博 士（情報科学） 寺 島 涼

学 位 論 文 題 名

コーパスに基づく機械翻訳における対訳コーパスの有効利用に関する研究

学位論文内容の要旨

近年，非常に多くの機械翻訳システムを利用することが可能となっている．しかし現状では，十分な翻訳品質をユーザに提供するには至っていない．古くから研究されてきた解析型機械翻訳手法では，人手であらかじめ文法規則や変換規則などを記述し，それらの規則に基づいて解析的に翻訳を行う．解析型機械翻訳手法は多様な言語現象を規則として記述することが困難であること，また規則の追加などの改良により，新たな誤翻訳が生まれるという副作用の問題がある．

このような問題を解決するために，現在では，コーパスに基づく機械翻訳手法の研究が盛んに行われている．コーパスに基づく機械翻訳手法は，対訳文の集合である対訳コーパスに基づき翻訳知識を構築し，それらを利用して翻訳を行う．翻訳知識を構築するためには，対訳文の原言語文と目的言語文間における，部分的な対応関係を決定することが有効である．

コーパスに基づく機械翻訳手法は，用例ベースの機械翻訳手法と統計的機械翻訳手法に大きく分類することができる．用例ベース翻訳では，静的な言語知識である，品詞情報及び構文情報を用いることで，対訳文における部分対応を利用する．つまり，対訳コーパスを翻訳知識として利用するためには，対訳コーパスにおける原言語と目的言語それぞれに対する解析ツールが必要である．よって，解析ツールが十分でない言語に適用することは困難である．また，統計翻訳は，大規模な対訳コーパスより，統計情報に基づき，統計的言語モデルと，部分対応の集合である統計的翻訳モデルを構築する．しかし，統計翻訳は，大規模な対訳コーパスの使用を前提としているため，データスパースである場合や，少量の対訳コーパスにおいては，効率よく部分対応を決定することが困難となり，その結果，翻訳精度の低下を招く．

著者はデータスパース及び量が十分でない対訳コーパスにおいて，解析ツールに強く依存せずに翻訳知識を向上させるためには，対訳コーパスをより効率的に利用する必要があると考えた．そこで，本研究では，省略可能情報を用いた部分対応学習の提案と，対訳文一般化による翻訳ルールを導入した統計翻訳の提案を行った．

省略可能情報を用いた部分対応学習は，対訳文中の句に相当する部分対応を効率よく決定するための学習手法である．本手法は，対訳文中の省略可能な部分に着目することにより，対訳知識である抽出ルールを自動獲得する．獲得された抽出ルールは，対訳文中の句に相当する部分の探索範囲を限定するための情報を有している．この抽出ルールを様々な対訳文に適用することにより，対訳文中の句レベルの部分対応を効率よく決定できる．その結果，解析ツールに強く依存することなく，多言語への適用が可能となると考えられる．著者は，省略可能情報を用いた部分対応学習を，学習型機械翻訳に適用した．本論文で用いる学習型機械翻訳は，対訳コーパスより帰納的学習に基づき翻訳ルール

を自動獲得し，それらを用いて翻訳を行う．学習型機械翻訳に本手法を適用することで，学習能力の向上の観点より，本手法の有効性の検証を行った．本手法を学習型機械翻訳に適用することで翻訳ルールの効率的な獲得が可能となり，人手による評価において翻訳精度が向上した．更に，様々な自動評価手法を用いた場合においても，それらのスコアの向上が確認された．また，対訳文中の効率的な部分対応決定の観点での検証を行った結果，本手法は低頻度の部分対応を効率よく決定できることが明らかとなった．これらの結果より，本手法が対訳コーパスに基づく学習型機械翻訳に有効であることが確認された．

対訳文一般化による翻訳ルールを導入した統計翻訳は，フレーズベース統計翻訳に対し，文の形を有する翻訳ルールを組み合わせた手法である．従来のフレーズベース統計翻訳は，対訳コーパスより構築された翻訳知識である，統計的言語モデルと統計的翻訳モデル (フレーズテーブル) を用いて翻訳を行う．フレーズテーブルによるフレーズ翻訳と，言語モデルによるフレーズ翻訳の並び替えにより，翻訳文を生成するため，翻訳文を生成するという点においては非常に頑健性が高い．しかしながら，それぞれの翻訳知識は文の形を保持してはいないため，不自然な翻訳文を生成することがある．そこで，本手法では，フレーズベース統計翻訳に対し，文の形を保持した翻訳ルールを組み合わせることで，この問題の解決を図った．本手法を用いた翻訳システムでは，翻訳対象文が入力されると，対訳コーパスに含まれるそれぞれの対訳文と翻訳対象文における差異・共通部分を決定する．次に，翻訳対象文に対し差異・共通部分を有する対訳文を選択し，単語の共起頻度に基づく統計情報を用いて，選択された対訳文における部分対応を決定する．対訳文より，決定された部分対応を一般化することで翻訳ルールを自動獲得する．以上より，解析ツールに強く依存することなく，対訳コーパスをより効率的に活用することが可能となり，翻訳対象文に適した翻訳ルールを獲得することができる．性能評価実験の結果，フレーズベース統計翻訳に対して，対訳文一般化による翻訳ルールを導入することで，人手による評価，様々な自動評価，それぞれにおいて翻訳精度の向上が確認された．

このように，本論文で著者は，対訳コーパスをより有効利用することで，翻訳知識の向上を図った．本論文で提案を行った，省略可能情報を用いた部分対応学習，対訳文一般化による翻訳ルールを導入した統計翻訳，それぞれの手法において，解析ツールを新たに使用することなく，対訳コーパスの有効利用に基き，翻訳知識の向上を実現した．これらの手法は，特定の言語に強く依存しない手法であるため，様々な言語へ適用することが可能であると考えられる．今後は，様々なコーパスを利用した実験を通じて，多言語翻訳の実現へ向けた取り組みが必要であると考えられる．

学位論文審査の要旨

主　査　教　授　荒　木　健　治
副　査　教　授　山　本　　強
副　査　教　授　長谷山　美　紀

学位論文題名

コーパスに基づく機械翻訳における対訳コーパスの有効利用に関する研究

近年，非常に多くの機械翻訳システムを利用することが可能となっている．しかし現状では，十分な翻訳品質をユーザに提供するには至っていない．古くから研究されてきた解析型機械翻訳手法では，人手であらかじめ文法規則や変換規則などを記述し，それらの規則に基づいて解析的に翻訳を行う．解析型機械翻訳手法は多様な言語現象を規則として記述することが困難であること，また規則の追加などの改良により，新たな誤翻訳が生まれるという副作用の問題がある．

このような問題を解決するために，現在では，コーパスに基づく機械翻訳手法の研究が盛んに行われている．コーパスに基づく機械翻訳手法は，対訳文の集合である対訳コーパスに基づき翻訳知識を構築し，それらを利用して翻訳を行う．翻訳知識を構築するためには，対訳文の原言語文と目的言語文間における，部分的な対応関係を決定することが有効である．

コーパスに基づく機械翻訳手法は，用例ベースの機械翻訳手法と統計的機械翻訳手法に大きく分類することができる．用例ベース翻訳では，静的な言語知識である，品詞情報及び構文情報を用いることで，対訳文における部分対応を利用する．つまり，対訳コーパスを翻訳知識として利用するためには，対訳コーパスにおける原言語と目的言語それぞれに対する解析ツールが必要である．よって，解析ツールが十分でない言語に適用することは困難である．また，統計翻訳は，大規模な対訳コーパスより，統計情報に基づき，統計的言語モデルと，部分対応の集合である統計的翻訳モデルを構築する．しかし，統計翻訳は，大規模な対訳コーパスの使用を前提としているため，データスパースである場合や，少量の対訳コーパスにおいては，効率よく部分対応を決定することが困難となり，その結果，翻訳精度の低下を招く．

著者はデータスパース及び量が十分でない対訳コーパスにおいて，解析ツールに強く依存せずに翻訳知識を向上させるためには，対訳コーパスをより効率的に利用する必要があると考えた．そこで，本研究では，省略可能情報を用いた部分対応学習の提案と，対訳文一般化による翻訳ルールを導入した統計翻訳の提案を行った．

省略可能情報を用いた部分対応学習は，対訳文中の句に相当する部分対応を効率よく決定するための学習手法である．本手法は，対訳文中の省略可能な部分に着目することにより，対訳知識である抽出ルールを自動獲得する．獲得された抽出ルールは，対訳文中の句に相当する部分の探索範囲を限定するための情報を有している．この抽出ルールを様々な対訳文に適用することにより，対訳文中の句レベルの部分対応を効率よく決定できる．その結果，解析ツールに強く依存することなく，多言語への適用が可能となると考えられる．著者は，省略可能情報を用いた部分対応学習を，学習型機械翻

訳に適用した. 本論文で用いる学習型機械翻訳は, 対訳コーパスより帰納的学習に基づき翻訳ルールを自動獲得し, それらを用いて翻訳を行う. 学習型機械翻訳に本手法を適用することで, 学習能力の向上の観点より, 本手法の有効性の検証を行った. 本手法を学習型機械翻訳に適用することで翻訳ルールの効率的な獲得が可能となり, 人手による評価において翻訳精度が向上した. 更に, 様々な自動評価手法を用いた場合においても, それらのスコアの向上が確認された. また, 対訳文中の効率的な部分対応決定の観点での検証を行った結果, 本手法は低頻度の部分対応を効率よく決定できることが明らかとなった. これらの結果より, 本手法が対訳コーパスに基づく学習型機械翻訳に有効であることが確認された.

対訳文一般化による翻訳ルールを導入した統計翻訳は, フレーズベース統計翻訳に対し, 文の形を有する翻訳ルールを組み合わせた手法である. 従来のフレーズベース統計翻訳は, 対訳コーパスより構築された翻訳知識である, 統計的言語モデルと統計的翻訳モデル (フレーズテーブル) を用いて翻訳を行う. フレーズテーブルによるフレーズ翻訳と, 言語モデルによるフレーズ翻訳の並び替えにより, 翻訳文を生成するため, 翻訳文を生成するという点においては非常に頑健性が高い. しかしながら, それぞれの翻訳知識は文の形を保持してはいないため, 不自然な翻訳文を生成することがある. そこで, 本手法では, フレーズベース統計翻訳に対し, 文の形を保持した翻訳ルールを組み合わせることで, この問題の解決を図った. 本手法を用いた翻訳システムでは, 翻訳対象文が入力されると, 対訳コーパスに含まれるそれぞれの対訳文と翻訳対象文における差異・共通部分を決定する. 次に, 翻訳対象文に対し差異・共通部分を有する対訳文を選択し, 単語の共起頻度に基づく統計情報を用いて, 選択された対訳文における部分対応を決定する. 対訳文より, 決定された部分対応を一般化することで翻訳ルールを自動獲得する. 以上より, 解析ツールに強く依存することなく, 対訳コーパスをより効率的に活用することが可能となり, 翻訳対象文に適した翻訳ルールを獲得することができる. 性能評価実験の結果, フレーズベース統計翻訳に対して, 対訳文一般化による翻訳ルールを導入することで, 人手による評価, 様々な自動評価, それぞれにおいて翻訳精度の向上が確認された.

このように, 本論文で著者は, 対訳コーパスをより有効利用することで, 翻訳知識の向上を図った. 本論文で提案を行った, 省略可能情報を用いた部分対応学習, 対訳文一般化による翻訳ルールを導入した統計翻訳, それぞれの手法において, 解析ツールを新たに使用することなく, 対訳コーパスの有効利用に基づき, 翻訳知識の向上を実現した. これらの手法は, 特定の言語に強く依存しない手法であるため, 様々な言語へ適用することが可能であると考えられる. 今後は, 様々なコーパスを利用した実験を通じて, 多言語翻訳の実現へ向けた取り組みが必要であると考えられる.

これを要するに, 著者は, 機械翻訳について, 翻訳知識の向上に関する新知見を得たものであり, 自然言語処理工学における機械翻訳技術の発展に貢献するところ大なるものがある. よって著者は, 北海道大学博士 (情報科学) の学位を授与される資格あるものと認める.