氏　　名	星野　翔
学位(専攻分野)	博士 (情報学)
学 位 記 番 号	総研大甲第 2083 号
学位授与の日付	平成 31年 3 月22日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Syntax-based Preordering for Statistical Machine Translation
論 文 審 査 委 員	主　査　教授　相澤　彰子 教授　神門　典子 教授　佐藤　健 准教授　小町　守　首都大学東京システムデザイン学部 教授　宮尾　祐介　東京大学大学院情報理工学系研究科

（様式３）

博士論文の要旨

氏名　　星野　翔

論文題目　Syntax-based Preordering for Statistical Machine Translation

This thesis focuses on a major yet unsolved machine translation difficulty called a word ordering problem. Since every language has its own word order, machine translation systems have to translate one's word order into another, in addition to translation of words. However, practical machine translation systems do not translate most of word orders due to computational complexity of the word ordering problem. This makes machine translation between distant language pairs, such as English and Japanese, inaccurate, because they have exceptionally dissimilar word orders as their nature. It is therefore the final goal of this study to establish a machine translation system for distant language pairs, using a practical reordering model that can handle accurate word ordering.

Our challenge primarily involves two types of obstacles we need to overcome. One obstacle is the computational complexity of the word ordering problem that has given limits to practical machine translation systems. Another obstacle is the problem of exceptionally dissimilar word orders in distant language pairs, such as English and Japanese, which has reduced machine translation accuracy to date.

In order to tackle our challenges, we make use of a promising approach called syntax-based preordering for statistical machine translation. In this approach, we use syntactic parsers that automatically parse input texts and output parse trees. We then modify the parse trees so that they represent much similar word orders to translation output than before. After that, the modified parse trees are used as input to a statistical machine translation system, which automatically learns a machine translation model from large data sets as similar to real world applications. Since this approach effectively divides the problem of complex machine translation into a separated reordering step and a translation step, we can fully focus on our challenges with two novel proposals.

The first proposal is a rule-based approach. We present two rule-based preordering methods named a two-stage method and a three-stage method. The two-stage method reorders a Japanese parse tree as similar to English using deep syntax information obtained with a predicate-argument structure analyzer. The three-stage method mimics the two-stage method by using little or no syntax. We eventually demonstrate the state-of-the-art performance in Japanese-to-English translation to date as a rule-based preordering approach.

The second proposal is a statistical approach. The statistical approach automatically learns reordering rules, unlike the rule-based approach. We present a simple yet effective statistical preordering method. In this method, we employ a greedy optimization strategy for modifying parse trees so that the modified parse trees maximize our objective function for reordering. We achieve the state-of-the-art accuracy in both English-to-Japanese and Japanese-to-English translation.

Form8・Separate Sheet
(様式８・別紙)

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏　　名　星野　翔

Title
論文題目　Syntax-based Preordering for Statistical Machine Translation

申請者は、統計的機械翻訳を高精度化するために入力単語列を翻訳先言語の語順に近づける手法である事前並べ替え手法を提案し、その有効性を大規模日英・英日翻訳データを用いた実験において実証した。

本論文は、全５章から構成される。第１章では、機械翻訳における重要な問題として言語間の語順の違いを解決する語順並べ替え問題について指摘し、機械翻訳の歴史を概観しながら語順並べ替え手法の重要性を議論している。そして、本論文の貢献として、入力文の構文構造を利用した事前並べ替え手法の提案と、機械翻訳実験による有効性の検証を主張している。

第２章では、本研究の背景として、構文構造・構文解析技術といった自然言語処理の基礎理論・技術、統計的機械翻訳の数理モデル、機械翻訳や事前並べ替えで用いられる機械学習手法、および機械翻訳における評価手法について網羅的な説明を行なっている。

第３章では、係り受け構造および述語項構造を利用したルールベースの事前並べ替え手法として２つの手法を提案している。第１の手法（二段階手法）は、第１段階で各動詞について主語と目的語を認識して動詞を主語と目的語の間に移動し、第２段階で文節内の内容語と機能語の順番を入れ替えることで、日本語文の語順を英語に近づける。第２の手法（三段階手法）は、構文解析誤りの影響を軽減するために、第１段階として述語項構造を利用して並列句の認識を行い、次に表層的情報を利用して第２段階（文節の入れ替え）と第３段階（文節内の入れ替え）の処理を行う。これらの手法の有効性を、大規模日英翻訳データを用いた実験により検証し、いずれもベースライン翻訳システムおよび既存の並べ替え手法より高い精度を達成することを示した。

第４章では、統計的機械学習を応用した事前並べ替え手法を提案し、その有効性を実証している。第３章の研究により事前並べ替えの有効性がある程度示されたものの、残された問題の分析から、並べ替えルールの追加やチューニングが避けられないことが指摘され、その解決法として並べ替えルールを統計的に学習するアイディアが導入されている。入力文を構文解析することで得られる句構造木の各ノードを走査し、子ノードの順番を入れ替えるか否かを機械学習で分類する。この時、学習のために正解データが必要となるが、ノードを入れ替えるか否かの正解データを直接作成することは事実上不可能である。そこで、ノードが支配する単語の並び順と翻訳先の文の語順との順位相関係数 Kendall's τ を計算し、それを最大化するようにノード入れ替えの正解ラベルを決定する手法を提案した。これに基づき各ノードについて最適なラベルを求めれば文全体に対する最適な並べ替えが得られることを示し、本手法に対する理論的根拠を与えた。さらに、分類器の精度向上のために新たな特徴量を提案した。大規模日英および英日翻訳データを用いた実験において、

本手法により既存手法を上回る翻訳精度が得られることを示した。

第5章では、以上の研究結果をまとめ、将来課題について議論している。

博士論文の内容については、提案手法、評価実験・分析、既存研究のサーベイなど、博士論文として十分なオリジナリティとクオリティがあるとの評価がなされた。本論文の内容は、査読付きジャーナル「情報処理学会論文誌」への採録が決定されている。以上のことから、全審査委員一致で、本論文は学位授与に値するとの判断に至った。