

携帯端末における帰納的学習を用いた 文字列情報の縮退および復元に関する研究

学位論文内容の要旨

プログラミング言語など、人為的に規定された言語を人工言語と呼ぶのに対して、日本語や英語など、人間が日常使用している言語を自然言語と呼ぶ。自然言語処理とは、この自然言語を計算機上で取り扱うことである。人工言語は曖昧性がない、すなわち完全データの集合であるのに対して、自然言語は社会の中で自然に発生し、自然に用いられている言語であるため、一般に多くの曖昧性を含んでおり、不完全データの集合であるといえる。従って、自然言語を計算機上で取り扱うことは非常に困難である。しかしながら、人間はこの自然言語の持つ曖昧さを解消し、データの不完全な部分を復元することにより、完全なデータとして処理することが可能である。この能力を解明し、計算機上で同様の処理を実現することが本研究の最終目的である。

一般に、人間が自然言語を理解する過程においては、背景知識や文脈情報、あるいは目や耳など五感からの情報も利用していると考えられる。これらすべての情報を計算機上で取り扱うことは、現在の技術では困難である。そこで、本研究においてはテキストデータのみを取り扱うこととする。テキストデータを可逆な不完全データに変換し、この変換の際に失われた情報を復元することを考える。この変換された不完全データは、人間にとっても今までに取り扱ったことのないデータである。そのため、このようなデータに対して、背景知識や文脈情報を利用することは困難であると考えられる。しかしながら、人間はこのようなデータに対しても処理を繰り返し学習を行うことにより、字面情報のみからでも失われた情報を復元し、次第に元のテキストデータへの変換を正しく行うことができるようになっていく。このような処理を計算機上で実現するのが、本研究の目的である。

携帯電話等、携帯端末の性能の進歩に伴い、このような小型端末上で電子メール等のテキストデータを処理する機会と必要性が増大している。携帯電話は現在広く使用されている携帯端末である。一般的な携帯電話は、その大きさに制約があり、大きなキーボード、多数のキーを備えることができない。そこで、この少ないキーを用いて文字列を入力するために、文字列情報の縮退を考える。端末へ入力すべき情報を減少させることにより、携帯電話上の少ないキーからでも迅速な入力が可能となる。そして、この入力の際に縮退した情報を復元することを目指す。携帯電話はその大きさの制約から装備可能なキー数が

限られるが、最低でも0~9,#,*の12個のキーは装備しているのが普通である。この12キーを用いて、例えば日本語を入力することを考える。日本語には約50個のかな文字が存在するので、1つのキーに複数の文字を割り当てる必要がある。一般的には、あ行、か行などの一行が1つのキーに割り当てられている。また、一般的な入力方式としては、文字循環指定方式が採用されているが、この入力方式では、1文字の入力に複数回の打鍵が必要となる。よって本研究においては、迅速な入力を可能とするために文字情報縮退方式を採用し、1文字の入力を1打で行うものとする。これにより、例えば「野球」を入力する場合、文字循環指定方式では、 $1+2+5+3=11$ 回の打鍵数が必要なのに対し、文字情報縮退方式では、 $1+1+1+1=4$ 回の打鍵数で入力完了する。しかしながら、入力された数字1文字は意図したかな文字以外にも、それと同一行の他のかな文字にも対応することになり、結果として入力数字列は多数の日本語文に対応し、曖昧性が増大している。従って、このような入力数字列は上述した不完全データであるといえる。この不完全性を解消し、縮退した文字列情報を復元することが本研究の具体的な目的である。

これを実証するために、本研究では、取り扱う対象を徐々に高度化し、それに耐えうる応用システムを工学的に実現し、その性能評価を行っている。そこで最初に取り扱った対象は、携帯端末向けの日本語入力手法である。この手法における入力データは、かな文字列の母音情報が縮退した数字列である。これを漢字かな混じり文である日本語文に正しく変換することが本手法の目的である。入力数字列と日本語文とは、表層表現は一部異なるが記述体系は同じである。すなわち、入力数字列と日本語文においては文字の対応関係を左から順に決定可能なので、表層表現が一致する部分を手掛かりとして、左から順に単語に相当するセグメントを抽出することができる。このように字面情報に基づき、対応関係を変換ルールとして随時獲得することが本手法の基本であり、これを帰納的学習と呼ぶ。このようにして獲得された変換ルールを用いて変換を行っていくので、本手法においては入力データの増加に伴い、次第に変換精度を向上させることができる。評価実験の結果、80[%]以上の変換精度が確認された。

次に、携帯端末向けの中国語入力手法を対象とした。この手法の入力データは、中国語のピンインに対応した数字列である。ピンインとは、中国語の読みに対応し、ローマ字アルファベットにより表現される。入力数字列と中国語文とは、表層表現はまったく異なるが、記述体系は同じである。従って、左から順に1対1の対応を決定することができる。しかしながら、日本語とは異なり、表層表現が一致する部分が存在しないため、対応を決定する手掛かりが1組の変換例中には存在しない。そのため、ここで取り扱う対象は日本語文に対して高度化されているといえる。そこで、本手法においては、2組の変換例を使用し、それらの間で一致する文字列を手掛かりとして対応関係を獲得し、変換に利用していく。評価実験の結果、65[%]程度までの変換精度の上昇が確認された。

次に、携帯端末向け機械翻訳手法を対象とした。日英機械翻訳を前提としているので、本手法の入力データは日本語のかな文字列に対応した数字列である。この数字列を英語文に変換する。よって、入力数字列と英語文とは、表層表現も記述体系もまったく異なっ

おり、手掛かりとなる部分が存在しても対応関係を1対1に決定できない。そのため、この対象は上記2つの対象に比べて高度化されたデータであるといえる。そこで、この手法においての対応関係は単語列単位で決定する。すなわち、単語の語順を含めて翻訳ルールとして獲得し、翻訳に利用する。評価実験の結果、他手法と比較して高い翻訳効率が得られ、本手法の有効性が確認された。

以上のように、高度化したデータに対しても本手法が有効であることが示され、人間と同等に情報を復元する能力を、計算機においても実現できる可能性が示された。

学位論文審査の要旨

主 査 教 授 荒 木 健 治

副 査 教 授 青 木 由 直

副 査 教 授 北 島 秀 夫

学 位 論 文 題 名

携帯端末における帰納的学習を用いた 文字列情報の縮退および復元に関する研究

著者は、携帯端末において迅速な文字列入力を可能とするために、文字列の情報を縮退させて入力を行い、その際に失われた情報を、統計情報を用いた帰納的学習により復元する手法について提案した。

著者は、近年の携帯電話の普及とそれに伴う電子メール利用者数の増大を考慮し、携帯電話における文字列処理に着目した。一般的な携帯電話は、その大きさに制約があり、大きなキーボード、多数のキーを備えることができない。そこで、この少ないキーを用いて文字列を入力するために、著者は文字列の情報を縮退させることを考えた。端末へ与える情報を減少させることにより、携帯電話上の少ないキーからでも迅速な入力が可能となる。そして、この入力の際に縮退した情報を復元することを研究の目的としている。

携帯電話はその大きさの制約から装備可能なキーの数が限られるが、最低でも0~9,#,*の12個のキーを装備している。この12キーを用いて、例えば日本語を入力する場合、日本語には約50個のかな文字が存在するので、1つのキーに複数の文字を割り当てる必要がある。一般的には、あ行、か行などの一行が1つのキーに割り当てられている。この割り当てに従い、かな1文字の入力を1打で行う。このようにして、文字列の情報を縮退させることにより迅速な入力を可能としている。しかしながら、入力された数字1文字は意図したかな文字以外にも、それと同一行の複数のかな文字にも対応することになり、結果として入力数字列は多数の日本語文に対応し、曖昧性が非常に増大している。この縮退した文字列情報を復元することが本研究の目的である。

これを実現するために、提案手法においては実例から表層表現が一致する部分を手掛かりとして、対応関係を変換ルールとして獲得する。さらに、獲得された変換ルール同士から再帰的に変換ルールを獲得することにより、抽象度の異なるさまざまな変換ルールを効率よく獲得可能である。このように表層表現に基づき、対応関係を変換ルールとして随時獲得することが提案手法の基本であり、これを帰納的学習と呼んでいる。このようにして獲得された変換ルールを用いて次回からの変換を行うので、提案手法においては入力

データの増加に伴い、次第にその変換精度を向上させることができる。また、変換の際に同一箇所において変換候補が競合した場合には、各変換ルールが持つゆう度を評価し、その時点で最適な変換候補を決定する。このゆう度評価には、隣接する文字列とのつながりを考慮した変換を行うために、文字 n-gram により獲得される統計的な値が加味されている。このように、提案手法では学習機能と統計情報を利用して文字列の曖昧さを解消し、縮退した文字列情報の復元を行っている。

提案手法の持つ情報復元能力の高さを実証するために、著者は取り扱う対象を徐々に高度化し、それに耐えうる応用システムを工学的に実現し、その性能評価を行った。最初に取り扱った対象は、携帯端末向けの日本語入力手法である。この手法における入力データは、かな文字列の母音情報が縮退した数字列である。これを漢字かな混じり文である日本語文に正しく変換することがこの手法の目的である。入力数字列と日本語文とは、表層表現は一部異なるが記述体系は同じである。すなわち、入力数字列と日本語文においては文字列の対応関係を左から順に決定可能なので、表層表現が一致する部分を手掛かりとして、左から順に単語に相当する変換ルールを抽出することができる。このようにして獲得された変換ルールを用いて変換を行うので、提案手法においては入力データの増加に伴い、次第に変換精度を向上させることができる。評価実験の結果、80[%] 以上の変換精度が確認された。

次に著者は、携帯端末向けの中国語入力手法を対象とした。この手法の入力データは、中国語のピンインに対応した数字列である。ピンインとは、中国語の読みに対応し、ローマ字アルファベットにより表現される。入力数字列と中国語文とは、表層表現はまったく異なるが、記述体系は同じである。従って、左から順に1対1の対応を決定することができる。しかしながら、日本語とは異なり、表層表現が一致する部分が存在しないため、対応を決定する手掛かりが1組の変換例中には存在しない。そのため、ここで取り扱う対象は日本語文に対して高度化されている。そこで、この手法においては、2組の変換例を使用し、それらの間で一致する文字列を手掛かりとして対応関係を獲得し、統計情報を考慮して変換に利用する。評価実験の結果、80[%] 程度までの変換精度の上昇が確認された。

次に著者は、携帯端末向け機械翻訳手法を対象とした。日英機械翻訳を前提としているので、この手法の入力データは日本語のかな文字列に対応した数字列である。この数字列を英語文に変換する。よって、入力数字列と英語文とは、表層表現も記述体系もまったく異なっており、手掛かりとなる部分が存在しても対応関係を1対1に決定できない。そのため、この対象は上記2つの対象に比べて高度化されている。そこで、この手法においては対応関係は単語列単位で決定する。すなわち、単語の語順を含めて翻訳ルールとして獲得し、翻訳に利用する。評価実験の結果、他手法と比較して10[%] 程度高い翻訳効率が得られ、提案手法の有効性が確認された。このように高度化したデータに対しても提案手法が有効であることが確認された。

著者は論文全体を通じて、研究領域の現状の分析、新規提案内容の記述、有効性の主張、研究領域における位置付けを正確に行ったと判定する。

以上を要約すると、著者は文字列情報を縮退させて入力を行い、その際に失われた情報

を統計情報を用いた帰納的学習により復元する手法を提案し、少数のキーのみを備えた携帯電話などの小型端末において、容易かつ迅速に日本語の入力、中国語の入力、さらに日本語文から英語文への翻訳を行なうことが可能であることを示した。本研究を通じての、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。