氏　　名	BUI THANH HUNG			
学位の種類	博士(情報科学)			
学位記番号	博情第 283 号			
学位授与年月日	平成 25 年 9 月 24 日			
論文題目	A study on statistical machine translation of legal sentences(法令文の統計的機械翻訳に関する研究)			
論文審査委員	主査	島津　明	北陸先端科学技術大学院大学	教授
		飯田　弘之	同	教授
		東条　敏	同	教授
		Ho Tu Bao	同	教授
		LE Anh Cuong	ベトナム国家大学	准教授

論文の内容の要旨

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.

論文審査の結果の要旨

本論文は、法律条項の機械翻訳について、新技術の提案とその効果を示している。本論文の機械翻訳手法は、近年盛んに研究されている統計的機械翻訳である。法律条項の翻訳には以下の問題がある。先ず、所得税法など法律条項の文は、一般に大変長く、正確に解析することが難しい。次に、日本語と英語など、言語構造が異なる言語は句の順序が異なり、翻訳システムのデコーダが適切な変換規則を選択するのは難しい。また、法律用語の翻訳が難しい。このような問題のそれぞれについて、本論文は、入力文の扱い、デコーダの変換規則の選択、統計的機械翻訳のための言語データの生成などに着目し、新提案を行い、その効果を実証的に示した。

入力文の扱いに係る提案は、原文を分割して、処理を正確にして翻訳の質を向上させようというものである。従来、長文を節に分割して各節を翻訳する方法が研究されている。本論文は、そのような従来研究に対して、法律条項の性質を考慮し、条項の論理構造に基づいて条項の文を分割して統計的機械翻訳により翻訳する方法を提案し、日英、英日翻訳実験により有効性を示している。実験対象は借地借款法など 4 法律の条項文 516 である。BLEU などの 3 つの尺度で、ベースラインの節分割の翻訳より良い結果となっている。日本人 10 名、及び英語話者 8 名による質の主観的評価も行い、良い評価となっている。語の翻訳誤りも減っている。法律条項の要件や効果の可読性も良くなっている。

デコーダの変換規則選択に係る提案は、適用可能な同期文脈自由規則が複数あるとき、言語素性や文脈素性を利用して、適切な規則を適用しようとするものである。提案法により、句の順序変換や訳語選択が改善される。上記法律の日英翻訳の BLEU 尺度は、本論文の提案を適用しない木構造の統計機械翻訳と比べ、良い結果となっている。

統計的機械翻訳のための言語データの生成に係る提案は、統計的機械翻訳の訓練文をパラフレーズして訓練データに加えること、及びパラレルコーパスを介して原言語の固有名から目的言語の固有名を獲得することにより、統計的機械翻訳システムの性能を高めようとするものである。これらにより、上記法律の英日翻訳の BLEU 尺度は良い結果となっている。

以上、本論文は、統計的機械翻訳について新技術と効果を示したものであり、学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。