Automatic Rating of Machine Translatability

Kiyotaka Uchimoto[†] **Naoko Hayashida**[‡] **Toru Ishida**[‡] **Hitoshi Isahara**[†]

[†]National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
{uchimoto,isahara}@nict.go.jp

[‡]Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
hysd@kuis.kyoto-u.ac.jp,ishida@i.kyoto-u.ac.jp

Abstract

We describe a method for automatically rating the machine translatability of a sentence for various machine translation (MT) systems. The method requires that the MT system can bidirectionally translate sentences in both source and target languages. However, it does not require reference translations, as is usual for automatic MT evaluation. By applying this method to every component of a sentence in a given source language, we can automatically identify the machine-translatable and non-machine-translatable parts of a sentence for a particular MT system. We show that the parts of a sentence that are automatically identified as non-machine-translatable provide useful information for paraphrasing or revising the sentence in the source language, thus improving the quality of the final translation.

1 Introduction

Machine translation (MT) systems are becoming more widely used by ordinary people as well as by expert translators, with numerous web sites offering free translation services. In view of this situation, an international research project called the ICE (Intercultural Collaboration Experiment) project was launched to investigate the use of MT systems (Nomura et al., 2002) [1]. This joint research project is being undertaken by universities, and research institutes and societies in Asia. The goal of the project is to support intercultural and multilingual collaboration by using MT systems to aid human-to-human communication across international borders. As the first step in achieving the goal, multinational Asian teams conducted an experiment on open-source software development. In the experiment, each team member wrote a message in his/her first language and translated it into the other members' first languages using an MT system. Each member who received a message read it in his/her first language. During the experiment, however, they often found that translation errors resulted in incomprehensible messages or possible misunderstandings. They therefore had to exchange messages several times to fix the errors and understand what the writer meant. The sender of an incomprehensibly translated message had to paraphrase the original message to make it more machine translatable. The problem here is that the receiver may have difficulty in detecting the incomprehensible or misleading parts of a message and in letting the sender know which parts need paraphrasing because in many cases one error affects other parts of the translation and the whole phrase or sentence becomes incomprehensible. Therefore, the sender has to identify the part that needs paraphrasing through trial and error.

In this paper, we define machine translatability as a measure that indicates how well a given sentence can be translated by a particular MT system, and propose a method for automatically rating its machine translatability. The machine translatability of a given sentence is estimated as high when the quality of the MT result is good. Generally, reference translations are required to evaluate the quality of the MT result. However, our proposed method dose not use reference translations to rate the machine translatability of a sentence. Instead, it only requires the MT system to bidirectionally translate the sentence into both the source and target languages. We consider that the availability of an effective rating method will improve communication between people who speak different languages. In this study, the MT system was used like a black-box tool because our aim was to produce a better translation than the original one without modifying the MT system itself.

2 Confidence Measure (C-measure)

The C-measure is defined as the similarity between a source-language sentence and its *back translation*. A back translation is defined as the source-language sentence that is obtained by

[1] http://ice.kuis.kyoto-u.ac.jp/ice/

translating a sentence into the target language and then retranslating that sentence translated into the target language back into the original language. The calculation of similarity is described in the next section. The similarity is ideally rated high when the original sentence and the back translation of the sentence have the same meaning. In this paper, we assume that the higher the C-measure, the more stable and reliable the translation. Of course, this assumption is not always true and exceptions are discussed in Section 3.2.

We used a commercial MT system that translates Japanese into English and English into Japanese to obtain the back translations in our experiments.

2.1 Features of the C-measure

Here we assume that the higher the similarity, the higher the machine translatability of a sentence. The similarity is calculated based on BLEU (Papineni et al., 2002), which is often used to evaluate automatic MT. The C-measure is calculated using the following equation:

$$\mathrm{CM} \;=\; \frac{2 \times \mathrm{CM_{bleu}(B|S)} \times \mathrm{CM_{bleu}(S|B)}}{\mathrm{CM_{bleu}(B|S)} + \mathrm{CM_{bleu}(S|B)}}, \quad (1)$$

where S and B in $\mathrm{CM_{bleu}(B|S)}$ indicate the original sentence and its back translation, respectively. $\mathrm{CM_{bleu}(B|S)}$ is derived from the equation for calculating the BLEU score by substituting the original sentence and its back translation for the reference translation and translation, respectively. The equation is as follows:

$$\log(\mathrm{CM_{bleu}(B|S)}) \;=\; min\left(1 - \frac{s}{b}, 0\right) + \sum_{n=1}^{N} \frac{1}{N} \mathrm{log} p_n(B|S), \quad (2)$$

where s, b, and N indicate the number of words in the original sentence, the number of words in its back translation, and the maximum number of words in the considered word n-gram. $p_n(B|S)$ is represented as follows:

$$p_n(B|S) \;=\; \frac{\sum_{wn \in B} Count_{clip}(wn)}{\sum_{wn' \in B} Count(wn')}, \quad (3)$$

where $Count(wn')$ indicates the frequency of the word n-gram wn' in B. $Count_{clip}(wn)$ is represented as follows:

$$Count_{clip}(wn) \;=\; \min(Count(wn), Count(wn|S)), \quad (4)$$

where $Count(wn|S)$ represents the frequency of the word n-gram wn in S. There are some differences between our concept of similarity and that of BLEU with our system having the following additional features:

- Tree-based word n-grams

  Word order is relatively free in several languages such as Japanese and Korean. For example, several sets of word order are possible for the following Japanese dependency structure: "*Taro to Hanako wa tenisu wo shita*," which means "Taro and Hanako played tennis."

  - *Taro to Hanako wa tenisu wo shita*
  - *tenisu wo Taro to Hanako wa shita*

  The dependency structure of this sentence is shown in Figure 1. Each node represents a *bunsetsu*. *Bunsetsu*s are minimal linguistic units obtained by segmenting a sentence naturally in terms of semantics and phonetics; each one consists of one or more morphemes. The BLEU score for the above two sentences is not 1 because BLEU is based on word n-grams. However, the similarity of the sentences should be 1 because the two sentences have the same meaning. In our measure, therefore, word n-grams are extracted from dependency trees as shown in Figure 1. A word unit is defined as a morpheme and the definition of a morpheme follows that of JUMAN (Kurohashi and Nagao, 1999). All word dependencies within a *bunsetsu* are assumed to be between adjacent morphemes. The direction of all word dependencies between *bunsetsu*s is assumed to be from the rightmost word in a modifier *bunsetsu* to the leftmost word in the modified *bunsetsu*. For example, the word bi-grams extracted from Figure 1 are "*Taro to*", "*Hanako wa*", "*tenisu wo*", "*to Hanako*", "*wa shita*", "*wo shita*". Here, word 3-grams were used as word n-grams, based on the results of our preliminary experiments.

- Harmonic-mean

  BLEU is based on the word n-gram precision of an automatic translation. Therefore, if an automatic translation and reference translation are substituted for each other, the BLEU score differs from the original one. However, they must have the same similarity. Our measure, therefore, uses not only the original BLEU score,

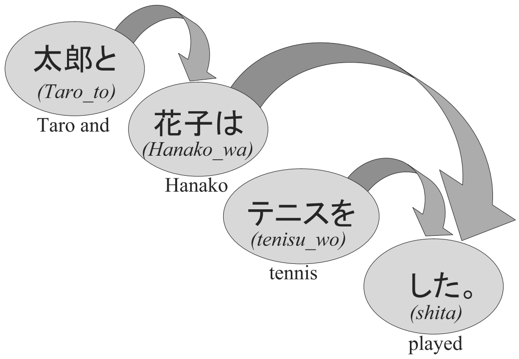

Figure 1: Example of Japanese dependency structure: "Taro and Hanako played tennis."

but also the BLEU score calculated when the automatic translation and a reference translation are substituted for each other. The latter BLEU score is based on the word n-gram recall of an automatic translation. Therefore, the F-measure, namely, the harmonic-mean of the precision and recall of both types of BLEU scores, as shown in equation (1), is used as our similarity measure.

- Generalization

  BLEU is based on surface words. Synonyms are therefore regarded as different words. However, when the differences between two sentences involve synonyms, the similarity should be 1. Therefore, the words are replaced with the appropriate word classes. When a word belongs to two or more classes, a quasi-optimal sets of word classes are found greedily in the sense the rate of agreement on word classes between the source-language sentence and its back translation is as high as possible. Word classes are defined based on a thesaurus *Bunrui goihyou* developed by the National Institute for Japanese Language (for Japanese Language (NIJL), 2004). The *Bunrui goihyou* has a tree structure and consists of seven layers. We used the upper fifth layers of the *Bunrui goihyou* as word classes. The leaves of the tree contain words, and each word has a figure indicating its category number. There are 101,070 words in the *Bunrui goihyou*. Words that belong in the conjunctive particle or numeral part-of-speech (POS) categories are generalized according to their POS categories. A series of numeral words are replaced with one numeral word, and all punctuation marks are ignored.

  Phrase-level and clause-level classes must also be considered and in future we plan to use state-of-the-art paraphrasing technologies.

In this paper, we used a BLEU-based measure as a C-measure. However, edit distance or other measures based on kernel methods could also be used as C-measures. In future work, we plan to investigate which C-measure is best.

2.2 Relationship to MT Evaluation Metrics

This section examines the relationships between the C-measure and the automatic MT evaluation metrics, BLEU and NIST (NIST, 2002), and between the C-measure and subjective human evaluation of MT results. The C-measure is calculated for a given sentence and the back translation produced by an MT system to rate the machine translatability of the Japanese input sentences. BLEU and NIST scores were calculated based on English translations and English reference translations to evaluate the English translations produced by the commercial MT system.

As a test set, we used an MT test set provided by NTT [2] (Ikehara et al., 1994). This set, which is used to evaluate Japanese-to-English MT systems, consists of 3,718 Japanese sentences with English translations. We used this set because it includes sentences that are difficult for a current leading commercial MT system to translate. The same problems found by the ICE project described in Section 1 were observed. Japanese sentences were used as input sentences. For each Japanese sentence, one English reference translation was used. The BLEU and NIST scores were calculated using the MT scoring software, mteval (version v11a) [3].

Figures 2 to 5 show the relationships between the C-measure and the BLEU score, between the C-measure and the NIST score, between the C-measure and the subjective human evaluation score for the fluency of the MT results, and between the C-measure and the subjective human evaluation score for the adequacy of the MT results, respectively. These figures were obtained by calculating the average scores for sets of original sentences and their English translations. The sets were constructed by classifying the original sentences into 10 groups according to their C-measures at intervals of 0.1.

[2] http://www.kecl.ntt.co.jp/icl/mtg/resources/index.php

[3] http://www.nist.gov/speech/tests/mt/resources/scoring.htm

The instructions for the subjective human evaluation followed that used for TIDES (TIDES, 2002). For subjective human evaluation, odd-numbered sentences were selected from the top of the test set. There were 950 sentences. Each line in each figure shows the result obtained when each feature described in Section 2.1 was used as the C-measure. For example, "bleu" shows that the original BLEU was used as a C-measure, and "tree-ngram+generalization" shows that BLEU with tree-based word n-grams and generalization features was also used as a C-measure.

When BLEU, with all the features as shown in Section 2.1, was used as the C-measure, the correlation coefficients between the C-measure and the BLEU and NIST scores, between the C-measure and the fluency, and between the C-measure and the adequacy in Figures 2 to 5 were 0.9128, 0.9133, 0.8112, and 0.7966, respectively, which are relatively high. This indicates that the C-measure can be used to select translations whose average quality is high. It also indicates that by collecting sentences with low C-measures, we can automatically find sets of sentences that are difficult for MT systems to translate without having to prepare reference translations. These sets could be used effectively to improve MT systems, avoiding the high cost of preparing reference translations. We believe that it is possible to avoid communication misunderstandings by relying on translations with a high C-measure and with only small differences between the source-language sentence and its back translation. This system could also reduce the cost of revising translations when automatic translation results require revision because the quality of the translations could be improved by revising sentences with low machine translatability.

In Figures 2 to 5, the best average correlation coefficient achieved for BLEU and NIST was obtained when BLEU with generalization and harmonic-mean features was used as the C-measure. The correlation coefficients achieved for BLEU and NIST were 0.9408 and 0.9346, which were both very high. The best average correlation coefficients achieved for subjective human evaluation was also obtained when BLEU with the generalization feature and harmonic-mean feature was used as the C-measure. The correlation coefficients achieved for fluency and adequacy were 0.9089 and 0.8770. This result suggests that a C-measure that correlates well with subjective human evaluation could be selected according to the correlation coefficients achieved for the automatic MT evaluation metrics. The correlation coefficients decreased when the tree-based word n-grams feature was used though this may have been due to errors in the analysis of the dependency structure.

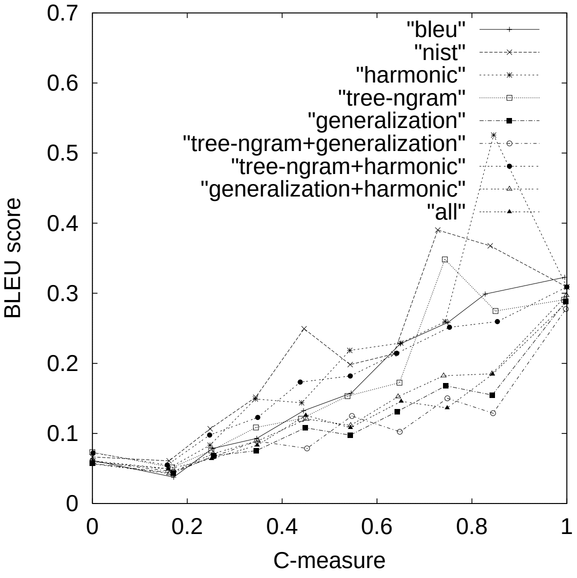

Figure 2: Relationship between C-measure and BLEU score

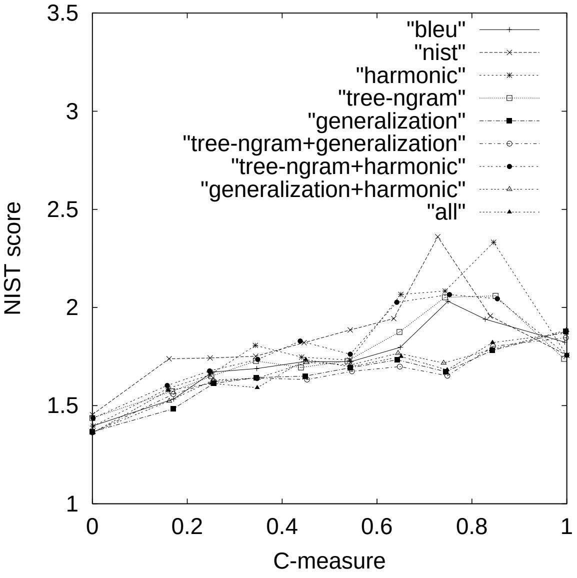

Figure 3: Relationship between C-measure and NIST score

In this section, the machine translatability of the sentences is calculated. However, there are still difficulties in identifying the non-machine-translatable parts of a sentence. In the next section, we describe a method for detecting the

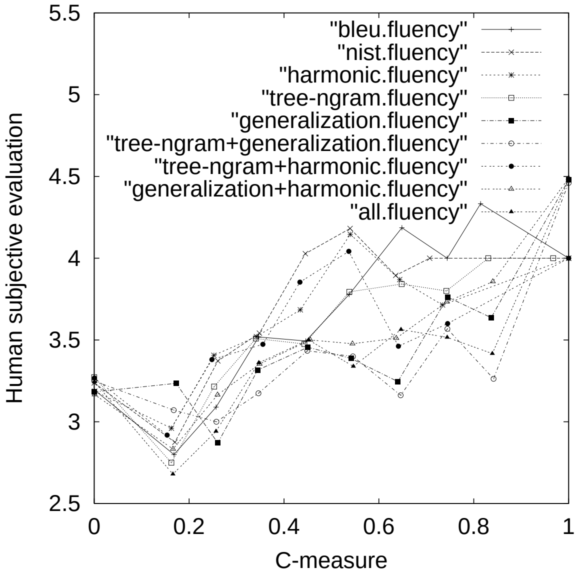

Figure 4: Relationship between C-measure and fluency

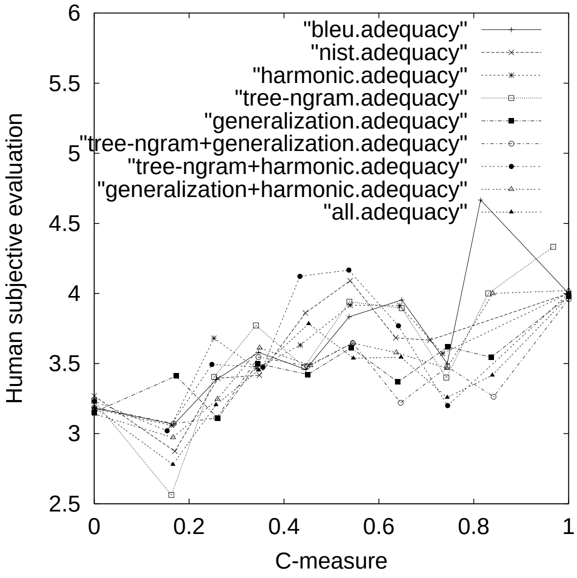

Figure 5: Relationship between C-measure and adequacy

non-machine-translatable parts of text excerpts.

3 Using Semi-Automatic Translation to Aid Translation

3.1 Detection of Non-Machine-Translatable Parts of a sentence

As shown in the previous section, the machine translatability of a given sentence can be ranked using a C-measure. Therefore, it may be possible to detect the machine-translatable parts of a sentence by calculating the C-measure for each part of the sentence. As sample parts, we used the back translation of each subtree for given sentences. That is, we calculated the C-measures for all subtrees in the given sentence. Here, let us assume that the sentences themselves belong to the subtree set SST. The dependency trees of a Japanese sentence can be derived using JUMAN (Kurohashi and Nagao, 1999) and KNP (Kurohashi, 1998). Subtrees were extracted from the dependency trees thus obtained.

The confidence score for a subtree $st_i (\in SST)$, $Scr(st_i)$ can be defined as follows:

$$Scr(st_i) \;=\; (\text{CM of } st_i) \times \frac{\#\text{ of } bunsetsus \text{ in } st_i}{\#\text{ of } bunsetsus \text{ in a given sentence}}. \quad (5)$$

Thus, the non-machine-translatable part is detected by finding the best subset of SST, ST_{best} as follows:

$$ST_{best} \;=\; \underset{ST}{\text{argmax}} \sum_{st_i \in ST} Scr(st_i), \quad (6)$$

where ST is a subset of SST and any *bunsetsu*s of subtrees in ST that do not overlap. That is, the original sentence can be generated by joining all the subtrees in ST. When several subtrees have the same confidence score, the longest one is preferred. The length is defined as the number of *bunsetsu*s in a subtree. When the confidence score of a given sentence is the highest of all the subtrees, that sentence is selected as ST_{best} by this equation.

A greedy algorithm is used to search for the optimal subset of S. Therefore, a quasi-optimal subset is sometimes selected rather than the optimal subset. This search algorithm will be improved in future work.

The best subset of subtrees with C-measures are presented to users. When there are non-machine-translatable parts, ST_{best} consists of subtrees with both high and low C-measures. There is a high possibility that subtrees with a low C-measure are non-machine-translatable parts. Possible non-machine-translatable parts are detected as follows:

1. When all the C-measures of the subtrees in the best subset are lower than a predetermined threshold, the subtree with the lowest C-measure is extracted from the best

subset and presented to users as a possible non-machine-translatable part. In this case, the rightmost part of the given sentence is often non-machine-translatable, or some information required for MT, such as the subject word of the sentence, may be missing. When several subtrees have the same C-measure, the longest one is preferred to the others.

2. When a subtree with a C-measure above the threshold is found in the best subset, the subtree is often machine-translatable, and the remaining subtrees with low C-measures are often non-machine-translatable. Then, all the subtrees with C-measures below the threshold are extracted from the best subset and presented to users as possible non-machine-translatable parts.

   We sometimes find a subtree from which possible non-machine-translatable parts have been extracted, but its super-subtree has a C-measure that is above the threshold. In this case, the difference between the subtree and its super-subtree indicates the non-machine-translatable part. Therefore, the super-subtree with the highest C-measure is presented to users as reference information. Here, the super-subtree of a subtree *st* means the subtree that includes *st*.

The example output is shown in Figure 6. “Partial translation” indicates the best subset of subtrees. “Check!” indicates non-machine-translatable parts. The threshold was set at 0.5 for the experiment. In this experiment, the parameters and thresholds were not tuned to the test set. Better results would be achieved if they were tuned and this could be done automatically.

In the example shown in Figure 6, the set of the subtree consisting of the *bunsetsu* “鉛筆は、” with a C-measure of 0 and the subtree consisting of the *bunsetsus* “２ＢかＨＢを使ってください。” with a C-measure of 0.77 were selected as the best subset. The second step in detecting non-machine-translatable parts, as mentioned above, is then applied since the C-measure of the second subtree in the best subtree is above the 0.5 threshold. The first subtree is then presented to users as a possible non-machine-translatable part.

```
#ORIGINAL: 鉛筆は、２ＢかＨＢを使ってください。
#----------------------------------------------------------------------
#   Original sentence             Back translation     Confidence score
#----------------------------------------------------------------------
#---Subtrees-->
２ＢかＨＢを使ってください。          ２ＢまたはＨＢを使ってください。    0.58
ＨＢを使ってください。               ＨＢを使ってください。            0.5
鉛筆は、２Ｂか使ってください。        ２Ｂまたは鉛筆を使います。         0.26
使ってください。                    使ってください。                  0.25
２Ｂか使ってください。               ２Ｂまたは使用。                  0.23
鉛筆は、２ＢかＨＢを使ってください。  鉛筆使用２ＢまたはＨＢ。           0.22
鉛筆は、ＨＢを使ってください。        鉛筆使用ＨＢ。                    0
鉛筆は、使ってください。             鉛筆を使ってください。             0
鉛筆は、                           鉛筆                            0
２Ｂか                             それは２Ｂですか？                 0
ＨＢを                             ＨＢの                           0
#<--Subtrees---
#----------------------------------------------------------------------
#   Original sentence            Back translation            C-measure
#----------------------------------------------------------------------
#---Partial translation-->
２ＢかＨＢを使ってください。          ２ＢまたはＨＢを使ってください。    0.77
［鉛筆は、                          鉛筆                            0   ］
（鉛筆は、２Ｂか使ってください。      ２Ｂまたは鉛筆を使います。         0.26）
#<--Partial translation---
#---Check!-->
［鉛筆は、                          鉛筆                            0   ］
#<--Check!---
EOD
```

Figure 6: Example of detection of non-machine-translatable part of a sentence.

3.2 Experimental Results and Discussion

We conducted experiments to examine whether detecting the non-machine-translatable parts of a sentence and the best subset of subtrees could help improve the machine translatability of the original input sentence. The first 100 sentences used in the experiment shown in Section 2.2 were selected and the best subset of subtrees and non-machine-translatable parts of the 100 sentences were presented to a human subject. For the C-measure, BLEU with generalization and harmonic-mean features, which achieved the best average correlation coefficients for both subjective human evaluation and the automatic MT evaluation metrics in Figures 2 to 5 was used. The subject revised the original Japanese sentences by referring to the information presented, as shown in Figure 6 (no information was presented on the target language). For example, if the human subject referred to the detected non-machine-translatable part “鉛筆は、”, which is indicated by “Check!” in Figure 6, and revised the original sentence “鉛筆は、2BかHBを使ってください。(As for a pencil, use 2B or HB.)” to “2BかHBの鉛筆を使ってください。(Use a 2B or HB pencil.)”, then we achieved an acceptable MT result, i.e., “Use the pencil of 2B or HB”, while the initial MT result was “The pencil use 2B or HB.” After the subject had revised the sentences, we found that the quality of the MT results improved as shown in Table 1. Forty-three sentences out of 100 were actually revised. Examples of original and revised sen-

Table 1: Evaluation of translations

	BLEU	NIST	Average grade	Acceptable translation
Before revision	0.1739	3.3162	2.73	54% (54/100)
After revision	0.2161	3.6674	3.52	75% (75/100)

Table 2: Examples of original and revised sentences and detected non-machine-translatable parts

Original sentences (non-machine-translatable parts are underlined)	Reference translations	Revised sentences
私は 最中を 食べた。 (MT: I ate time.)	I ate a monaka.	私はモナカを食べた。 (MT: I ate bean-jam-filled wafers.)
大抵の人が 帽子をかぶっていた。 (MT: The most person put on a hat.)	Most persons wore hats	ほとんどの人々が帽子をかぶっていた。 (MT: Most people put on a hat.)
彼は 飛んでいる 鳥を撃ち落とした。 (MT: He shot down the bird to be flying in.)	He shot down a bird in flight.	彼は飛行中の鳥を撃ち落とした。 (MT: He shot down the bird of the flying.)
船が暗礁に 乗り上げる。(MT: The ship reaches a deadlock.)	A ship runs aground.	船が座礁する。(MT: The ship strands.)
彼は 魚を 釣りに行った。 (MT: He went fishing in the fish.)	He went fishing.	彼は釣りに行った。 (MT: He went fishing.)
彼は手を 合わせた。(MT: He adjusted a hand.)	He placed his hands together.	彼は合掌した。 (MT: He joined one's palms together.)
両力士は胸を 合わせた。 (MT: Both sumo wrestlers adjusted a chest.)	The two sumo wrestlers came to grips.	両力士は組み合った。 (MT: Both sumo wrestlers grappled.)
彼は 仕事に 身を入れた。(MT: He attended to the work.)	He put his heart into his work.	彼は仕事に専念した。 (MT: He concentrated on the work.)

tences and non-machine-translatable parts are shown in Table 2. By assessing the MT results obtained before and after revision, we found that 29 of the revised sentences were translated into higher quality English sentences than the initial translations, and the quality of the MT results for the remaining 14 sentences did not decrease. For 18 (62%) of the 29 sentences, the detected non-machine-translatable parts were revised appropriately. Two sentences of the remaining 11 sentences were complemented with subject words. The remaining 9 sentences were revised because the machine translatability of the whole sentence was low and the human subject judged that the back translation included an incorrect part, although the subtree was not automatically detected as a non-machine-translatable part. These results indicate that detecting the non-machine-translatable parts of sentences and the best subsets of subtrees helps to improve the translatability of the original input sentence. The quality of the MT results was evaluated by a human subject using five grades: 1 (very poor) to 5 (very good). The translation was considered acceptable when the grade was 3 or better. We found that the average grade improved from 2.73 to 3.52 following revision, as shown in Table 1. For the 43 revised sentences, the average grade improved significantly from 1.63 to 3.47. Fifty-seven sentences out of 100 were not revised because the human subject decided that (A) no revision was needed for 29 of the 57 sentences, or (B) that it was difficult to revise the original sentences for the 28 remaining sentences, although the meaning of the back translation did not match that of the original sentence. The number of MT results that received grades of 2 or less was 2 for (A) and 9 for (B). After revision, the number of MT results receiving grades of 2 or less decreased from 35 to 14. Therefore, 46 translations with grades of 2 or less were originally unacceptable and should have been revised, and 21 (46%) of the 46 were rated as acceptable after the original sentences were revised. That is, the number of acceptable translations improved from 54 (54%) to 75 (75%) after revision.

Finally, we conducted an experiment to compare the MT results obtained by referring to the non-machine-translatable parts of a sentence and the best subset of subtrees with those obtained without referring to the additional information. One hundred sentences from number 201 to 300 of the MT test set, as shown in Section 2.2, were used in the experiment. The average grade of the MT results for the original sentences was 2.27. We found that the average grade improved to 3.06 when the additional information was used, and improved to 2.63 without any additional information. This result shows that the information provided by our system helped improve the machine translatability of the original input sentence.

In this paper, we used an MT system that is considered to produce one best translation. To detect non-translatable parts, however, two or more MT systems could be combined. A TM (Translation Memory) system could also be used.

We assumed that the higher the C-measure, the more reliable the translation. However, the back translation of a literal translation is often similar to the original sentence, even though the literal translation is incorrect. In this case, the sentence is estimated to be highly machine translatable. To avoid overestimating the machine translatability, we are planning to use language models in the target language. Note that in the experiments described in Section 2.2, the percentage of overestimated input sentences, i.e., the percentage of the sentences that received subjective human evaluation, fluency and adequacy scores of less than 3 but C-measures of over 0.5, was less than 4%, which is not significant.

There are other ways of improving the performance of MT systems, for example, by developing MT-system-dependent rules. However, these rules would be superseded if the system was improved. The proposed method is MT-system-independent, and thus will not be affected by improvements in MT systems. Rather, any improvements in MT systems will help improve the reliability of machine translatability ratings.

4 Conclusion

We proposed a method for automatically rating the machine translatability of a given text for a particular MT system. The method provides: (1) a confidence measure (C-measure) that estimates the machine translatability of a given input for an MT system, and (2) information to identify the machine-translatable and non-machine-translatable parts of an input sentence.

We found that the C-measure correlated well with subjective human evaluation of the results of machine translation as well as with the automatic MT evaluation metrics, BLEU and NIST. These results indicate that the C-measure could be used to automatically evaluate translation results, avoiding the need for reference translations.

Although we used a single MT system in one translation direction, we are planning to use multiple MT systems and TM to further improve the measure for rating machine translatability, and to produce higher quality translations. We are also planning to use the C-measure to automatically revise a given source text. In recent years, there has been intensive research on paraphrasing technology, and free software is available for paraphrasing sentences. If we used this technology to automatically select the optimal paraphrase in the source language for translation by a particular MT system, we could improve the quality of the resulting translation without changing the MT system itself. Such a system would also have potential applications in identifying complete texts or stretches of texts that might cause problems. We are planning to investigate these possibilities in future work.

Acknowledgments

The authors would like to thank Mr. Masatoshi Tsuchiya, Mr. Mitsuhiro Kida, Dr. Kazuhiro Takeuchi, and the members of the Ishida Lab. at Kyoto University for their assistance in software development, and Dr. Takehito Utsuro, Prof. Shoichi Yokoyama, and anonymous reviewers for their helpful comments.

References

The National Institute for Japanese Language (NIJL), editor. 2004. *Word List by Semantic Principles*. Dainippon-tosho. (in Japanese).

Satoru Ikehara, Satoshi Shirai, and Kentaro Ogura. 1994. Criteria for Evaluating the Linguistic Quality of Japanese to English Machine Translations. *Transactions of the JSAI*, 9(4), 7. (in Japanese).

Sadao Kurohashi and Makoto Nagao, 1999. *Japanese Morphological Analysis System JUMAN Version 3.61*. Department of Informatics, Kyoto University.

Sadao Kurohashi, 1998. *Japanese Dependency/Case Structure Analyzer KNP Version 2.0b6*. Department of Informatics, Kyoto University.

NIST. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Technical report, NIST.

Saeko Nomura, Toru Ishida, Kaname Funakoshi, Mika Yasuoka, and Naomi Yamashita. 2002. Intercultural Collaboration Experiment 2002 in Asia: Software Development Using Machine Translation. *Transactions of Information Processing Society of Japan*, 44(5):503–511. (in Japanese).

Kishore Papineni, Salim Roukos, Todd Ward, and Weiing. Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

TIDES. 2002. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations. http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess02.pdf.