Abstract

Machine translation is the task of automatically translating a text from one natural language into another. Statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora (Philipp Koehn, 2010). Many translation models of statistical machine translation such as word-based, phrase-based, syntax-based, a combination of phrase-based and syntax-based translation, and hierarchical phrase-based translation are proposed. Phrase-based and hierarchical-phrase-based model (tree-based model) have become the majority of research in recent years, however they are not powerful enough to legal translation. Legal translation is the task of how to translate texts within the field of law. Translating legal texts automatically is one of the difficult tasks because legal translation requires exact precision, authenticity and a deep understanding of law systems. The problem of translation in the legal domain is that legal texts have some specific characteristics that make them different from other daily-use documents as follows:

- Because of the meticulous nature of the composition (by experts), sentences in legal texts are usually long and complicated.
- In several language pairs such as English-Japanese the target phrase order differs significantly from the source phrase order, selecting appropriate synchronous context-free grammars translation rule (SCFG) to improve phrase-reordering is especially hard in the hierarchical phrase-based model
- The terms (name phrases) for legal texts are difficult to translate as well as to understand.

Therefore, it is necessary to find ways to take advantage to improve legal translation. To deal with three problems mentioned above, we propose a new method for translating a legal sentence by dividing it based on the logical structure of a legal sentence, using rule selection to improve phrase-reordering for the tree-based machine translation, and propose sentence paraphrasing and named entity to increase translation.

A legal sentence represents a requisite and its effectuation (Tanaka et al. 1993). If each part of the legal sentence is shown separately, the readability will increase especially for a long

sentence as seen in administrative laws. Such parts are recognized automatically by dividing a legal sentence according to the requisite-effectuation structure as described in this thesis. Furthermore, each fragment obtained by the dividing is shorter than the original sentence and the translation quality is expected to be improved. For the first problem mentioned above, we propose dividing and translating legal text basing on the logical structure of a legal sentence. The existing methods for dividing a sentence are mainly based on clause splitting and not be based on the requisite-effectuation structure. We recognize the logical structure of a legal sentence using statistical learning model with linguistic information. Then we segment a legal sentence into parts of its structure and translate them with statistical machine translation models. In this study, we applied the phrased-based and the tree-based models separately and evaluated them with baseline models.

Rule selection is important to tree-based statistical machine translation systems. This is because a rule contains not only terminals (words or phrases), but also nonterminals and structural information. During decoding, when a rule is selected and applied to a source text, both lexical translations (for terminals) and reorderings (for nonterminals) are determined. Therefore, rule selection affects both lexical translation and phrase reorderings. For the second problem, we propose a maximum entropy-based rule selection model for the tree-based model, the maximum entropy-based rule selection model combines local contextual information around rules and information of sub-trees covered by variables in rules.

For the last problem, we propose sentence paraphrasing and named entity approaches. We apply a monolingual sentence paraphrasing method for augmenting the training data for statistical machine translation systems by creating it from data that is already available. We generate named-entity recognition (NER) training data automatically from a bilingual parallel corpus, employ an existing high-performance English NER system to recognized named entities at the English side, and then project the labels to the Japanese side according to the word alignment. We split the long sentence into several block areas that could be translates independently.

We integrate dividing a legal sentence based on its logical structure into the first step of rule selection as well as sentence paraphrasing and named entity. With this method, our experiments on legal translation show that the method achieves better translations.

Keywords: *phrase-based machine translation*; *tree-based machine translation*; *logical structure of a legal sentence*; *CRFs*; *Maximum Entropy Model*, *rule selection*; *linguistic and contextual information*; *sentence paraphrasing*, *NER*