

学 位 論 文 題 名

コーパスからの翻訳知識の自動獲得手法に関する研究

学位論文内容の要旨

人間が持つ能力の一つとして、未知の状況に対する頑健さがある。実環境において人間は、未知の情報に頻繁に遭遇している。例えば、人間の幼児は、日々未知の言語現象に接していると考えられる。そのような環境において、人間の幼児は自らが持つ言語及び知識獲得能力により、未知の言語現象に内在する知識を動的に習得することができる。本研究では、人間が有する生得的能力と考えられる、データに内在する規則を動的に獲得する能力をコンピュータ上に実現することを目指す。その際に、コンピュータ上においては「2種類の記号列中の対応関係を決定する能力」が有効であるとし、この能力をコンピュータ上に実装することで未知の情報に対して頑健なシステムの実現を図る。また、「2種類の記号列中の対応関係を決定する能力」の向上を図ることは、記号列中に内在する規則の効率的な獲得をもたらすため、より高い学習能力を有するシステムを構築することにもなる。そこで本研究の目的は、「2種類の記号列中の対応関係を決定する能力」を実現するためのメカニズムの検証とその有効性及び問題点を性能評価実験を通して明らかにすることである。

本研究を行うにあたり、2種類の記号列には異なる言語の単語列を用いる。すなわち、システムは異なる2つの言語からなる単語列中から対応関係を決定することにより、そこに内在する規則を自動獲得する。その場合、コンピュータには「2種類の記号列中の対応関係を決定する能力」と同時に、「二つの事物が同じか異なるかを判断する能力」が備わっていることを前提としている。本研究では、このような能力より獲得される規則を翻訳知識と呼ぶこととする。ここで、異なる2つの言語からなる単語列とは自然言語の文であり、この自然言語文の電子化された集合体、すなわち、言語データはコーパスと呼ばれる。したがって、本研究では、コーパスに対して、「2種類の記号列中の対応関係を決定する能力」を適用することにより、翻訳知識を自動獲得する。このような能力は、特定の言語には依存しないため、対象言語が変更されてもコンピュータ上に同様なシステムを再構築する必要はなく、労力の削減の観点においても有効である。さらに、理論的には、このような能力は言語データに限定されるものではなく、様々な未知のデータに対して有効であると考えられる。しかし、本研究では、提案する学習能力を評価するために言語データを対象とし、その有効性について検証することとする。

本研究で用いる「2種類の記号列中の対応関係を決定する能力」を用いて獲得される翻訳知識は、機械翻訳というタスクにおいては、翻訳処理に使用する翻訳ルールと位置付けることができる。ここで翻訳ルールとは、対応関係にある2つの異なる言語文からなる翻訳例に対して、上述した能力を適用することで自動獲得される、ある言語の文（原言語文）から異なる言語の文（目的言語文）に変換するための規則である。本研究では、この

翻訳ルールはシステムの持つ学習能力により、与えられた翻訳例から動的に獲得される。したがって、未知の翻訳例が与えられた場合、システムはそこに内在する規則を自動獲得し、新たな翻訳規則として翻訳処理のために逐次蓄えていく。一方、現在の機械翻訳手法の主流となっている解析的なアプローチによる機械翻訳手法では、文を解析するための文法規則、2言語間を結ぶ変換規則、多義性を解消するための選択制約などの翻訳処理に必要な規則は、予め人手により与えられる静的な知識である。このようなアプローチでは、新たな言語現象に直面した場合、それらに対処可能な新たな規則の追加や改良が必要となる。しかし、その結果、それまでの規則集合との整合性が崩れ、それまでの翻訳精度に対して悪影響を及ぼすなどの副作用が大きな問題点として指摘されている。したがって、機械翻訳においても、未知のデータに対して、動的に翻訳規則を獲得する学習能力は非常に有効と考えられる。

本研究では、「2種類の記号列中の対応関係を決定する能力」を用いて翻訳ルールを自動獲得するが、その獲得方法として様々なメカニズムを用いている。その第一段階として、2つの翻訳例間の共通部分と差異部分を多段階に抽出することで翻訳ルールを獲得する帰納的学習に対し、遺伝的アルゴリズムの枠組みを適用した。すなわち、翻訳例を染色体として位置付け、その翻訳例に対して遺伝的アルゴリズムの基本操作である選択交配、突然変異を発生させることで、多様な翻訳例を自動生成する。その結果、帰納的学習が抱えていた、良質な翻訳ルールを獲得するためには類似した多くの翻訳例が必要になるという問題が解消される。さらに、誤った翻訳ルールに対して淘汰処理を行うことで、システム全体が良質な翻訳ルールを保持するために進化していく。中学1年生用レベルのコーパスを用いて行った性能評価実験の結果、遺伝的アルゴリズムの適用が、52.8%から61.9%の翻訳率の向上をもたらしたことを確認した。

さらに、本研究では、より効率的な翻訳ルールの獲得を目的に、上述した手法で獲得された翻訳ルールを起点として、翻訳ルールの自動獲得における連鎖を引き起こすメカニズムを導入した。すなわち、翻訳ルールの獲得が新たな翻訳ルールの獲得をもたらすという翻訳ルール獲得の連鎖反応が起こるということである。これは、「2種類の記号列中の対応関係を決定する能力」を翻訳例全体に対してではなく、局所部分に着目して適用することで実現している。このメカニズムの導入により、様々な翻訳例が混在するコーパス、すなわち、データスパースであるコーパスにおいても、良質な翻訳ルールの獲得が可能となった。中学2年生レベルのコーパスを用いて行った性能評価実験の結果、翻訳率として61.1%が得られ、言語データが高度化した場合でも、学習能力の向上により翻訳率が低下しないことを確認した。また、解析的なアプローチによる商用の機械翻訳システムとの比較においても、同等な翻訳率を導き出した。

そして、本研究で獲得される翻訳知識は、機械翻訳の基礎的な資源である対訳辞書の自動構築というタスクにおいては、対訳語と位置付けることができる。コーパスからの対訳語の自動抽出においては、提案する学習能力の汎用性を検証することを目的に、複数の言語のコーパスを用いて性能評価実験を行った。5つの異なる言語のコーパスからの対訳語の抽出率は、どの言語のコーパスにおいても60%前後となり、その汎用性を確認した。

本研究では、生得的能力として「2種類の記号列中の対応関係を決定する能力」をコンピュータ上に実装し、それを言語データに適用することでその有効性を検証した。性能評価実験を通して、未知の言語データが与えられた場合、そこに内在する規則の獲得に本メカニズムが有効であることを確認した。

学位論文審査の要旨

主 査 教 授 荒 木 健 治

副 査 教 授 北 島 秀 夫

副 査 教 授 山 本 強

学 位 論 文 題 名

コーパスからの翻訳知識の自動獲得手法に関する研究

著者は、言語及び知識獲得能力の工学的実現の観点から、動的に知識を自動獲得するための学習手法を提案した。そして、その学習手法の有効性を確認するために言語処理システムである機械翻訳システムと対訳語の自動抽出システムに応用した。著者の提案する学習手法では、実例からそこに内在するルールを自動獲得する。具体的には、実例として、意味的には近いが体系が大きく異なる2つの言語間の翻訳例を用いた場合、言語間の意味的な対応関係を有する翻訳知識を自動獲得する。翻訳例の集合体であるコーパスから動的に獲得された翻訳知識が、言語処理システムにおいて有効であることを性能評価実験より確認し、提案する学習手法の有効性を示した。

知識を獲得する人間の言語及び知識獲得能力を工学的に実現することは非常に重要な研究である。なぜなら、そのことにより、環境に応じて動的な知識の獲得が可能となる。したがって、そのような能力をコンピュータ上に実現できれば、汎用的なシステムを構築することが可能である。著者は、人間が有する言語及び知識獲得能力の工学的な実現の観点から、実例より動的に知識を自動獲得する学習手法を提案した。学習手法としては、核となるものとして、遺伝的アルゴリズムを用いた帰納的学習を提案した。遺伝的アルゴリズムを用いた帰納的学習では、「2つの事物が同じか異なるかを判断する能力」を生得的な能力として仮定し、この生得的な能力に基づきルールを自動獲得している。具体的には、2つの実例間において、共通部分と差異部分を多段階に抽出することでルールを獲得する。さらに遺伝的アルゴリズムを用いた帰納的学習では、より多くの良質な翻訳ルールを獲得するために不可欠な実例を自動生成するために、遺伝的アルゴリズムの基本操作を適用している。すなわち、実例に対し、交叉と突然変異を行うことで、新たな実例の自動生成を行っている。さらに、精度を適応度として淘汰処理を行っている。次に著者は、遺伝的アルゴリズムを用いた帰納的学習の能力向上を目的とした、再帰チェーンリンク型学習を提案した。再帰チェーンリンク型学習では、既存のルールを新たな翻訳ルールを獲得するための知識源として有効利用することで、データスパースな実例の集合から効率よくルールを獲得する。また、著者は実例中の局所部分に関するルールのみを獲得するものとして、隣接情報学習と帰納的連鎖学習の提案を行った。隣接情報学習と帰納的連鎖学習では、遺伝的アルゴリズムを用いた帰納的学習と再帰チェーンリンク型学習の考え方に基づき、実

例の局所部分に関するルールを自動獲得する。著者は、実例として、意味的に近いが体系が大きく異なる2つの言語間の翻訳例を用いることで、言語間の対応関係についてのルールの自動獲得を試みた。そして、獲得されるルールを翻訳知識と位置づけた。提案する学習手法を用いることで、言語間の対応関係に関する静的な知識に依存することなく、翻訳知識を動的に獲得することができる。

次いで著者は、提案する学習手法の有効性を検証するために、本手法を言語処理システムに応用した。遺伝的アルゴリズムを用いた帰納的学習と再帰チェーンリンク型学習においては、英日の機械翻訳システムへの応用を行った。そこで獲得される翻訳知識は、翻訳結果の生成に有効となる翻訳ルールである。遺伝的アルゴリズムを用いた帰納的学習を適用する際には、翻訳例を染色体に、単語を遺伝子と位置づけている。そして、翻訳例に対して交叉、突然変異を行うことで、翻訳例の自動生成を行っている。さらに、2つの翻訳例間において、共通部分と差異部分を多段階に抽出することで翻訳ルールを自動獲得する。再帰チェーンリンク型学習は、既存の翻訳ルールを有効利用することにより、連鎖的に翻訳ルールを自動獲得するためのメカニズムである。この再帰チェーンリンク型学習を導入することにより、データスパースな翻訳例からなるコーパスに対しても、効率よく翻訳ルールを獲得する。著者は、性能評価実験により、有効な翻訳率として61.1%、また、学習データに不足がない状況においては、有効な翻訳率として85.0%が得られたことを確認し、提案する学習手法の有効性を示した。

次に著者は、隣接情報学習と帰納的連鎖学習を、複数の言語の対訳語の自動抽出システムに応用した。そこでは、翻訳知識として、対訳語を効率よく自動抽出するために有効となるテンプレートを動的に獲得する。自動獲得されるテンプレートは、データスパースネスの問題に有効である。すなわち、テンプレートを用いることで、対訳文に対して探索範囲を限定することができ、効率よく対訳語を自動抽出することが可能である。著者は、5つの言語のコーパスを対象に対訳語の自動抽出を行った結果、Dice係数に基づくシステムに隣接情報学習を導入することで対訳語の抽出率が8ポイント向上したことを確認し、提案する学習手法の有効性を示した。さらに、隣接情報学習に対して、より効率よく、そして、より多くのテンプレートを自動獲得することを目的とした帰納的連鎖学習が、様々な類似度尺度に基づくシステムに対して有効であることを性能評価実験を通して示した。

以上を要約すると、著者は、言語及び知識獲得能力の工学的実現の観点より、実例から知識を動的に獲得するための学習手法を提案し、その有効性を確認するために、機械翻訳システムと対訳語の自動抽出システムへの応用を行った。その結果、提案する学習手法により、言語間の対応関係に関する静的な知識を必要とすることなく、動的に様々な言語間の翻訳知識を獲得可能であることを示した。また、このような学習手法は汎用性の高い手法と考えられるため、今後、非タスク依存の観点から、提案手法が様々なタスクに応用可能となることが期待される。このような研究は、情報メディア工学、自然言語処理工学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(工学)の学位を授与される資格あるものと認める。