, τ was slightly better than ρ, but ρ was improved by precision. For the second question, it turned out that BLEU's Brevity Penalty was counter-productive. A precision-based penalty gave a better solution. With this precision-based penalty, both ρ and τ worked well and they outperformed conventional methods for NTCIR-7 data. For similar language pairs, our method was comparable to conventional evaluation methods. Fu-

ture work includes extension of the method so that it can outperform conventional methods even for similar language pairs.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for MT evaluation with improved correlation with human judgements. In *Proc. of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and Summarization*, pages 65–72.

Alexandra Birch and Miles Osborne. 2010. LRscore for evaluating lexical and reordering quality in MT. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 327–332.

Alexandra Birch, Miles Osborne, and Phil Blunsom. 2010. Metrics for MT evaluation: evaluating reordering. *Machine Translation*, 24(1):15–26.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluatiing the role of Bleu in machine translation research. In *Proc. of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Chrstof Monz, and Josh Schroeder. 2007. (Meta-)Evaluation of machine translation. In *Proc. of the Workshop on Machine Translation (WMT)*, pages 136–158.

Etienne Denoual and Yves Lepage. 2005. BLEU in characters: towards automatic MT evaluation in languages without word delimiters. In *Companion Volume to the Proceedings of the Second International Joint Conference on Natural Language Processing*, pages 81–86.

Hiroshi Echizen-ya and Kenji Araki. 2007. Automatic evaluation of machine translation based on recursive acquisition of an intuitive common parts continuum. In *Proceedings of MT Summit XII Workshop on Patent Translation*, pages 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 *Proceedings of the 3rd Workshop on Patent Translation*, pages 9–16.

Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2008. Overview of the patent translation task at the NTCIR-7 workshop. In *Working Notes of the NTCIR Workshop Meeting (NTCIR)*, pages 389–400.

Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh. 2010. Head Finalization: A simple reordering rule for SOV languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 250–257.

Maurice G. Kendall. 1975. *Rank Correlation Methods*. Charles Griffin.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 71–78.

Dan Melamed, Ryan Green, and Joseph P. Turian. 2007. Precision and recall of machine translation. In *Proc. of NAACL-HLT*, pages 61–63.

Kishore Papineni, Salim Roukos, Todd Ward, John Henderson, and Florence Reeder. 2002a. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish Results. In *Proc. of the International Conference on Human Language Technology Research (HLT)*, pages 132–136.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, pages 311–318.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*.