Paraphrasing Rules for Automatic Evaluation of Translation into Japanese

KANAYAMA Hiroshi
Tokyo Research Laboratory, IBM Japan, Ltd.
1623-14 Shimo-tsuruma, Yamato-shi, Kanagawa 242-8502, Japan
kanayama@trl.ibm.com

Abstract

Automatic evaluation of translation quality has proved to be useful when the target language is English. In this paper the evaluation of translation into Japanese is studied. An existing method based on n-gram similarity between translations and reference sentences is difficult to apply to the evaluation of Japanese because of the agglutinativeness and variation of semantically similar expressions in Japanese. The proposed method applies a set of paraphrasing rules to the reference sentences in order to increase the similarity score for the expressions that differ only in their writing styles. Experimental results show the paraphrasing rules improved the correlation between automatic evaluation and human evaluation from 0.80 to 0.93.

1 Introduction

Evaluating natural language processing applications' output is important both for users and developers. Tasks such as sentential parsing, morphological analysis and named entity recognition are easy to evaluate automatically because the "right answer" can be defined deterministically under a specific grammar or assumed criterion.

The evaluation of machine translation is not so straightforward since there are infinite ways to output similar meanings and one can not enumerate the right answers exhaustively. In spite of that, automatic translation evaluation is practically important because the evaluation is laborious work for humans and evaluation by humans tends to be arbitrary. Automatic evaluation is more reliable than human evaluation because of its consistency for the same translations.

BLEU (Papineni et al., 2002b) is one of the methods for automatic evaluation of translation quality. It uses the ratio of co-occurring n-grams between a translation and single or multiple reference sentences. High correlation is reported between the BLEU score and human evaluations for translations from Arabic, Chinese, French, and Spanish to English (Papineni et al., 2002a).

This paper investigates how to apply BLEU to the evaluation of English-to-Japanese translation. The main goal of this paper is to design a reliable method of evaluation for translations from another language to Japanese (henceforth we call this *Japanese translation evaluation*). There are some difficulties in adjusting BLEU for Japanese: BLEU uses n-grams of words, so words in a sentence are assumed to be separated by spaces, while Japanese does not use spaces between words. Moreover, Japanese has more variation in writing styles than English. A major difference in these languages is that Japanese has polite forms expressed by inflections or auxiliary verbs. If the style of the translations is not the same as that of the reference sentences, the evaluation score becomes low even though the translations are accurate in their meanings and grammar. To solve these problems, we apply paraphrasing rules to the reference sentences so that the differences in writing styles do not affect the evaluation score.

Another goal is derived from this application of paraphrasing: to define a "good paraphrase". Here paraphrasing means rewriting sentences without changing their semantics. Several methods of paraphrasing have been studied. Some of them aim at the preprocessing of machine translation (Mitamura and Nyberg, 2001; Takahashi et al., 2001). They use paraphrasing to transform the input sentences so that the language-transferring routines can handle them easily. Another application of paraphrasing is to canonicalize many expressions that have the same semantics, supporting information retrieval or question answering (Zukerman and Raskutti, 2002; Torisawa, 2002). Paraphrasing techniques in these studies are considered to be useful, but they are difficult to evaluate.

Machine translation evaluation requires methods to judge whether two sentences have the same meaning even when they are syntactically different. Therefore if a set of paraphrasing rules contributes to more reliable translation evaluation, it can be said to be "good" paraphrasing. Thus the study in this paper also presents a new paradigm for evaluating paraphrases.

Section 2 overviews the BLEU metric. Section 3 presents the proposed method of Japanese translation evaluation, and its performance is evaluated in Section 4. Based on the experimental results, Section 5 discusses qualitative and quantitative features of paraphrasing.

2 Background: Overview of BLEU

This section briefly describes the original BLEU (Papineni et al., 2002b)[1], which was designed for English translation evaluation, so English sentences are used as examples in this section.

2.1 N-gram precision

BLEU evaluation uses a parallel corpus which consists of sentences in the source language and their translations to the target language by professional translators. We call the professional translations *reference sentences*. It is preferable if the corpus has multiple reference sentences translated by multiple translators for each source sentence.

Sentences in the source language are also translated by the translation systems to be evaluated. The translations are called *candidate sentences*. Below is an example.

[1]See the cited paper for more detailed definitions.

Example 1

Reference 1
I had my watch repaired by an office worker.

Reference 2
A person in the office repaired my watch.

Candidate 1
I had a man in the office repair a watch.

Candidate 2
I had the person of an office correct a clock.

The BLEU score is based on n-gram precision shown in Equation (1). It is the ratio of n-grams which appear both in the candidate sentence and in at least one of the reference sentences, among all n-grams in the candidate sentence.

$$p_n = \frac{\sum_{s \in cand} \sum_{ngr \in s} \min(C(ngr), C_r(ngr))}{\sum_{s \in cand} \sum_{ngr \in s} C(ngr)} \qquad (1)$$

$cand$: candidates s : sentence ngr : n-gram
C : count in the candidate sentence
C_r : count in a corresponding reference sentence

Candidate 1 in Example 1 contains 11 unigrams including punctuation. 8 unigrams out of these also appear in Reference 1 or Reference 2: 'I', 'had', 'a', 'in', 'the', 'office', 'watch' and '.', therefore, the unigram precision of Candidate 1 is 8/11. The bigram precision is 4/10 since 'I had', 'in the', 'the office' and 'watch .' are found. The only matched trigram is 'in the office', so the trigram precision is 1/9.

On the other hand, the unigram, bigram, and trigram precisions of Candidate 2 are 8/11, 2/10, 0/9, respectively, which are lower than those of Candidate 1. Indeed Candidate 1 is a better English translation than Candidate 2.

In practice the n-gram precision is calculated not for each sentence but for all of the sentences in the corpus.

2.2 Brevity Penalty

The n-gram precision is calculated by dividing the number of matched n-grams by the number of n-grams in the candidate sentence. Therefore, a short candidate sentence which consists only of frequently used words can score a high n-gram precision. For example, if the candidate sentence is just "The", its unigram precision is 1.0 if one of reference sentences contains at least one 'the', and that is usually true.

To penalize such a meaningless translation, the BLEU score is multiplied by the *brevity penalty* shown in (2).

$$\mathrm{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ e^{(1-r/c)} & \text{if} \quad c \leq r \end{cases} \tag{2}$$

where c and r are the total numbers of words in the candidate sentences and the reference sentences which have the closest numbers of words in each parallel sentence.

2.3 BLEU score

The BLEU score is calculated by Equation (3) below. It is the geometric average of the n-gram precisions multiplied by the brevity penalty. The geometric average is used because p_n decreases exponentially as n increases. The BLEU score ranges between 0 and 1.

$$\mathrm{B{\scriptsize LEU}} = \mathrm{BP} \cdot \left(\prod_{n=1}^{N} p_n \right)^{\frac{1}{N}} \tag{3}$$

The evaluations use unigrams up to N-grams. If a large n is used, the fluency of the sentences becomes a more important factor than the correctness of the words. Empirically the BLEU score has a high correlation with human evaluation when $N = 4$ for English translation evaluations (Papineni et al., 2002b).

3 Japanese Version of BLEU and Its Extension

This section describes how to adapt BLEU for Japanese translation evaluation. The adaptation consists of three steps.

3.1 Use of Morphological Analyzer

The first modification is mandatory for using the n-gram metric as in the original BLEU implementation. Since Japanese has no spaces between words, the words have to be separated by morphological analysis as in Example 2.

Example 2

Kare	*ga*	*hon*	*wo*	*yo*	*mi*	*mashi*	*ta*	.
He	SUBJ	book	ACC	read	INF	POLITE	PAST	.

'He read a book.'

3.2 Distinguish between Different Parts-of-speech

Many English words can be used as various parts-of-speech (POSs), but BLEU doesn't distinguish between the words with the same surface form in terms of their POSs, since the sentences are not processed by a tagger, so the system can't handle POSs. This doesn't cause a problem because most of the multi-POS words have conceptually similar meanings, as exemplified by the adverb 'fast' and the adjective 'fast' which have the same basic concept, so matching them between the candidate and references reasonably reflects the quality of the translation.

On the other hand, Japanese homonyms tend to be completely different if their POSs are different. For example, the postpositional phrasal particle '*ga*' and the connective particle '*ga*' should be distinguished from one another since the former acts as a subject case marker, while the latter connects two clauses that normally contradict each other. Fortunately the morphological analyzer outputs POS information when the sentence is separated into words, and therefore the words are also distinguished by their POSs in the described method.

3.3 Paraphrasing Rules

Example 3 is another possible translation of the source sentence of Example 2.

Example 3

Kare	*ga*	*hon*	*wo*	*yo*	*n*	*da*	.
He	SUBJ	book	ACC	read	INF-EUPH	PAST	.

'He read a book.'

The only difference here is the ending of the sentence has a less polite form. However, when Example 2 is the only reference sentence, the BLEU evaluation of Example 3 does not score high: 6/8 for unigrams, 4/7 for bigrams, 3/6 for trigrams, and 2/5

1	\$1(verb-c) : *n* : *da*	↔	\$1 : *mi* : *masi* : *ta*
2	*nai*(adj) : .	↔	*ari* : *mase* : *n* : .
3	\$1(noun) : *da*	↔	\$1 : *dearu*
4	*ni* : *yo* : *t* : *te*	↔	*ni* : *yo* : *ri*

Table 1: Examples of paraphrasing rules. \$1 denotes a wild card shared by both sides. ':' is a boundary of morphemes. '(verb-c)' means a consonant verb such as '*yomu*'. Actually these rules have conditions not described here so that they are not overused.

for 4-grams, while its meaning is same as that of the reference sentence.

Basically BLEU copes with this problem of variation in writing styles by relying on the number of reference sentences available for each source sentence and by reflecting the total size of corpus. That is, if the corpus has multiple reference sentences translated by different translators, multiple writing styles will tend to be included, and if the corpus is very large, such inconsistencies of writing style are statistically not a problem.

In Japanese translation evaluation, however, this problem can not be resolved using such a quantitative solution because the influence of the differences in writing styles are too large. For example, whether or not the translation is given in the polite form depends on the translation system[2], so the evaluation score is strongly affected by the degree of matching of the writing styles between the translation system and the reference sentences.

To cancel out the differences in writing styles, we apply some paraphrasing rules to the reference sentences to generate new sentences with different writing styles. The generated sentences are added to the reference sentences, and therefore, n-grams in the candidate sentences can match the reference sentences regardless of their writing styles. Table 1 shows examples of paraphrasing rules.

These rules are applied to the reference sentences. If a reference sentence matches to a paraphrasing rule, the sentence is replicated and the replica is rewritten using the matched rule. For example, the Japanese sentence in Example 2 matches Rule 1 in Table 1 so the Japanese sentence in Example 3 is produced. In this case, the evaluation is done as if there are two reference sentences, therefore, a candidate sentence gets the same score regardless of its politeness.

To avoid applying the same rules repeatably, the rules are applied in a specific order. How to generate the rules is described in Section 4.1.

[2]Some translation systems allow us to specify such writing styles but some systems don't.

4 Experiments

4.1 Environments

To see how much the three extensions above contribute to the evaluation of translation, the correlation between the automatic evaluation and the human evaluation is calculated. We used a bilingual corpus which consists of 6,871 English sentences on a technical domain and their translations into Japanese.

100 sentences were randomly selected and translated by 5 machine translation systems S1-S5 and a human H1 who is a native Japanese speaker but does not have strong knowledge of the technical domain. These 6 translations were evaluated by five methods: B1 to B4 are Japanese versions of BLEU with the extension described in Section 3 and M1 is a manual evaluation.

B1: Morphological analysis is applied to translated Japanese sentences. Only the technique described in Section 3.1 is used.

B2: Functional words are distinguished by their POSs. This corresponds to the technique in Section 3.1 and 3.2.

B3: Paraphrasing rules are applied to the reference sentences as described in Section 3.3. Here the applied rules are limited to 51 rules which rewrite polite forms (*e.g.* 1 and 2 in Table 1).

B4: All 88 paraphrasing rules including other types (*e.g.* 3 and 4 in Table 1) are applied.

M1: Average score of the manual evaluation of all translations in the corpus. The sentences were scored using a 5-level evaluation: 1 (poor) to 5 (good). The evaluator was different from the translator of H1.

	B1	B2	B3	B4	M1
S1	0.115	0.114	0.132	0.135	2.38
S2	0.130	0.129	0.149	0.151	2.74
S3	0.134	0.132	0.148	0.152	2.77
S4	0.137	0.135	0.148	0.158	3.16
S5	0.183	0.177	0.179	0.180	3.38
H1	0.170	0.166	0.179	0.187	4.40
correl	0.797	0.803	0.865	0.931	(1.0)

Table 2: BLEU scores evaluated by each method. 'correl' means the correlation of each method with the manual evaluation (M1).

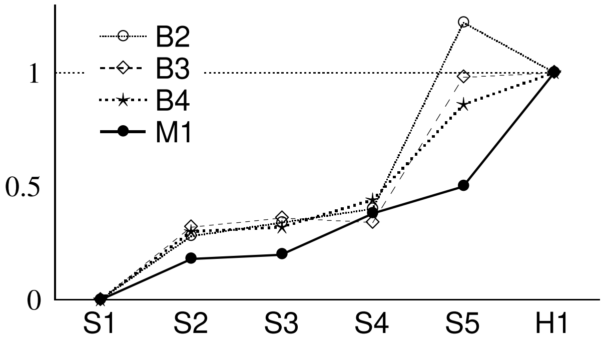

Figure 1: BLEU scores normalized as S1 is 0 and H1 is 1. B1 is omitted since it is close to B2.

The paraphrasing rules used in B3 and B4 were prepared manually by comparing the candidate sentences and the reference sentences in the reminder of the corpus which are not used for the evaluation. The application of the rules are unlikely to produce incorrect sentences, because the rules are adjusted by adding the applicable conditions, and the rules that may have side effects are not adopted. This was confirmed by applying the rules to 200 sentences in another corpus. A total of 189 out of the 200 sentences were paraphrased in at least a part, and all of the newly created sentences were grammatically correct and had the same meaning as the original sentences.

4.2 Experimental Results

Table 2 shows the result of evaluation using the five methods. Comparing the correlation with M1, B2 slightly outperformed B1, thus the POS information improves the evaluation. B3 was better than B2 in correlation by 0.06. This is because the scores by the B3 evaluation were much higher than the B2 evaluation except for S5, since only S5 tends to output sentences in polite form while the most of reference sentences are written in polite form. Further improvement was observed in B4, by applying other types of paraphrasing rules.

Figure 1 graphically illustrates the correlation between the BLEU evaluations and the human evaluations, by normalizing the results so that S1 is 0, H1 is 1, and the rest of scores are linearly interpolated. We can see that only B4 ranks all six systems in the same order as the manual evaluation.

5 Discussion

5.1 Lexical or Structural Paraphrasing Rules

The paraphrasing rules used here have no lexical rules that rewrite content words into other expressions as in Example 4.

Example 4

dokusho : *suru*	↔	*hon* : *wo* : *yo* : *mu*
'read'		'read a book'

The main reason why we don't use such rules is that this type of rules may produce incorrect sentences. For instance, (a) in Example 5 is rewritten into (b) by the rule in Example 4, but (b) is not correct.

Example 5

(a) *Kare ha watashi no hon wo yo mu.*
'He reads my book.'
(b)* *Kare ha watashi no dokusho suru.*
'He my reads.' (literally)

This error can be decreased if the paraphrasing rules have more strict conditions about surrounding words, however, using such lexical rules contradicts the original BLEU's strategy that the differences in expressions should be covered by the number of reference sentences. This strategy is reasonable because complicated rules tend to make the evaluation arbitrary, that is, the evaluation score strongly depends on the lexical rules. To verify that the lexical rules are unnecessary, we added 17,478 word-replacing rules to B4. The rules mainly replace Chinese characters or Kana characters with canonical

Paraphrasing rule	Δcorrel
da(aux) $\leftrightarrow$ *de* : *a* : *ru*	0.025
\$1(verb-v) : *ru* $\leftrightarrow$ \$1(verb-v) : *masu*	0.022
\$1(noun) : (dot) : \$2(noun) $\rightarrow$ \$1 : \$2	0.020

Table 3: The three best paraphrasing rules which contributed to the translation evaluation. The column 'Δcorrel' means the decrease of the correlation in the translation evaluation when the rule is removed. '(verb-v)' denotes a vowel verb.

ones. With the rules, the correlation with M1 was 0.886, which is much lower than B4.

This result implies the differences in content words do not affect the evaluations. More specifically, BLEU's misjudgments because of differences in content words occur with almost equal probability for each translation system. Thus it is enough to use the structural (*i.e.* non-lexical) paraphrasing rules which rewrite only functional words.

5.2 Evaluation of Each Paraphrasing Rule

The contribution of the paraphrasing was measured by the increase of reliability of the translation evaluation, as described in Section 4.2. In the same way, the effect of each single paraphrasing rule can be also evaluated quantitatively. Table 3 shows the three best paraphrasing rules which contributed to the translation evaluation. Here the contribution of a rule to the automatic evaluation is measured by the increase of correlation with the human evaluation when the rule is used.

6 Conclusion and Future Work

This paper has proposed an automatic translation evaluation method applicable to Japanese translation evaluation. The paraphrasing rules that cancel out the differences in writing styles contributed to improve the reliability of the automatic evaluation. The proposed evaluation method with paraphrasing rules achieved a high correlation of 0.93 with the human evaluation, while the correlation was 0.80 without the rules.

The experiments clarified how much the paraphrasing rules improved the evaluation by comparing the correlations. This means our system can evaluate not only the translation quality but also the paraphrasing rules under the assumption that the more properly the semantically similar sentences are judged as close sentences the more reliable the translation evaluation is. Therefore the translation evaluation gives us an objective evaluation method of the paraphrasing quality that has been difficult to evaluate.

This paper focuses on non-lexical paraphrasing since lexical paraphrasing rules make the translation evaluation inconsistent, but if an exhaustive and precise set of paraphrasing rules can be generated, it will be useful for translation evaluation, and its appropriateness should be shown by the reliability of the translation evaluation. In order to develop such desirable paraphrasing rules, the automatic acquisition of paraphrasing rules will be our next research direction.

Acknowledgments

I am grateful to Dr. Kishore Papineni for the instruction of BLEU. I would like to thank people in Yamato Research Laboratory for helping the evaluation.

References

Teruko Mitamura and Eric Nyberg. 2001. Automatic rewriting for controlled language translation. In *Proc. of NLPRS2001 Workshop on Automatic Paraphrasing*, pages 1–12.

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 HLT2002*, pages 124–127.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, pages 311–318.

Tetsuro Takahashi, Tomoya Iwakura, Ryu Iida, and Kentaro Inui. 2001. Kura: A revision-based lexico-structural paraphrasing engine. In *Proc. of NLPRS2001 Workshop on Automatic Paraphrasing*, pages 37–46.

Kentaro Torisawa. 2002. An unsupervised learning method for associative relationships between verb phrases. In *Proc. of COLING 2002*, pages 1009–1015.

Ingrid Zukerman and Bhavani Raskutti. 2002. Lexical query paraphrasing for document retrieval. In *Proc. of COLING 2002*, pages 1177–1183.