# 第 1 章

# 序論

## 1.1 研究の背景

計算機を用いた自然言語処理研究の目的の一つは、計算機と人間との間のマン・マシン・インタフェイスを容易にし、計算機の機能を最大限に引き出すことにある。これまでの研究の蓄積の成果として、ワードプロセッサやパーソナルコンピュータは低価格化が進み、広く一般に普及することとなった。それには習熟に時間がかかるキーボード操作を介してのデータ入力が必要となる。近時、日本語でも音声入力装置が実用化され、家庭やオフィスにおける文書の入力・作成・編集作業は大きな変革期を迎えつつある。これは、キーボードという計算機側から提供される機器の操作に人間が慣れていくのではなく、計算機が人間の声という自然言語に慣れていく(学習していく)、すなわち機械の側から人間の方へ歩みよろうとする試みであり、計算機と人間とのインタフェイスという観点からは大きな前進をとげたことになる。視点を変えてこれを評価すれば、機械の言語理解能力が人間のレベルに近付いた、つまり「機械がより人間らしくなった」と言える。これにより、長年の基礎研究の成果が「音声認識」という形で応用分野における研究にまで達し、計算機の利用価値の可能性を広げたことになる。

最近では、インターネットやWWW上にある膨大な量の文書を検索して必要な文書を探したり、大量の異なる言語で書かれた文書どうしを関連付けて文書に内包されている抽象化された概念を同定したり、比較したりする必要性が高まってきている。特に最近、自然言語処理の分野で注目を浴びている文書要約やキーワード抽出においては、ノイズ(誤りや不必要な情報)を含む大量のデータから必要な情報のみを抽出する際、日本語を計算機で処理することの困難さが指摘されている。べた書きされた日本語文章の文節区切り、形態素の定義とその認定、文書要約やキーワード抽出のために必要とされる形態素の重要度・類似度等、各形態素のもつデータの量と質やアルゴリズムの精密さのために必要な辞書構築の労力は非常に大きなものとなっている。人間が日常会話で用いている単語をカバーするためには約数万、専門用語を含むと約数十万の語彙(すなわち形態素)をもつ辞書が必要であると言われている。このような大規模な辞書を人手を介さずに自動

的に構築できるか、あるいはまた、そのような大規模辞書を用いずに精度の高い文節区切りが行なえれば、精度の高い高速な文節区切りにより、WWWによる検索速度の向上や、より厳密な検索(ノイズの除去)、よりニーズの高い情報への効率的なアクセスが可能になるからである。

このように言語処理の発展に重要な役割を担う文節区切りは、計算機による日本語処理の最初の段階である形態素解析として理論体系化され、ワードプロセッサの各種機能の充実や、かな漢字変換の精度向上という実用上のニーズから、大学や企業で既に１５年以上にわたり研究されてきている。しかし、今までの研究成果は汎用性の問題と知的所有権の問題からそれぞれの大学や企業において単独で用いられてきており、形態素解析の精度の高い辞書を構築するために、それぞれが膨大な時間と人手をかけてきた。また、これらの辞書はそれぞれの研究目的、例えばワードプロセッサのかな漢字変換の精度を上げるため、といった目的に特化して構築されているため、かな漢字変換用に構築された辞書はそのままでは機械翻訳や文書要約等、別の用途には利用できないということも多い。

ワードプロセッサによって作成された文書にはかな漢字変換の際に誤った漢字に変換されてしまった誤変換や漢字に変換されずにひらがなのまま残ってしまった未変換の文字を含む可能性があり、それらが文中に存在するかどうか、また、存在しているとしたらどこからどこまでか、を決定するための有効な手法は未だに提案されていない。これは、例えば、「ワードプロセッサ野」という文字列があった場合、もともとユーザが「ワードプロセッサ分野」と書きたかったのに、「分」が欠落したものか、あるいは「ワードプロセッサの」と書きたかったのに、助詞の「の」が誤って漢字化したものか、あるいはまた、同じ誤変換でも「ワードプロセッサや」と書きたかったのに、助詞の「や」が誤って漢字化したのか、文法的な解析だけでは判断し難いためである。同様に未変換の例としても、「情報が狂い」という文は文法的には文の構造は正しいため、かな漢字変換が誤ったという判断を一意的に行うことは不可能である。しかし実際には、ユーザが「情報学類」と変換することを意図していたが、文書入力時に見落としたため未変換が生じたと判断することは現在の形態素解析手法では困難である。このような誤りを含む文章の形態素解析に統計的な手法を適用することを考えた。統計的な文字列処理は、一般に言語の認識アルゴリズムに用いられ、認識アルゴリズムのトレースをグラフの形で保存してグラフから解析結果を生成する。言語の認識は基本的には非決定性有限オートマトンのシミュレーションととらえることができ、一文字ごとに状態が遷移していく。このため、形態素解析は探索の方法によっては文の長さに依存して、指数関数的に時間がかかると言われるが、ここでは、漢字やかたかなといった特定の字種の直後に続くひらがなの並びに注目することにより、曖昧さなしに文字列の長さに比例した計算時間で文節に区切る方法を考えた。統計的手法は著明な著作家の著作物鑑定や音楽における楽曲の類似性、演奏者の推定などにも用いられており、自然言語処理においても有効な手段であると考えられる。

点字翻訳のための分かち書きについて、従来手法による形態素解析では不十分である理由につ

いて説明する。点字は表音文字体系であり、基本的には従来でいうところの「短単位」の文節区切りを行う必要がある。ここでいう「短単位の文節区切り」とは、「今日情報処理学会研究会に参加した」という文を考えたとき、「今日」「情報」「処理」「学会」「研究会に」「参加した」のように区切る、区切り方である。この例文では従来の形態素解析における「短単位の形態素」の単位が分かち書きにも有効に働く。しかし、「今日株式会社筑波で会議だった」というような文の場合には、「今日」「株式会社」「筑波で」「会議だった」というように区切り、「株式」「会社」とは、区切らない。これは、「株式会社」の読みが、「連濁」をおこして「かぶしきがいしゃ」と読まれるためであり、「連濁をおこす熟語の間では分かち書きを行わない」というのが、点字翻訳の分かち書きの規則である。このような分かち書きの例として、「心当り」、「心覚え」、「心憎い」は「心」「当り」、「心」「覚え」、「心」「憎い」のように分けるが、「心尽くし」、「心積もり」などは「連濁」をおこしているので分けない、というような例があげられる。分かち書きはまた、語の「泊数」にも関連しており、「アウトサイダー」、「アウトプット」、「アウトライン」等は「アウト」「サイダー」、「アウト」「プット」、「アウト」「ライン」のように区切るが、一方、「アウトロー」、「ツーアウト」などは区切らずに続けて書く。また、「ハッピーニューイヤー」などは、「ハッピー」「ニューイヤー」のように区切る。「複合名詞では、3泊以上の自立可能な意味の成分が2つ以上あれば、その境目で区切り、2泊以下の意味の成分は、そのどちらかに続けることを原則とする」という分かち書きの規則のためである。このように区切るためには「語の読み情報」、しかも、複数の語が連続して1つの語を構成している場合には、いくつの語から構成されているか、構成する際に、連濁をおこしているかいないか、それぞれの読みの「泊数」がいくつか、というような情報をもつ必要があり、従来の形態素解析では点字翻訳のための分かち書きに十分な情報が得られない。最近では、**knowledge-poor approach**といって、膨大な辞書や知識をもたずに、発見的で効果的な、少量の知識を基に、代名詞の照応関係を決定する方法が効果をあげているとの報告がある。このような手法は、従来の「知識を沢山持たせれば持たせるほど、より正確な解析が可能な」手法と対立するものではなく、今後の計算機の能力が拡大するに従って解析可能性が向上するであろう各種解析手法と組み合わせることにより、さらに効率良く問題を解くことが可能となる。ここでは、前述のように区切り方の特殊な「点字翻訳の分かち書き」のために、いくつかの特定の文字と見出し語のみの小さなテーブルを用意することで、精度の高い分かち書きを行う手法について検討した。

前述の通り文節区切りは現時点においても未だ発展途上にあり、実用性・経済性・簡便性をさらに高めるには新たな観点からのアプローチが期待されている。その一方で、ハードウェアの急速な発達とソフトウェアの効率化は計算機で大量のデータを高速に処理することを可能とし、自然言語データの蓄積、検索、統計調査等を容易なものとする段階に至っている。ここではこうした計算機が得意とする大容量のデータからの特徴抽出、あるいは統計量抽出といった計算機の機能をさらに活用することにより、膨大な人手の投入を大幅に削減する自然言語処理のためのいく

つかの簡便な文節区切り手法を提案する。

## 1.2 研究の目的

日本語は英語や他の欧米諸国語と異なり、単語と単語の間に空白がない。この、単語と単語の間に空白がないことが、計算機を用いて日本語を処理する際の大きな制約となっている。このために欧米諸言語では既に実用化されている商用機械翻訳のような応用研究も、日本ではまだ実験段階の域を出ていない。単語と単語を切り分けることを文節区切りといい、本研究の目的は従来の研究で用いられているような膨大な辞書を用いず、形態素解析も行わずに文節を区切る簡便な手法を提案することである。本論文で提案する文節区切り手法はすべて計算機において機械的に取り扱い可能で、人間の介在を必要としない手法である。この手法は、従来のように、その分野における人間のエキスパートが発見的にルールや辞書を構築していく手法に比べると、人手を介さず、機械的に処理されるために速度も速く、かな漢字変換の誤変換や未変換にも柔軟に対応可能であり、文節区切りの精度も実用的なレベルに達している。また、従来手法に比べると辞書も小規模でしかも辞書引きにかかる時間も短くてすむという利点も備えている。本論文ではさらに、これらの文節区切り手法と機械的に構築した辞書を実際に応用システム上に実現し、それぞれの手法の有効性を確認した。

現在、日本語を計算機に入力する際の最も一般的な方法は、キーボードを介してローマ字を入力し、それをかなに変換し、次にそのかなに対応する漢字に変換するという方法である。このとき同音異義語が存在すると、ユーザは複数の候補の中から意図する特定の漢字を選択するという手順をふまなければならない。その際、ユーザの知識の不足や不注意により、ローマ字、かな、漢字の各段階で入力誤り、変換誤りを含む可能性をもっている。ユーザの、変換操作のタイミングを含む全ての入力を記憶し、これを利用することも考えられるが、文節あるいは短い文単位のかな漢字変換による入力方法はユーザごとに異なり、必ずしも均質な文節区切りが得られて解析に利用できるとは限らない。このため、計算機による日本語処理では、通常、最初から「べた書きされた機械可読文書」を処理対象とし、それ以上の文書情報はもたないものとして解析を行う。

英文ワードプロセッサでは既に実現されているスペルチェック機能も、日本語入力方式の問題が複雑であるために日本語ワードプロセッサでは未だに実現されていない。これは日本語の文章の校正のためには一度入力された文章に対して再度解析を行なう必要があり、それにはかな漢字変換用とは異なる形態素解析技術が必要となるためである。かな漢字変換のためには、入力されたかな文字列にどのような自立語列と付属語列が対応するか、文法的にチェック可能である。しかし、ワードプロセッサで作成された文書に頻繁に出現する漢字の誤変換については、形態素解析のレベルでは発見が難しく、統語解析を行い文の意味が正しく認定されてからでないと誤りかどうか判断し難い。また、脱字のような誤りを含む未知語については、まず、どこからが未知語であ

るかの問題とどこまでが未知語であるかの問題から、未知語の発見とその認定は非常に困難である。辞書に載っていない語を発見した場合、通常の形態素解析はその語の品詞も接続情報ももたずに解析を進めなければならず、解析結果は曖昧となり解析の精度が低下する。そこで、辞書のカバーする語彙に拘らず、ノイズを含む文書に対して高い精度で文節区切りを行うことができれば、その後の解析処理への影響も少なく、辞書のメンテナンス等にも負荷がかからなくて済むと考えた。本論文では、形態素解析を行なわず、膨大な辞書も用意せずに、従来手法より簡便な手法を用いて日本語を文節（あるいは用途に応じた単位）に区切り、文節単位でのさまざまな言語処理を行なうための手法について考察すること、およびその有用性を検証することを目的とする。

## 1.3 本研究の工学的意義

近年のWWWの技術標準に基づくネットワーク利用の拡大にともない、伝統的な自然言語処理の手法と、誤りを含む可能性のある大量のテキストに対して頑健に働く自然言語処理の手法とを併用する高速な文書処理技術の実用化が望まれている。

このような要求に応えるため、自然言語処理の研究の流れを概括すれば、

1. 形態素解析を行い、精緻な形態素情報を基に各種応用処理を行う方法

2. 形態素解析を行わずに、ヒューリスティックを用いて応用処理を行う方法

との2つに大別される。1番目のアプローチは、ハードウェアの発展を前提として辞書情報の集積、アルゴリズムの高度化を図ることにより、大量テキストに対する頑健性向上を実現しようとするものである。この手法は日本語の「意味的特徴」、すなわち形態素がどのような形態で連結・組み合わされているかに注目し、言語学上の法則を基にした枠組みで言語処理を行おうとするものである。このアプローチにおいての頑健性の向上策としては、文書中に出現する未知語（誤り）に対し、左からの最短文字列を未登録語として処理を進める方法や、字種情報を用いて文字を読み飛ばして処理を進める方法が知られている。この方法では、未知語の存在によりその前後の形態素解析に失敗する等、未知語の影響が局所的に留まらない可能性がある。

一方、2番目のアプローチではカタカナ列、アルファベット列、ひらがな列、一文字の漢字をすべて同等に取り扱い、一文字ごとに自立語辞書を検索して文節区切りとなる可能性の大小を計算し、不必要な辞書検索を避けるために、ヒューリスティックを用いる方法が知られている。形態素解析を行わない手法では単語の前後の接続関係をチェックしないことから、一般に、処理効率は形態素解析を行う手法より優れているが精度においてその改善が望まれていた。

本稿における主要な研究は、形態素解析に依存しないという点でこのうちの2番目の潮流に属しており、その優位性である高い処理効率を担保しつつ、いかに精度を高めるかを課題としている。そしてその具体的な成果としては

1. 大量の生テキストを高速に処理すること、

2. 特定の分野、スタイルに依存しない非制限テキストを対象とすること、

を目指すものである。日本語処理の難しさ汎用性の低さの原因の1つが文節区切りの煩雑さにあるとの認識から、上記の課題を達成するための出発点として、まずは日本語を文節に区切る手法とその応用システムに着目した。

本研究に関し、日本語処理の潮流という観点から特に留意されるべき点の1つ目は、日本語がもつ漢字かな混じり文の特徴的抽象性に着目し、日本語を漢字やカタカナ、ひらがなの組み合わされ方に注目したという点である。最近の認知科学的研究の結果によれば、人間が文書言語処理を行なう際に、かならずしも論理的・意味的側面からのみ処理しているわけではなく、視覚的処理と意味的処理を同時に、あるいは、両者を組み合わせて処理していることが明らかになっている。とりわけ日本語は他の言語に比べて音の種類（音素）が非常に少ない言語であり、単語を短くすると音の種類と組合せの自由度の少なさから同音語が多くなるため、視覚的処理が重要である。日本語は同音語の多さにより生じる多義性や曖昧性を文字種を変えることによって相当程度解消していると考えることができる。本稿で提案する手法は、人間の文書上の言語処理において重要なファクターとなっている文字表記という視覚情報を計算機に導入しようというものである。このアプローチに基づく研究は発表当時にも例がなく、その後も新しい研究成果は発表されていない。

言語処理の潮流という観点から特に注目すべき第2の点は、日本語の言語的特徴を統計的手法を用いて定量化・特定化しようとした点にある。本研究で提言する手法をさらに拡大・発展させて日本語処理プログラムに組み込めば、日本語独自の言語処理の困難さは軽減されることになる。換言すれば、欧米諸言語と同等レベルの応用範囲を得られるということであり、インターネットをはじめとするコンピュータのネットワークに基づくコミュニケーションを促進することが期待される。本研究では、日本語の文書における視覚情報の活用、日本語における言語特性の統計処理による定量化により、これまでの形態素解析と同等程度の精度の実現と処理効率の向上を示した。また、応用分野として言語処理技術のうち、文書校正機能の提言、点字用分かち書き支援手法の提案を行い、本手法の今後の可能性について考察した。

## 1.4 研究の概要

前述の言語処理上の課題解決に向けた試みとして、(1)漢字とかなの組み合わせ統計情報を用いた日本語文文節区切り方式、(2)「構造化文書」という概念の導入と、それを用い、ワードプロセッサによって作成された文書に出現する誤りを発見する文書校正支援方式、(3)統計情報に基づく品詞制限を利用した形態素解析における曖昧さ解消方式、(4)字種情報と特定の助詞をキーとし、ルールベースを用いた点字翻訳のための分かち書き方式、といった言語処理を行うことに

ついて考えた。また、実際にそれぞれの処理方式について実験的なシステムを構築してその有用性について検討したので、その結果について述べる。

以下で各章における研究の概要を述べる。

第2章では、計算機による日本語処理の諸問題点について考察し、簡便な日本語処理技術の必要性について論述する。

第3章では、辞書も文法情報も用いずに、漢字とその後に続くひらがなの統計情報を用いて文章を文節に区切る手法について提案する。第3．2節では、取り扱う文書のドメインを限定せず、しかも辞書も用いずに従来程度の文節区切り精度を維持する手法のアイディアについて述べる。従来の形態素解析では単語の認定と単語どうしの接続関係のチェックのために膨大な辞書が必要とされるが、本手法では、それらを利用せずに文節に区切ることを試みる。第3．3節ではＪＩＣＳＴ文献抄録および読売新聞1カ月分のデータから文字の組合せの出現頻度統計をとり、日本語を文節単位に分割する際に有効なひらがな文字列のパターンを解析した。第3．4節では第3．3節の結果から、特定の2～3文字に注目するだけで精度良く日本語を文節に区切ることが可能であることを示す。第3．5節では第3．4節で示した方法を改良し、さらに文節区切りの精度を向上させる方法について述べる。第3．6節では本手法の評価とさらに残されている問題点について考察した。

第4章では第3章で提案した文節区切り手法を応用して、構造化文書という文書構造を提案し、構造化された文書構造の上の誤りを各レベルで指摘するルールをＰＲＯＬＯＧで記述し、実験システムの上で実現した。第4．2節では「ワードプロセッサによって作成された技術文書」に出現する誤り調査を行った結果について報告する。第4．3節では第4．2節で行った誤り調査に基づいてワードプロセッサによって作成された文書の特徴を調査・分類するとともに、それらの誤りを校正するためには従来のような形態素解析ではなく、独自の手法が必要であることを説明する。第4．4節では、構造化された文書構造に変換された文書の上では文書と校正ルールを同様の形式で表現可能であるため、構造化文書から多種のユーザインタフェイスが提供できることを実験システムにより示す。ここでは、文書が構造化されていることにより、「1単語ごと」あるいは、「1文節ごと」、「1文ごと」、「1段落ごと」といったような単位での文書処理が可能となる。第4．5節ではワードプロセッサによって作成された文書の誤りを指摘する校正知識について論じる。校正知識は上述した実験システム上の2つのユーザインタフェイス用に開発され、インタフェイスに応じて使い分けられる。第4．6節では第4．5節の校正知識の開発環境について説明する。第4．7節では文書校正支援の拡張機能の1つとして、キーワード抽出機能について考え、キーワードの抽出方法と実際の実験システム上での動作、推定される利用方法について述べる。第4．8節では提案した手法を実現した実験システムの機能評価を行い、市販のワードプロセッサの編集機能と比較検討し、校正支援システムに今後必要な機能について考察を行った。

第5章では、統計的な情報を用いて複数の品詞をもつ形態素の品詞を制限し、従来の形態素解析で問題となっている解析結果の曖昧性を減らす手法を提案する。第5．2節ではまず形態素解析における文節のうち、曖昧さを含む語句を分類した。第5．3節では第5．2節で取り出された曖昧な語や句の曖昧さを解消する方式について提案する。また、第5．4節では1つの語が多数の品詞をもつ、いわゆる多品詞語の問題をとりあげ、語の出現状況に応じて品詞を制限して形態素解析の曖昧さを減らすことを試みる。第5．5節では上記5．3節、第5．4節でとりあげた問題について実際に曖昧さがどの程度解消できるか、実験を行って評価した。第5．6節では形容詞における多義の解消のための方式を提案し、それを評価する。

第6章では、既存の辞書から自動的に辞書を構築して、字種と特定の助詞に注目して日本語の文章を分かち書きする方法について述べる。上記第3章で述べた手法では、従来手法で用いるような辞書は必要でない一方、大量の実験文書にあたって統計情報を得ることが必須である。このような手法は大量の文書が存在する場合は有効であるが、入手不可能で統計情報を抽出できない可能性も存在する。第6．2節では、点字翻訳における分かち書きの問題を考察し、従来の形態素解析では分かち書きに不十分な理由について述べる。第6．3節ではこの問題点を解決するため、小規模の見出し語のみからなるテーブルを用い、接続情報を必要とせず、字種の情報といくつかの特定の助詞に着目して点字翻訳のための分かち書きを行う手法を提案する。この手法では、使用するテーブルを既存の辞書から自動的に構築可能である上、テーブルの大きさを必要最小限にとどめることが可能である。第6．4節ではこの手法の有効性を確認するため、実験システムを構築したので、それについて詳述する。第6．5節では第6．4節で構築した実験システムで実際の分かち書きを行った結果について評価する。

第7章は結論であり、本研究で得られた成果を総括し今後の課題について述べる。