

帰納的学習を用いた多言語単語分割手法に関する研究

学位論文内容の要旨

分かち書きをされていない言語、例えば、日本語、中国語、タイなどのような言語の自然言語処理において文を単語に分割することは欠かすことができない処理過程であり、機械翻訳、情報検索、及び音声認識などは全て単語を基本的な単位として行われる。また、インターネット上で様々な言語のテキスト情報が行き交う時代において、自然言語処理技術の開発は多言語に対応できることが望ましいと考えられる。これらの言語処理技術はすべて単語を基本単位として行われているため単語の間にスペースを入れる習慣のない言語ではまず文を単語に分割する必要がある。

単語分割に関する従来の研究は大別して統計的な手法、規則に基づく手法、および統計的な情報と規則を融合する手法が存在する。統計的な手法では文字の **n-gram** 頻度や相互情報量などを利用して単語分割を行う。スパースネスの問題を避けて高精度モデルを構築するために大量のコーパスが必要である。また、規則に基づく手法ではあらかじめ単語辞書、同一文字列は複数の分割可能性があるあいまい分割を処理する規則辞書を用意する。規則に基づく手法の正解率は辞書の大きさと規則の量に依存する。また、辞書の構築、規則の抽出、整理及び更新には膨大な労力を必要とする。さらに、未知語に対処することは困難である。また、規則辞書と統計的な情報を融合する手法では単語を構成する文字列および単語が出現する文脈に関する確率モデルを用いて、文字列の単語らしさを判断し単語の境界を決定する。この方法は単語分割済みコーパスを必要とするので、これらが必ずしも容易に入手できないという問題点がある。

以上述べたような単語分割が自然言語処理においての必要性および従来研究での問題点から本研究では帰納的学習を用いた単語分割手法を提案し多言語処理を目指して研究行ってきた。本手法では帰納的学習を用いて文書中に重複して出現する文字列の共通部分と差異部分を再帰的に多段階に抽出することにより、単語として確実性の高いものを単語候補として抽出する。また、抽出状況により抽出された単語候補を分類し、辞書に登録する。そして、登録された単語候補を単語としての確実性の高い順に用いて文を単語に分割する。本手法においては、表層レベルの字面情報のみから単語分割が行えるので言語に依存せず多言語に対応で

きるという利点がある。さらに、本手法では予め単語辞書、あいまい分割を処理する規則などを用意する必要がない。また、文を単語に分割しながら未知語を推測し辞書を生成する。本論文では言語を一般化して表現するために記号で表すことにより本手法の言語ごとに依存せず多言語に対応できるという特性を示す。

本手法の多言語に対する汎用性を確認するために三つの言語の文書、中国語(3分野 374,837 単語)、日本語(389,230 単語)、韓国語(91,033 単語)の文書を用いて実験を行った。実験の初期状態は辞書が空で始まり、帰納的学習を用いて共通部分、差異部分を抽出することにより未知語を推測し辞書を生成することによって文を単語に分割する。実験の結果は中国語の総平均正分割率 90.6%、日本語の総平均正分割率 87.5%、韓国語の総平均正分割率 80.8%であった。これらの実験結果から本手法の多言語に対する適応性が確認できた。また、中国語の3分野の文書の実験結果からも本手法は異なる分野の文書に対して適応することが確認できた。

本手法の言語ごとの汎用性及び分野ごとの適応性の評価実験は全て同一アルゴリズムを用いて同じ条件で行った。本手法は文書中から共通部分、差異部分を抽出することにより分野ごと、ユーザごとに適応する単語辞書を生成し文を単語に分割するので、どのような分野、ユーザにも適応が可能であるという特徴がある。

本論文は七章からなる。第一章序論では本研究の背景と目的および本論文の構成などについて述べる。第二章帰納的学習を用いた多言語単語分割手法では、帰納的学習を用いて文書中に重複して出現する文字列を抽出することにより未知語を推測し文を単語に分割するアルゴリズムの詳細について述べる。第三章中国語に対する有効性では、中国語の三つの分野の文書を用いた単語分割実験の結果及び考察、異なる分野の文書に対しての本手法の分野ごとの適応性、帰納的学習手法の学習機能の性能評価実験、結果及び考察、未知語を推測する能力についての考察を述べる。第四章日本語に対する有効性では、日本語の単語分割実験の結果、考察、有効性について述べる。第五章韓国語に対する有効性では、韓国語の単語分割実験の結果、考察、有効性について述べる。第六章多言語に対する汎用性と関連手法との比較では、異なる言語に対しての本手法の言語ごとの汎用性、本手法の言語非依存性による多言語に対する有効性を述べ、また、誤分割の原因および校正処理の回数についての考察を述べる。また、本手法の有効性を確認するために行った他手法との比較実験について述べる。最後に第七章結論では、全体の結論を述べるが、ここでは、上記各章の実験結果の考察、提案手法の有効性の確認、今後の課題について述べる。

学位論文審査の要旨

主 査 教 授 枡 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

帰納的学習を用いた多言語単語分割手法に関する研究

日本語,中国語など,単語間の分かち書きをせずに書かれる言語の自然言語処理において,文を単語に分割することは欠かすことができない処理過程であり,機械翻訳,情報検索,及び音声認識など,自然言語処理のさまざまな応用では単語を基本的な単位として処理を行っている.また,インターネット上で様々な言語のテキスト情報が行き交う時代において,多言語に対応した自然言語処理技術の開発が必須となっており,単語分かち書きをしない言語を処理する必要性はさらに増大している.

単語分割に関する従来の研究の主要な流れとして,統計的な手法,単語辞書及び分割規則に基づく手法,統計的な情報と規則を融合する手法が挙げられる.統計的な手法では文字の **n-gram** や相互情報量を利用して単語分割を行う.この手法ではスパースネスの問題を避けて高精度なモデルを構築するために大量のコーパスが必要となる.

これに対して,単語辞書及び分割規則に基づく手法では,あらかじめ単語辞書,複数の分割可能性がある文字列の曖昧分割を処理するための規則辞書を用意する.本手法による分割の正解率は辞書の大きさと規則の量に依存し,辞書の構築,規則の抽出,整理及び更新には膨大な労力を必要とし,さらに,未知語に対処することが困難であるといった問題が存在する.

規則辞書と統計的な情報を融合する手法では単語を構成する文字列および単語が出現する文脈に関する確率モデルを用いて,文字列の単語らしさを判断し単語の境界を決定する.この方法は単語分割済みコーパスを必要とするので,これらが必ずしも容易に入手できないという問題点がある.

以上述べたような従来の研究手法における問題点を解決した,汎用性の高い単語分割手法が必要であるという観点から,著者は本論文において,多言語処理を対象として帰納的学習を用いた単語分割手法を提案した.提案手法では帰納的学習を用いて文書中に重複して出現する文字列の共通部分と差異部分を再帰的に多段階に抽出することにより,未知語を推測し辞書に登録する.そして,登録された単語候補を単語としての確実性の高い順に適用して文を単語に分割する.提案手法においては,表層レベルの字面情報のみから単語分割が行えるので言語に依

存せず多言語に対応できるという利点がある。さらに、本手法では予め単語辞書、曖昧分割を処理する規則などを用意する必要がなく、文を単語に分割しながら未知語を推測し辞書を生成するので、辞書構築、規則抽出、コーパス収集などのことを低減することが可能である。

著者は、提案手法の多言語に対する汎用性及び多分野に対する適応性の評価実験を全て同一の学習アルゴリズムを用い同条件で行った。中国語、日本語、韓国語の3つの言語の文書を用いた実験結果（中国語の総平均正分割率 90.6%，日本語の総平均正分割率 87.5%，韓国語の総平均正分割率 80.8%）により本手法の多言語に対する有効性を示した。また、中国語3分野の文書の実験結果から、本手法の異なる分野の文書に対しての適応性を確認した。

各章の概要は次の通りである。著者は、まず第2章で、帰納的学習を用いた多言語単語分割手法について述べ、本手法において文書中に重複して出現する文字列を単語として再帰的に抽出することにより未知語を推測し文を単語に分割するアルゴリズムの詳細について説明した。第3章では、中国語に対する有効性を確認するために、三つの分野の文書を用いた単語分割実験、実験の結果及び考察、異なる分野の文書に対して本手法の分野ごとの適応性、帰納的学習手法の学習機能の性能評価実験、結果及び考察、未知語を推測する能力について述べた。第4章では、日本語を用いた単語分割実験、実験の結果、考察、日本語に対する有効性について述べた。第5章では、韓国語を用いた単語分割実験、実験の結果、考察、韓国語に対する有効性について述べた。第6章では、多言語に対する汎用性について述べ、提案手法の多言語に対する有効性を主張した。また、誤分割の原因および校正処理の回数についての考察を述べた。第7章では、提案手法の有効性を確認するために行った他手法との比較実験について述べた。第8章では結論を述べ、さらに、今後の課題として帰納的学習を用いた単語に品詞情報を付与する研究について述べた。

以上を要約すると、著者は、帰納的学習を用いた多言語単語分割手法を提案し、予め単語辞書を用意せず各国語の文章データより帰納的学習を用いて単語を獲得し、それらを用いて未知語を推測しながら文を単語に分割するという手法が多言語に対して有効であることを実証した。さらに、提案手法が分野あるいはユーザごとに動的に適応することを示した。また、著者は従来研究との対比により、新規提案の内容及びその有効性を示しており、自然言語処理工学、情報メディア工学の発展に貢献するところが大きい。

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。