Abstract

The aim of this study is to improve the quality of machine-translated Japanese from an English source by optimizing the source content using a machine translation (MT) engine. We measured the improvement using the existing metrics, *bilingual evaluation understudy* (BLEU), and *translation edit rate* (TER) by comparing the translation from the improved source with an existing one.

We utilized the concept of *context-free grammar* (CFG) hypothesized by Noam Chomsky, an American linguist, for the translation improvement methodology: human linguistic ability is innate and can be regarded as an organ; an innate *language acquisition device* (LAD); and accordingly, every child possesses the knowledge of *universal grammar* (UG), which phrase structure rules can represent.

This theory can be applied to translation quality improvement, as the neural MT structure is similar to that of the human nerve system; however, it does not completely simulate the recognition system of the human brain. In contrast, CFG is an abstract model that simulates the human cognitive model. Therefore, it can be utilized for quality improvement in combination with a neural MT engine.

In this study, we assumed that the tree height of the CFG syntax of the source is highly correlated with the accuracy of the target translation output because of the nature of the CFG, which is considered to be an embodiment of the fundamental linguistic module representing the human mind and brain. That is, the sentence of relatively low tree height produces the accurate translation since the human cognition system can accept grammar with a simple structure.

Based on the assumption, if the pre-edited source follows the CFG phrase structure rules that possibly improve the target Japanese translation, we can then provide generic and concrete guidelines for the source content developers.

Prolog programming was used to analyze the tree height of a CFG syntax tree. Prolog is a well-known logic programming language belonging to first-order predicate logic and has a high affinity for linguistic structure analysis. For these reasons, Prolog has been used to develop artificial intelligence (AI). A Prolog program is a collection of

Implication and Unit clauses. Unlike many programming languages that handle propositional logic leading to the truth and false values using Boolean operations, such as “AND,” “OR,” and “NOT,” Prolog handles predicates that represent the state and nature of the object; for example, “A is B.” That is, Prolog holds not only Boolean values but also objects and predicates. Prolog also uses first-order predicate logic, which is a commonly applied mathematical model. For the reasons as mentioned above, it is called a logic language or logic programming language.

In this study, to script the sentence in the Chomsky normal form in Prolog, we used a definite clause grammar, for example, “`sentence :- noun_phrase, verb_phrase`” which means that a sentence consists of a noun phrase and a verb phrase. The sample code was also used in the structural analysis to determine the tree height in accordance with CFG.

Two translation quality metrics were used for the evaluation: BLEU and TER. BLEU is used as a de facto standard for the automatic calculation of translation quality. The advantage of using BLEU is to enable quality evaluation easily and immediately so that a comparison between systems can be made at a low cost without complex or manual calculations. BLEU also generates a *brevity penalty* (BP), a penalty given for extremely short sentences, resulting in a high score. The BLEU score is highly correlated with manual human evaluations. TER is a metric that measures how much the post-editor edited the machine-translated sentence to create an improved translation. The evaluation is performed by comparing a reference sentence with a machine-translated sentence. A higher score is better for BLEU, whereas a lower score is better for TER. Both metrics assume that the more similar the machine-translated sentence is to the reference sentence, the better the quality.

We collected 50 source English samples and reference translations (Reference) from the existing translation memory to conduct our experiment. The improved English source was created by polishing the original English source. Both the original and improved English sentences were input in Amazon Translate to obtain the respective Japanese translated results. Reference translation (Reference) was manually created from the original source English samples. Three types of Japanese translation results were compared with Reference: Baseline, Improved, and Postedited. Baseline is the raw result from the original source using the MT engine, Improved is from the improved English

source, and Postedited is the one polished from Improved.

Next, we applied the improved English sources to CFG using Prolog, analyzed the tree height level (level), and categorized them. As a result, level 3: eight samples; level 4: 14 samples; level 5: 19 samples; level 6: five samples; level 7: two samples; level 8: two samples were obtained. The most frequently occurring level was level 5; the next most frequent was level 4.

Analysis was performed on three groups: levels 3 and 4, levels 3 to 5, and all levels after tokenization using Janome, a library of Python and a morpheme analysis engine, and then Pickle, a Python library for saving the data. The intention behind comparing levels 3 and 4 and levels 3 to 5 is to verify that sentences with higher tree height create an adverse result in translation quality and produce low-quality translations. This fact assumes that simple and clear source English sentences create a more accurate MT.

As a result, when comparing Improved with Baseline using level 3 and 4 groups, the BLEU score of Improved was increased by 5.8. The Improved TER score is 49.1, which is better than 65.5 of Baseline by 16.4, the most significant improvement among the three results, meaning that it produces an excellent Postedited score with relatively lesser effort. In the case of the level 3 to 5 group, the BLEU score gap between Baseline and Improved was 2.6. The TER difference between Baseline and Improved was 10.7, which is the second-best rate among the three results. The results show that a lower CFG tree height significantly produces better results with higher BLEU and lower TER scores.

From the results of this study, it can be concluded that the source tree height based on CFG is generally correlated with translation improvements, as shown in the BLEU and TER scores. This means that source sentences with simple grammar are easily and accurately translated, which could result from the common language acquisition system of humans proposed by Chomsky.

Given these facts, source content creation based on CFG is an effective method for translation with an MT engine. As demonstrated earlier, the application of this method can reduce the manual workload and the post-editing cost. This is beneficial from at least two perspectives: First, we do not have to pay a large sum of money to the language vendor. Second, the MT result can be improved without re-training the MT engine, which

does not require specific technical expertise and consumes a considerable amount of computer resources. This achievement allows people not well trained in artificial intelligence to improve the MT results.

We believe that this study will promote and assist in developing effective new technologies in the future.