

Automatic Detection and Semi-Automatic Revision of Non-Machine-Translatable Parts of a Sentence

Kiyotaka Uchimoto*, Naoko Hayashida†, Toru Ishida†, and 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 developed a method for automatically distinguishing the machine-translatable and non-machine-translatable parts of a given sentence for a particular machine translation (MT) system. They can be distinguished by calculating the similarity between a source-language sentence and its *back translation* for each part of the sentence. The parts with low similarities are highly likely to be non-machine-translatable parts. We showed that the parts of a sentence that are automatically distinguished as non-machine-translatable provide useful information for paraphrasing or revising the sentence in the source language to improve the quality of the translation by the MT system. We also developed a method of providing knowledge useful to effectively paraphrasing or revising the detected non-machine-translatable parts. Two types of knowledge were extracted from the EDR dictionary: one for transforming a lexical entry into an expression used in the definition and the other for conducting the reverse paraphrasing, which transforms an expression found in a definition into the lexical entry. We found that the information provided by the methods helped improve the machine translatability of the originally input sentences.

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 Intercultural Collaboration Experiment project was launched to investigate the use of MT systems (Nomura et al., 2002)¹. This research project is being undertaken by universities and research institutes and societies in Asia. The goal is to support intercultural and multilingual collaboration by using MT systems to aid communication across international borders. As the first step toward achieving the goal, multinational Asian teams experimented 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 writers meant. 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. Even if the part is identified, he/she may have difficulty in paraphrasing it effectively. We developed methods for automatically detecting the non-machine-translatable parts of a given sentence for a particu-

lar MT system and for providing knowledge useful to effectively paraphrasing or revising the detected non-machine-translatable parts.

2. Machine Translatability

We used the definition of machine translatability of Uchimoto et al. (Uchimoto et al., 2005). Machine translatability is a measure that indicates how well a given sentence can be translated by a particular MT system, and we call the measure the “confidence measure”, or 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 translating a sentence into the target language and then retranslating that sentence into the original language. We calculated similarities using a method based on BLEU (Papineni et al., 2002). We assumed that the higher the C-measure, the more stable and reliable the translation. The C-measure is calculated using the following equation:

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

where S and B in $CM_{bleu}(B|S)$ indicate the original sentence and its back translation. The term $CM_{bleu}(B|S)$ 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. The equation is as follows:

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

¹<http://ice.kuis.kyoto-u.ac.jp/ice/>

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. The term $p_n(B|S)$ is calculated as follows:

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

where $\text{Count}(wn')$ indicates the frequency of the word n-gram wn' in B . The term $\text{Count}_{clip}(wn)$ is calculated as follows:

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

where $\text{Count}(wn|S)$ represents the frequency of the word n-gram wn in S . Our concept of similarity differs from that of BLEU, with our system having the following additional features:

- Tree-based word n-grams

In our measure, word n-grams are extracted from dependency trees. All word dependencies within a *bunsetsu* are assumed to be between adjacent words. The *bunsetsus* are minimal linguistic units obtained by segmenting a sentence naturally in terms of semantics and phonetics, and each of them consists of one or more words. The direction of all word dependencies between *bunsetsus* is assumed to be from the rightmost word in a modifier *bunsetsu* to the leftmost word in the modified *bunsetsu*. Here, word 3-grams were used as word n-grams, based on the results of our preliminary experiments.

- Harmonic mean

Our measure uses not only the original BLEU score, 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

The words are replaced with their word classes. When a word belongs to two or more classes, quasi-optimal sets of word classes are found greedily in the sense that 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 layer as word classes. The leaves of the

tree contain words, and each word has a figure indicating its category number. The *Bunrui goihyou* contains 101,070 words. 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 is replaced with one numeral word, and all punctuation marks are ignored.

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

3. Automatic Detection of Non-Machine-Translatable Parts of a Sentence

3.1. A Method for Detecting Non-Machine-Translatable Parts

Uchimoto et al. reported that the machine translatability of a given sentence can be ranked using a C-measure (Uchimoto et al., 2005). Therefore, the machine-translatable parts of a sentence may be detectable by calculating the C-measure for each part of the sentence. As sample parts, we used the back translation of each subtree for a given sentence. 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.

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

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

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

$$ST_{best} = \operatorname{argmax}_{ST} \sum_{st_i \in ST} Scr(st_i), \quad (6)$$

where ST is a subset of SST and any *bunsetsus* 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 *bunsetsus* 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} using this equation. A greedy algorithm is used to search for the optimal subset of S .

The best subset of subtrees with C-measures is presented to users. When parts cannot be machine translated, ST_{best} consists of subtrees with both high and low C-measures. Subtrees with low C-measures are highly likely to be non-machine-translatable parts. Then, possible non-machine-translatable parts are detected using the following steps:

```

#ORIGINAL: 彼は魚を釣りに行った。
#-----
# Original sentence      Back translation      Confidence score
#-----
#---Subtrees-->
彼は釣りに行った。      彼は釣りに行った。      0.75
彼は行った。            彼は行った。            0.5
彼は魚を釣りに行った。  魚を釣るために、彼は行った。 0.40
魚を釣りに              魚を釣るには。          0.23
魚を釣りに行った。      魚を釣るために、私は行った。 0.19
釣りに行った。          私は釣りに行った。      0.19
彼は                    彼?                      0
魚を                    釣りなさい。            0
釣りに                  釣りにおいて。          0
行った。                私は行った。            0
#<---Subtrees---
#-----
# Original sentence      Back translation      C-measure
#-----
#---Partial translation-->
彼は釣りに行った。      彼は釣りに行った。      1
[ 魚を                  釣りなさい。            0 ]
(彼は釣りに行った。    彼は釣りに行った。      0.75)
#<---Partial translation---
#---Check!-->
[ 魚を                  釣りなさい。            0 ]
#<---Check!---
EOD

```

Figure 1: Example of detection of non-machine-translatable part of sentence.

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 right-most part of the given sentence is often non-machine-translatable, or some information required for MT such as the subject word of the sentence is often 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 over 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 over 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*.

Example output is shown in Figure 1. “Partial translation” indicates the best subset of subtrees. “Check!” indicates non-machine-translatable parts. The threshold was set at 0.5 in the experiment. In the example shown in Figure 1, the set of the subtree consisting of the *bunsetsu* “魚を” with a C-measure of 0 and the subtree consisting of the *bunsetsu* “彼は釣りに行った。” with a C-measure of 1 was selected as the best subset. The first subtree, the *bunsetsu* “魚を”, would then be presented to users as a possible non-machine-translatable part.

3.2. Experimental Results and Discussion

We experimented 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 originally input sentences. As a test set, we used an MT test set provided by NTT² (Ikehara et al., 1994). This set consists of 3,718 Japanese sentences with English translations. Japanese sentences were used as input. The first 100 sentences of the test set were selected, and the best subsets of the subtrees and non-machine-translatable parts of the 100 sentences were presented to a human subject. For the C-measure, we used 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 (Uchimoto et

²<http://www.kecl.ntt.co.jp/icl/mtg/resources/index.php>

Original sentence (non-machine-translatable parts are underlined)	Reference translation	Revised sentence
私は <u>最中</u> 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.)
<u>大抵</u> 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 <u>飛んでいる</u> 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 <u>暗礁</u> 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 <u>魚</u> 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 <u>手</u> 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.)
<u>両力士</u> 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 <u>仕事</u> に <u>身</u> 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.)

Table 1: Examples of original and revised sentences and detected non-machine-translatable parts.

Human subject	Original (%)	Without information (%)	With information (%)
1	37	49	64
2	37	53	46
3	37	41	44
4	37	46	47

Table 2: The percentage of acceptable translations.

al., 2005). The subject revised the original Japanese sentences by referring to the information presented, as shown in Figure 1 (no information was presented on the target language). For example, when the human subject referred to the detected non-machine-translatable part “魚を”, which is indicated by “Check!” in Figure 1 and revised the original sentence “彼は魚を釣りに行った。” to “彼は釣りに行った。”, then we achieved an acceptable MT result, i.e., “He went fishing”, while the initial MT result was “He went fishing in the fish.” After the subject had revised the sentences, we found that the quality of the MT results improved. Examples of original and revised sentences and non-machine-translatable parts are shown in Table 1. 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 number of acceptable translations improved from 54 (54%) to 75 (75%) after revision.

Next, we experimented to compare the MT results obtained by referring to the non-machine-translatable parts of a sentence and the best subsets of subtrees with those obtained without referring to the additional information. The 100 sentences from number 201 to 300 of the MT test set were used in the experiment. The original sentences were revised by four human subjects. The results are shown in Table 2. We found that the number of acceptable translations improved from the initial 37 when the original sentences were revised with the additional information. We also found that the number of acceptable translations us-

ing the additional information was higher than that without using additional information, except in the case of one subject. This shows that the information provided by our system generally helped improve the machine translatability of the original input sentences.

4. Semi-Automatic Revision of Non-Machine-Translatable Parts of a Sentence

4.1. A Method for Semi-Automatic Revision of Non-Machine-Translatable Parts

In the experiments described in the previous section, the human subjects had to revise the original sentences without any knowledge of paraphrase candidates except the back translation. The subjects thus had no idea in many cases how to effectively revise the original sentences. If paraphrase candidates and the automatically detected non-machine-translatable parts could be provided, they would help the human subjects revise the sentences. This section describes a method for extracting knowledge to help us effectively paraphrase the non-machine-translatable parts. Investigating the paraphrase strategy of the human subjects in the experiments described in Section 3.2., we found that the strategy consisted of the following three actions:

1. Paraphrase a word in a *bunsetsu* as different one.
2. Paraphrase one or more *bunsetsus* as one *bunsetsu*.
3. Complement omitted case elements in each predicate.

Definition	Lexical Entry	Similarity
魚を釣ること (to fish)	魚釣りする (to fish)	0.62
魚を釣ること (to fish)	魚釣する (to fish)	0.62
魚を釣ること (to fish)	釣魚 (fishing)	0.62
魚を釣る人 (fisher)	つり手 (fisher)	0.62
魚を釣る人 (fisher)	釣り手 (fisher)	0.62
魚を釣る人 (fisher)	釣手 (fisher)	0.62
魚を釣ってとらえる (to catch fish by fishing)	釣る (to fish)	0.62
...

Table 3: Examples of candidates extracted from the EDR dictionary for “魚を釣る” (to fish).

We focused on supporting the first and second actions because they were major actions in our experiments. The third action will be supported in future work. We also found that less ambiguous expressions could be correctly translated more reliably than plain but ambiguous expressions. This is because the accuracy of text analysis improves when a given sentence consists of unambiguous words or has simple syntactic structure, and the improvement in text analysis contributes to improving the quality of MT results.

We used a dictionary to extract knowledge for paraphrasing. In conventional methods, also, dictionaries have been used for paraphrasing. For example, a method for paraphrasing verbs to plain expressions was proposed (Kaji et al., 2002). In this method, a verb found as a lexical entry of a dictionary is transformed into an expression found in the definition of the verb. The knowledge for transforming a lexical entry into an expression used in the definition is useful for paraphrasing; we call this *lex-to-def knowledge*. However, this kind of knowledge often generates plain but ambiguous expressions. To reduce the ambiguity, therefore, we also extract knowledge for conducting the reverse paraphrasing, which transforms an expression found in a definition into the lexical entry; we call this *def-to-lex knowledge*. We also extract notational variants and synonyms because they are often less ambiguous than the original word for a particular MT system. They are extracted by finding lexical entries that have the same definitions as the original word. We used the EDR dictionary (NIC, 2003), which has approximately 410,000 lexical entries.

Knowledge extraction and semi-automatic revision were conducted using the following steps:

1. Extraction of notational variants and synonyms and generation of paraphrase candidates

Content words are extracted from the automatically detected non-machine-translatable parts, and definitions whose lexical entries are the same as the content words are extracted from the EDR dictionary. Then, the lexical entries whose definitions are the same as the extracted definitions are extracted as candidates of notational variants and synonyms. For example, the definition of the lexical entry “釣魚” (fishing) is “魚を釣ること” (to fish)”, and two other lexical entries “魚釣りする” (to fish) and “魚釣する” (to fish) have the same definition in the EDR dictionary. Paraphrase candidates are generated by transforming the original content words into the notational variants and

synonyms.

2. Extraction of lex-to-def and def-to-lex knowledge and generation of paraphrase candidates

Content words are extracted from the automatically detected non-machine-translatable parts, and definitions and lexical entries including each content word are extracted from the EDR dictionary. Next, a *bunsetsu* in the non-machine-translatable parts and dependencies consisting of a *bunsetsu* in the non-machine-translatable parts and its modifiee or modifier are extracted from the original sentence. For each pair of the extracted strings, one extracted from the EDR dictionary and the other from the original sentence, the similarity is calculated. Then, all pairs are sorted according to the similarity, and the top 100 pairs are used as knowledge for paraphrasing. If several pairs have the same similarity, the pairs that include more content words from the original sentence are preferred. The similarity of each pair is calculated using the C-measure. Paraphrase candidates are generated using the pairs with high similarity in two ways: one, transforming the original strings into the definition whose lexical entry is paired with the original strings, and the other, transforming the original strings into the lexical entry whose definition is paired with the original strings.

For example, the detected non-machine-translatable part of the fifth example in Table 1 is “魚” (fish) and “を” (case-marker), and its modifiee is “釣る” (to fish). For the dependent expression, “魚を釣る” (to fish), the candidates shown in Table 3 can be extracted. In this table, “similarity” indicates the similarity between the expression “魚を釣る” and each definition in the EDR dictionary.

3. Ranking paraphrase candidates

For each paraphrase candidate generated using the above steps, a C-measure is calculated using the back translation. Then, the paraphrase candidates are ranked according to the C-measures.

By checking paraphrase candidates with their C-measures and back translations, an appropriate paraphrase is selected manually. In the above steps, paraphrase candidates are also manually generated. We plan to automate the generation of paraphrase candidates in the future.

Original sentence (non-machine-translatable parts are underlined)	Reference translation	Revised sentence
家には手が <u>ない</u> 。 (MT: There are no hands in a house.)	The house is short handed.	家は人手不足だ。 (MT: A house is shortage of labor.)
私は猫の手も <u>かりたい</u> 。 (MT: I'd like to be also aided by a cat.)	I am so busy that any help would be appreciated.	私は非常に忙しくて誰にでも応援してもらいたい。 (MT: I'm very busy and want everyone to support.)
彼は手のひらを <u>かえした</u> 。 (MT: He has returned a palm.)	He completely changed his attitude.	彼は態度をがらりと変えた。 (MT: He changed the attitude entirely.)
君は手が空いていたら、 <u>手伝ってほしい</u> 。 (MT: If you're free, you want you to help me.)	If you are free, we need your help.	君は手が空いていたら、 <u>手伝ってもらいたい</u> 。 (MT: If you're free, could you help me?)
彼は耳が <u>聞こえない</u> 。 (MT: He doesn't hear an ear.)	He is hearing impaired.	彼は聾者だ。 (MT: He's a person with hearing impairments.)
彼は蜂の巣を <u>つついた</u> 。 (MT: He picked a bee hive.)	He poked a honeycomb.	彼は蜂の巣を <u>こづいた</u> 。 (MT: He poked a bee hive.)

Table 4: Examples of original and revised sentences and detected non-machine-translatable parts.

4.2. Experimental Results and Discussion

We experimented to show that the extracted knowledge is useful for revising the detected non-machine-translatable parts. The 100 sentences from number 301 to 400 of the MT test set as described in Section 3.2. were used in the experiment. Examples of original and revised sentences and non-machine-translatable parts are shown in Table 4. We found that the numbers of acceptable translations improved from 68 (68%) to 76 (76%) using the paraphrase candidates, although the numbers of acceptable translations improved to 72 (72%) using only the information described in Section 3.. This shows that the information provided by semi-automatic paraphrasing helped improve the machine translatability of the originally input sentences.

5. Conclusion

We developed a method for automatically detecting the non-machine-translatable parts of a given text and a method for providing useful knowledge to effectively paraphrase or revise the detected non-machine-translatable parts. We found that the information provided by the methods helped improve the machine translatability of the originally input sentences.

Although we used a single MT system in one translation direction, we are planning to use multiple MT systems and translation memories to find the best measure for rating machine translatability and make the best translation. In recent years, research on paraphrasing technology has been intensive, and free software is now available for paraphrasing sentences. Therefore, we are planning to use this technology and our method to automatically generate paraphrase candidates and select the optimal paraphrase in the source language for translation by a particular MT system.

6. References

- The National Institute for Japanese Language (NIJL), editor. 2004. *Word List by Semantic Principles*. Dainippon-tohsho. (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):569–579. (in Japanese).
- Nobuhiro Kaji, Daisuke Kawahara, Sadao Kurohashi, and Satoshi Sato. 2002. Verb Paraphrase based on Case Frame Alignment. In *Proceedings of the 40th ACL*, pages 215–222.
- 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.
- NICT (National Institute of Information and Communications Technology), 2003. *EDR Electronic Dictionary Technical Guide*.
- 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 Wei jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th ACL*, pages 311–318.
- Kiyotaka Uchimoto, Naoko Hayashida, Toru Ishida, and Hitoshi Isahara. 2005. Automatic Rating of Machine Translatability. In *Proceedings of the MT Summit X*, pages 235–242.