氏　　名	Han Dan
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1723 号
学位授与の日付	平成26年9月29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 Pre-reordering for Chinese-to-Japanese Statistical Machine Translation
論 文 審 査 委 員	主　査　准教授　宮尾　祐介 教授　神門　典子 准教授　Nigel COLLOER 教授　相澤　彰子　国立情報学研究所 教授　磯崎　秀樹　岡山県立大学

(別紙様式 2)
(Separate Form 2)

論文内容の要旨

Summary of thesis contents

Bilingual phrases are the main building blocks in statistical machine translation (SMT) systems. At training time, the most likely word-to-word alignment is computed and several heuristics are used to extract these bilingual phrases. Although this strategy performs relatively well when the source and target languages have a similar word order, the quality of extracted bilingual phrases diminishes when translating between languages structurally different, such as Chinese and Japanese. Syntax-based reordering methods in preprocessing stage have been developed and proved to be useful to aid the extraction of bilingual phrases and decoding. For Chinese-to-Japanese SMT, we carry out a detailed linguistic analysis on word order differences of this language pair to improve the word alignment. Our main contribution is threefold: (1) We first adapt an existing pre-reordering method called Head-finalization (HF) (Isozaki et al. 2010) for Chinese (HFC) (Han et al. 2012) to improve Chinese-to-Japanese SMT system's translation quality. HF is originally designed to reorder English sentences for English-to-Japanese SMT and it performs well. However, our preliminary experiments results reveal its disadvantages on reordering Chinese due to particular characteristics of languages. We thus refine HF to HFC based on a deep linguistic study. To obtain the required syntactic information, we use a head-driven phrase structure grammar (HPSG) parser for Chinese. Nevertheless, the follow-up error analysis from the pre-reordering experiment explores more issues that bring difficulties for further improvement on HFC, such as the tree operation restriction of binary tree, inconsistency on definition of linguistic term and so on. (2) We then propose an entire new pre-reordering framework which is using an unlabeled dependency parser to achieve additional improvements on reordering Chinese sentences to be like Japanese word orders. We refer to it as DPC (Han et al. 2013a) for short. In this method, we first identify blocks of Chinese words that demand reorderings, such as verbs and certain particles. Then, we detect the proper position which is the right-hand side of their rightmost object dependent, since our reordering principle is to reorder a Subject-Verb-Object (SVO) language to resemble a Subject-Object-Verb (SOV) language. Other types of particles are relocated in the last step. Unlike other reordering systems, the boundaries of verbal blocks and their rightmost object in DPC are defined only by the dependency tree and part-of-speech tags. Additionally, dismissing of using structural and punctuation border is another benefit for the reordering of the reported speech frequently occurring in news domain. The experiments show advantages of DPC over the SMT baseline (Moses) and our HFC systems. Important advantages of this method are the applicability of many reordering rules to other SVO and SOV language pairs as well as the availability of dependency parsers and POS-taggers for many languages. Considering our pre-reordering methods of HFC and DPC are linguistically-motivated, both are sensitive to parsing errors, even though DPC is designed to be more fault-tolerant parsing method by reducing the use of syntactic information, i.e.,

dependency labels. For future work on improving DPC or other reordering methods, it is meaningful to observe how parsing errors influence reordering performance. (3) We hence take a deep observation about the effects of parsing errors on reordering performance (Han et al. 2013b). We combine empirical and descriptive approaches to carry out a three-stage incremental comparative analysis on the relationship between parsing and pre-reordering. Our conclusion can be used to benefit not only for the improvements of syntax-based pre-reordering methods, but also for the developments of POS taggers and parsers.

(別紙様式 3)
(Separate Form 3)

博士論文の審査結果の要旨

Summary of the results of the doctoral thesis screening

本審査は、出願者による約 45 分間のプレゼンテーションの後、20 分程度の質疑応答が行われた。また、その後出願者により博士論文の改善点についての説明が行われた。

博士論文は全 7 章から構成される。第 1 章では、機械翻訳の社会的重要性、特に中国語と日本語の翻訳の必要性が説明され、その技術的困難について議論している。現在は統計的機械翻訳が主流であるが、中国語と日本語のように文法的に遠い言語ペアでは単語間の翻訳関係や対訳フレーズの学習が困難である。これを解決する手法として、中日機械翻訳のための事前並べ替え手法を提案している。

第 2 章では、研究の背景として、統計的機械翻訳の基盤技術や数理モデル、構文解析技術、および実験に使用した対訳テキストデータについて説明している。

第 3 章では、関連研究として、様々な言語ペアにおける事前・事後並べ替えおよび単語並べ替えの統計モデルの既存研究、および構文解析や機械翻訳のエラー分析に関する既存研究について議論を行っている。さらに、博士論文で提案している技術の基礎となっている Head Finalization（主語後置並べ替え）手法について詳細な説明を行っている。Head Finalization は、構文解析を行った後、各フレーズの主辞を後ろ側へ移動するという手法で、英日翻訳の精度を大幅に向上させることが知られている。

第 4 章では、Head Finalization を中国語へ適用するための手法を提案している。Head Finalization をそのまま適用した場合は、語順が日本語とは大きく異なってしまうケースがあり、翻訳精度の向上が限定的であった。翻訳結果の分析により、これは中国語と日本語では主辞の定義が異なることが原因であるとの知見を得た。これに基づき、主辞の定義が異なる場合（例えば、中国語では動詞の否定を表す「不」は修飾語であるが、日本語では「無い」が主辞になる）を列挙し、Head Finalization 手法を改良した。特許および新聞の大規模対訳データを用いた実験において、中日翻訳の精度が大幅に向上することを示した。

第 5 章では、係り受け構造を用いた事前並べ替え手法を提案している。第 4 章の手法は、フレーズ間の順番を入れ替えるという操作によって語順を変えるため、フレーズ境界を破るような並べ替えは不可能であった。そこで、係り受け構造を手がかりにし、動詞とその付属語を目的語の右側に移動する手法を提案した。中国語と日本語は文の構造は大きく異なる（中国語は主語の次に動詞が来る）が、名詞句などの構造はあまり差がない。そこで、動詞だけを移動することで、日本語の語順に近づけることができる。この手法は、構文木の構造に制約されずに移動を行うことができ、さらに係り受け構造の情報を必要最低限しか使わないため構文解析誤りに対して頑健であるという利点がある。第 4 章と同じデータで実験を行い、提案手法はさらに翻訳精度を向上させることを示した。

第 6 章では、構文解析誤りと事前並べ替え精度との関係について詳細な分析を行っている。本研究の提案手法はいずれも構文解析を利用しているため、構文解析誤りによって並べ替えが間違ってしまうという問題がある。そこで、どのような構文解析誤りが致命的な並べ替え誤りを引き起こしてしまうのかを定量的・定性的に分析した。その結果、構文解析誤りはわずかではあるが事前並べ替えの精度を下げてしまうこと、特に動詞の係り先・係り元の誤りと、文全体の主辞の認識誤りが並べ替えの精度に大きく影響することが分かった。

第 7 章では、以上の知見に基づき本論文の貢献を述べ、将来の研究課題について議論している。

(別紙様式 3)
(Separate Form 3)

質疑応答では、実験結果の詳細や解釈、既存研究との差分、提案手法をさらに改善するためのアイディアや他の言語ペアへの応用可能性などについて質問がなされ、おおむね的確に答えた。予備審査において指摘された修正すべき点についても的確に対応しており、博士論文の内容についてもオリジナリティ、評価実験、分析など十分であるとの評価がなされた。本論文の内容はジャーナル「自然言語処理」に採録されており、また一部の内容は Pacific Asia Conference on Language, Information, and Computation などの国際会議に採択されている。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。