（ 続紙 １ ）

京都大学	博士（情報学）	氏名	Fabien Cromieres
論文題目	Using Scalable Run-Time Methods and Syntactic Structure in Corpus-Based Machine Translation （スケーラブルな実行時手法と構文木に基づくコーパスベース機械翻訳）		

（論文内容の要旨）

本論文は、コーパスベース機械翻訳に関して、数百万文規模の大規模な対訳コーパスを扱うことができるスケーラブルな手法で、かつ、事前の計算量を抑え、余分な情報を保存する必要がない実行時手法について論じている。さらに、日英や日仏などの言語間の構造の違いを吸収するために構文木を利用する手法に関する議論を行っており、全６章から構成されている。

第１章は序論であり、コーパスベース機械翻訳手法の概要を述べた後、それを大規模な対訳コーパスに適用するために解決すべき問題について議論している。また、高精度な翻訳を実現するために翻訳において構文情報を利用することが考えられるが、そのために解決すべき課題を整理している。

第２章ではコーパスベース機械翻訳手法の歴史を概観し、そこでの問題点などを指摘している。またコーパスベース手法の例である用例ベース翻訳と統計翻訳との差が近年小さくなってきていることについて議論している。

第３章では、まず対訳コーパスからの翻訳知識獲得に必要な技術である対訳文アライメントの現状の問題点を整理し、確率的アライメントモデルの新たな枠組みを提案している。近年のほとんどのアライメントアルゴリズムは統計的なモデルを構築し、与えられた対訳文に確率を与えるものである。モデルのパラメータは対訳コーパスを対象とした最尤推定により学習され、各文のアライメントは尤度が最大となるものを探索することで行われる。しかし、一般的にアライメントモデルは複雑で、最適なパラメータの計算を完全に行うことは困難であり、様々な部分でアドホックなヒューリスティクスが利用される。これを改善するために、因子グラフで表現された対訳文上で確率伝搬法を用いることにより、周辺確率の計算や最尤状態を求めることが不可能な場合であっても、効率的にそれらの近似を求める枠組みを提案する。周辺確率が計算できればEMアルゴリズムを用いたモデルの学習が可能となり、最尤状態が求まれば対訳文内の最適な対応関係を求めることができる。２０万文の英仏対訳コーパスを用いた実験により、アライメントの精度がベースラインと比較してF値で６ポイント向上することを示した。次にこれを拡張し、やはり確率伝搬法を利用することにより，文の構文情報を取り入れたアライメントモデルを提案している。最後に、単語分割に依存しないアライメントアルゴリズムを提案している。このアルゴリズムでは事前のトレーニング段階ではなく、実行時にパラメータの計算が可能であることが特徴である。

第４章では、木構造で表現された大規模用例データベースから、クエリの木構造にマッチするものを効率的に検索する手法を提案している。木構造ではなく単語列を扱う翻訳モデルでは接尾辞配列を利用して効率的に用例を検索する手法が提案されているが、これを木構造でも扱えるように一般化することは難しい。これは、ある単語列の部分単語列は単語列のサイズに対して２次関数的にしか増加しないが、木構造では連続な部分木のパターンが指数関数的に増加するためである。そこで、この問題を解

決し、大規模データベースでも効率的に木構造の検索が行える手法を提案する。これはクエリ木構造の部分木の出現を、大きさ１から徐々に大きくしていき、再帰的に求めるという考えに基づいている。この手法ではラベル付けされた木構造に対して接尾辞配列の考え方を適用することにより、大規模なデータベースであっても効率的かつ高速に検索が行える。実験により、全ての部分木をハッシュテーブルとして保存するナイーブな手法と比較して、提案手法はディスクスペースを５％以下に抑えることができ、さらに検索速度は５倍高速であることを示した。さらにこの手法を拡張し、木構造の一部を品詞などのレベルに汎化することにより、語彙のレベルで完全に一致しない場合であっても、近い用例を検索することを可能にした。またn-best構文解析結果を用いる方法や、木構造として不連続なものを検索する方法についても述べている。

第５章では、構文木を利用したコーパスベース機械翻訳手法に同期木置換文法(STSG)を用いることを考え、この文法による既存のデコーディングアルゴリズムの改善を提案している。STSGは効率的にデコーディングが行えるため広く利用されているが、ある種の言語間での構文構造の相違が表現できないなど、文法の表現力に限界がある。本章ではこれを克服するための新たな同期文法を提案している。その特徴の一つは文法の非対称性、つまり原言語側と目的言語側で異なる文法を用いることである。原言語側では線形時間で解析が可能な木置換文法を用い、目的言語側ではより複雑な言語現象を扱うことができる木配列文脈自由文法を用いる。実験により、提案した文法はSTSGよりも柔軟に多くの翻訳ルールを学習することが可能であり、翻訳精度が向上することを示した。

第６章は結論であり、本論文を総括し、今後の展望について議論している。

（続紙 ２ ）

（論文審査の結果の要旨）

本論文はコーパスベース機械翻訳手法において、学習に用いる対訳コーパスを大規模にすることで生じる問題、および各言語文の構文情報を取り入れるにあたり解くべきいくつかの問題について、それらの解決方法を提案したものであり、得られた主要な成果は以下のとおりである。

１．グラフィカルモデルの一つである因子グラフで表現された対訳文上で確率伝播法を用いた確率的アライメントモデルの新たな枠組みを提案した。アライメントモデルを因子グラフで表現することが容易であることを示し、モデルの学習が確率伝播法で効率的に行えることを示した。２０万文の英仏対訳コーパスを用いた実験により、提案手法を既存のアライメントモデルに適用し、アドホックなヒューリスティクスを用いた手法よりも効率的に学習を行うことができ、かつアライメントの精度もベースラインと比較してF値で６ポイント向上することを示した。

２．木構造で表現された大規模用例データベースから、クエリ木構造にマッチするものを効率的に検索する手法を提案した。クエリ木構造の部分木のうち、小さいものの出現のANDを取ることにより、徐々に大きな部分木の出現を検索する手法であり、単語列を扱う手法で利用されている接尾辞配列の考え方を応用することにより高速なマッチングを実現した。さらに１つの対訳文から用例として抽出可能な全ての部分木を保存しておく必要がないため、使用するディスクスペースやメモリサイズを抑えることが可能である。実験では、全ての部分木をハッシュテーブルとして保存するナイーブな手法と比較して、提案手法はディスクスペースを５％以下に抑えることができ、さらに検索速度は５倍高速であることを示した。またこの手法を拡張し、木構造の一部を品詞などのレベルに汎化することにより、語彙のレベルで完全に一致しない場合であっても、近い用例を検索することを可能にした。またn-best構文解析結果を用いる方法や、木構造として不連続なものを検索する方法も提案した。

３．構文情報を利用したコーパスベース機械翻訳手法における同期木置換文法（STSG)を用いた既存のデコーディングアルゴリズムについて、STSGの表現力の限界を示し、これらの限界を克服するための新たな同期文法を提案した。その特徴は、原言語側と目的言語側で異なる文法を用いることであり、原言語側では線形時間で解析が可能な木置換文法を用い、目的言語側ではより複雑な木配列文脈自由文法を用いた。実験により、提案手法はSTSGと比較して、入力に対してより柔軟に多くの翻訳パターンを学習することができ、翻訳精度も向上することを示した。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、平成２３年２月２４日実施した論文内容とそれに関連した試問の結果合格と認めた。