博士（工　学）　タンテリ　アンドリアマナカシナ

学 位 論 文 題 名

A Study on Recursive Division Method in Example-Based Machine Translation

（実例に基づく機械翻訳における再帰分割手法に関する研究）

学位論文内容の要旨

近年、社会の国際化及び情報化の進展に伴い、機械翻訳の需要が増大し様々な研究が行なわれて来た。しかし、これらの研究には、辞書などの言語資源の不足、解析ツールの不完全性、さらに翻訳手法そのものなど、種々の問題点があり、特に研究があまり行なわれていない言語間ではこの問題点は一層深刻である。機械翻訳は当初、文法など、文を構成する規則を利用する手法が主流であったが、巨大化した規則や辞書の作成およびそれらの保守の労力が膨大であるという問題が存在した。この問題を克服するために、実例に基づく手法が提案された。その基本的な考え方は、翻訳例コーパスを用意して、入力文と類似している原文を持つ翻訳例をコーパスから抽出し、それらを用いて類推し、翻訳を行なうというものである。近年、インターネットの普及などにより、電子化された二か国語コーパスが容易に入手できるようになったことから、本手法は実用的手法として有望視されている。この手法では、入力文の大部分をカバーできる例文が存在すれば翻訳が可能であるが、そのような翻訳例がなければ正確に翻訳することが困難である。また、入力文と類似文の対応関係の誤りがあると、翻訳の失敗を引き起こす。さらに、構文解析を利用する実例に基づく機械翻訳手法では構文解析ツールの不完全性も翻訳結果に悪影響を与える。

それらの問題を解決するために、本論文では、原文と訳文の対応関係を有する品詞情報付きコーパスを用いて、文を繰り返し分割して翻訳を行う手法を提案する。本手法は類似文検索部及び訳文生成部より構成される。類似文検索部では、入力文の各部分ごとにその部分が存在する翻訳例の原文の中で一番高い類似度を有する翻訳例をコーパスから抽出する。訳文生成部では、入力文を翻訳例と共通する部分とその右側部分と左側部分に分割する。共通部分は翻訳例のそれに対応する部分を用いて翻訳する。同様に他の翻訳例を用いて、翻訳されていない左側部分や右側部分を分割し翻訳する。そして、次にこれらの部分を共通部分を含む翻訳例を用いて組み合わせる。このような処理により文を分割することによって翻訳例の原文でカバーされない文を翻訳することが可能となる。また、翻訳を段階的に行うことによって異なる部分で発生する可能性がある誤りも回避できる。構文解析ツールの不完全性を回避するためにタグ付きコーパスを利用する。本研究では品詞情報のみを利用しているが、本手法は同様のアルゴリズムにより構文、意味などのより高度なタグにも対応できる。タグを高度化すると翻訳精度も向上するものと考えられる。なお、翻訳手法とともに、対応関

係付きコーパスの構築方法として、類推による半自動式の対応関係推定手法も提案する。

本翻訳手法を用いた実験では２，５００翻訳例をコーパスとして利用し、新しい４００文をシステムに翻訳させ、その性能を評価した。その結果、平均６２％の正翻訳率が得られた。これは翻訳例コーパスの少なさや非文が多い会話の世界を対象としたことを考えると良好な結果と考えられる。また、一つの翻訳例でカバーできない入力文が複数の翻訳例を用いて正確に翻訳されることが確認できた。

本論文は５章からなる。第１章では、機械翻訳の必要性および様々な従来手法とその問題点を概観し、提案手法の概要を説明する。

第２章では、本翻訳手法が使用する単語間の対応関係付コーパスの構築のために開発された類推による半自動式の対応関係推定手法について述べる。対応関係付の既存のコーパスを用いて、新しい一組の翻訳例の対応関係を推定し、決定する。その結果が誤りを含む場合、人手で校正し、新しい翻訳例を翻訳例コーパスに追加する。このような方法により翻訳例コーパスが最初空であっても、対応関係付きコーパスを少しずつ構築することができる。従来の統計的手法では大量なコーパスを必要とし、さらに、一対多や多対多の対応関係や一文に数回出現する同じ単語の対応関係も正確に決定できない。このような問題を解決するために本手法を提案した。

第３章では、本翻訳手法について述べる。本章は二部よりなる。第一部は本手法に適切な類似文検索アルゴリズムとその類似度について述べる。入力文と共通している部分を検索するために部分マッチング手法を使用する。なお、できるだけ入力文と構造的に類似している例文を抽出するために、比較は共通部分に留まらず、文頭や文末まで続ける。第二部は再帰分割翻訳手法について述べる。文を分割して、類似文と共通する部分を翻訳しながら入力文を書き換えるという基本的な考え方やその繰り返しの流れを説明し、抽出された翻訳例の優先順位及び未対応語の処理方法について述べる。

第４章では、本手法を用いた実験方法とその有効性を確認した実験結果およびシステムの出力結果の例を示し、考察する。翻訳例の少なさにも関わらず良好な結果が得られたことおよび一例文でカバーされない文をうまく翻訳できた結果について述べる。提案された翻訳例の優先順位および未対応語の処理方法から良好な結果が得られた。また、その問題点についても述べる。

第５章では、本論文で提案した手法をまとめ、さらに、今後の展望として、コーパスから抽出される統計的な情報や品詞情報以外のタグの利用について述べる。

学位論文審査の要旨

主　査　教　授　栃内香次
副　査　教　授　青木由直
副　査　教　授　北島秀夫
副　査　助教授　荒木健治

学位論文題名

A Study on Recursive Division Method in Example-Based Machine Translation

(実例に基づく機械翻訳における再帰分割手法に関する研究)

社会の国際化、情報化の進展に伴い、機械翻訳の実用化に対する要求が高まり、それを受けて活発な研究が行われているが、未だ種々の未解決な問題点がある。従来、文法を中心とし、主として文を構成する規則を利用する翻訳手法が主流であったが、多様な文を翻訳するために膨大な規則や辞書を作成し、それらを保守する必要があることと、会話文など文法的に不正確な文を含む場合のロバスト性などの問題が存在する。

これらの問題に対処する手法として、近年、実例に基づく翻訳手法が提案された。その基本的な考え方は、大量の翻訳例コーパスを用意し、翻訳対象文と類似している原文を持つ翻訳例をコーパスから抽出し、それからの類推によって翻訳を行うというものである。最近、インターネットの普及などにより、電子化された二か国語コーパスが容易に入手できるようになったことから、本手法は実用的手法として有望視されている。本手法は、多数の入力文について類似した翻訳例が存在することが良質な翻訳を可能とするための必須の条件であり、大量の翻訳例コーパスの構築と、入力文と翻訳例文との類似度計算アルゴリズムの開発が重要となる。さらに、入力文の構文解析を行ってから翻訳例との類似度計算を行う手法では構文解析アルゴリズムの不完全性が問題となる。

本論文は、これらの問題を解決するために、品詞情報付き対訳コーパスを用い、入力文の分割をくり返して翻訳を行う再帰的分割手法を提案している。本手法は類似文検索部及び訳文生成部より構成される。類似文検索部では、入力文の各部分ごとに、翻訳例原文中で最も高い類似度を有する翻訳例をコーパスから抽出する。訳文生成部では、まず入力文を翻訳例と共通する部分、その左側部分および右側部分に分割する。共通部分は翻訳例の対応する部分を用いて翻訳し、ついで、他の翻訳例を用いて左側部分と右側部分を同様に分割し翻訳する。この操作をくり返し、最後に全体を組合わせて翻訳を完成する。このように文を分割して再起的に処理することによって、1文の翻訳例のみではカバーされない文を翻訳することが可能となる。さらに本論文では、翻訳例コーパスの構築方法として、

類推による半自動式単語間対応関係推定手法を提案している。

本論文は５章からなる。第１章では、機械翻訳の必要性および様々な従来手法とその問題点を概観し、提案手法の概要を説明している。

第２章では、本翻訳手法で使用する単語間対応関係付対訳コーパスの構築のために開発された、類推による半自動式の対応関係推定手法について述べている。

第３章では、二部に分けて翻訳手法について述べている。第一部では類似文検索アルゴリズムについて述べ、第二部では再帰的分割手法について述べている。

第４章では、本手法の有効性を確認するための実験について述べている。実験では、翻訳例コーパスとして２５００文を用い、会話文を含む４００文を翻訳し、性能評価を行った。その結果、平均６２％の正翻訳率が得られた。これは翻訳例コーパスが極めて少量であることと、翻訳対象に非文が含まれていることを考慮すると、かなり良好な結果と考えられる。また、一つの翻訳例では翻訳できない入力文が複数の翻訳例を用いて正確に翻訳されることを確認している。

第５章では、本論文で提案した手法をまとめ、さらに、今後の展望として、コーパスから抽出される統計的な情報や品詞情報以外のタグの利用について述べている。本論文では品詞情報のみを利用しているが、同様のアルゴリズムにより構文、意味などのより高度な情報を有するコーパスにも対応でき、翻訳精度向上の可能性を有することを述べている。

これを要するに、著者は、実例に基づく機械翻訳手法において、品詞情報付き対訳コーパスを用い、翻訳対象文を分割して対訳コーパスと部分的に一致する部分をくり返し求め、全文の翻訳を完成させる再帰的分割手法と、品詞付き対訳コーパスの半自動的構築法を提案し、実験によりその有効性を実証するとともに、さらに高度な構文、意味情報付き対訳コーパスを用いてより高精度な翻訳品質を得ることの可能性を示したもので、自然言語処理工学ならびに情報メディア工学の発展に寄与するところ大である。

よって著者は、北海道大学博士（工学）の学位を授与される資格あるものと認める。