日本語処理における頑健な文節区切り手法とその応用

第 2 章

計算機による日本語処理の諸問題

2.1 はじめに

「計算機による言語処理」といった場合、情報科学あるいは計算機科学の分野では、プログラミング言語、形式言語、そして自然言語という、3種類の「言語」が研究の対象となり得る。本章では、これらのうち、人間が日常用いている「言葉」すなわち「自然言語」を対象として、計算機による言語処理の問題点について述べる。

2.2 日本語形態素解析の諸問題

ここでは日本語における形態素の定義、各種形態素解析手法とその問題点、新たな形態素解析手法の必要性について論じる。

2.2.1 形態素解析の定義

形態素とは一般に、意味の単位を担う最小の単位であり、日本語の形態素解析は、以下の2つの目的のために行われる。

1. 形態素の区切りを認定すること。

2. 形態素の品詞を認定すること。

欧米諸言語では単語は空白で区切られており、日本語処理における形態素解析の1番目に行われる、形態素の区切り処理は必要でない。単語の語形変化を伴う場合でも、規則的な変形が多く、元の単語の形にもどすことは比較的容易である。たとえば英語の場合、"cars" という単語を考えたとき、「車」を意味する"car" と、それが複数であることを示す "-s" という2つの形態素から構成されることは明白であり、この状態から辞書をひいて必要な情報を得ることが可能である。しかし、日本語の場合、語と語の間に空白がないこと、また、漢字と漢字の組合せから容易に新しい

単語を創造し得ることから、語の定義は非常に曖昧である。たとえば、「電子計算機」は「電子」「計算機」の2語からなる、とも考えられるし、「電子」「計算」「機」の3語からなる、とも考えられる。このように、「語」の定義が明白でないという理由により、日本語では「形態素」と「語」とは同じような意味で用いられることが多い。

たとえば、「コンピュータは昔は電子計算機と呼ばれていました」という文章に対して形態素解析を行うと、以下のように出力されるのが一般的である。

形態素 (語)	読み	辞書記載項目	品詞
コンピュータ	こんぴゅーた	コンピュータ	普通名詞
は	は	は	副助詞
昔	むかし	昔	普通名詞
は	は	は	副助詞
電子計算機	でんしけいさんき	電子計算機	普通名詞
と	と	と	格助詞
呼ばれて	よばれて	呼ばれる	動詞
い	い	い	動詞性接尾辞
ました	ました	ます	動詞性接尾辞

これ以外にも、例えば、文節の区切りやアクセント情報、活用語の活用型と活用形などを出力する解析プログラムもある。

いずれにせよ、形態素解析はその利用目的に応じて解析方法を選択する必要がある。これは、形態素解析の解析結果を統語解析の入力として利用するのか、音声合成用の入力として利用するのか、あるいは、キーワード抽出に用いるのかといった応用によって、必要となる解析結果が異なるためである。

2.2.2 形態素解析技術

日本語の形態素解析は、計算機による自然言語処理の分野では比較的盛んに研究されている技術であり、この技術を利用した様々なシステムが提案されている [13, 41]。

一般に日本語の形態素の並びは、少数の例外を除いては、ほとんど直前の1形態素によって限定されるといわれている。そこで、従前より、それぞれの形態素にどのような形態素が接続し得るかを、接続表によってモデル化し、このモデルを用いたシステムが数多く報告されている [20, 27, 31, 41, 53, 80, 82, 63]。しかし、接続表を用いて解析を行っても、必ずしも形態素解析結

果は一意に定まらず、形態素の数の組合せ数だけ、曖昧な結果を出力する。そのような曖昧な解析結果に対し、最長一致法、文節数最小法、最小コスト法、最尤法などのヒューリスティクスを用いて曖昧さを除去し、高い精度で正しい解析結果を選択する方法が提案されている。

日本語の形態素解析では、上述した、「形態素の区切り」と「品詞の認定」の2つの段階で、それぞれ誤りを含む可能性をもっている。「区切り」の誤りとして、「東大通り (ひがしおおどおり)」を「東大 (とうだい)」「通り (どおり)」と区切ってしまうような例があげられる。また、「品詞」については、品詞の分類の細かさと数え方に依存した誤りが存在する。上記の例では、「東大通り」を一つの固有名詞として誤りを1つと数える方法と、「東」と「大通り」の2つの普通名詞の認定を誤ったとして数える方法とがあり、辞書の大きさと品詞のもち方によって誤りの数が異なってくる。また、誤りとは言えないが、「大規模計算機設備」が、「大規模計算機」の「設備」か、あるいは「大規模」な「計算機設備」なのかは、語単位の区切り方の問題ではなく、意味のレベルの解析の問題に関わってくる。これらの問題点に関しては未だに統一的な基準がなく、共通な、公開された自然言語処理環境の整備が望まれている。

形態素解析技術の応用としては機械翻訳のための文解析、文書校正支援、キーワード抽出、音声合成のための漢字かな変換、音声認識、 文字認識等々が考えられるが、もっとも成功した応用の1つがワードプロセッシングのためのかな漢字変換である。かな漢字変換では人間が適宜変換命令を与え、かな漢字変換の曖昧性は人間とのインタラクションによって解消される。たとえば、「ここではきものをぬいでください」と入力した場合、「ここで履物を脱いで下さい」なのか、「ここでは着物を脱いで下さい」なのかは、人間が選択・決定する。入力されたかな文字列に対し、形態素の区切りと品詞が与えられることで、かな漢字変換プログラムは容易に漢字かな混じり文を出力することが可能である。人間は同音異義語の選択をするだけで済む。

最近の動向として、解析の目的を情報のグローバリゼーションに対応した機械翻訳等においては、より長い単位で単語をとらえ、その後の構文解析での曖昧さを減らそうという動きが主流になりつつある。こうした流れに沿った研究が、春野 [61]、颯々野 [28]、松本 [72, 73] らによってなされ、これまで以上に実用的な形態素解析システムが成果として報告されている。具体的には、例えば「延長であり」のような句に対し、

形態素 (語)	読み	辞書記載項目	品詞
延長	えんちょう	延長	サ変名詞
で	で	だ	助動詞
あ	あ	ある	動詞
り	り	り	動詞性接尾辞

という解析と、

形態素 (語)	読み	辞書記載項目	品詞
延長	えんちょう	延長	サ変名詞
であり	であり	だ	助動詞

という2つの典型的な形態素解析のうち、より、長い単位で形態素を捉える、第2番目のような解析結果を出力するものである。

形態素を短い単位で捉えるにせよ、長い単位で捉えるにせよ、現行の言語処理プロセスの枠組みにおいては文中の形態素が認定できなければ次のプロセスは不可能であり、形態素解析の精度の高低が次のプロセスの精度を支配すると言っても過言ではない。

近年、インターネットやイントラネット、WWWの急速な普及・拡大に伴い、大量の情報の創出、発信、流通が個人でも簡単に行われるようになってきた。こうした流れの中で、情報の加工、利用のために、文解析や文章要約、キーワード抽出が、迅速に、しかも効果的に行われることが望まれている。この際に、最も大きな問題となるのが、日本語の文書中に存在する未知語の問題である。現在の形態素解析技術では未知語の存在を認定し、それをいかに正しく切り出すかが非常に大きな課題である。

2.2.3 日本語形態素解析の新たなアプローチ

前述した問題点を解消するため、日本語を形態素に分割する新たな手法を提案する。ここでは、形態素が正しく区切られれば、その後は処理目的に応じた情報をもつ大規模辞書を用意することで、容易に形態素の品詞は認定できるし、形態素の認定さえ正しければ、どのような文法による解析も適用可能であると仮定し、意味レベルまでの解析を行うことなく形態素解析を行う手法について考察する。

アプローチの第一は、形態素解析を、接続関係をもたない文節区切りととらえ、統計的な文字列処理として扱う手法である。統計的手法は、音声認識や音声理解システムにおいて広く用いられている [60]。言語モデルの単位としては、単語、統語カテゴリ、音節、音韻、かな・漢字の連鎖モデルなどが用いられており、音声認識や音声理解の認識精度向上に有効であることが知られている。ここでは処理の単位をかな・漢字の文字とし、漢字とひらがなの組合せ統計をとり、その統計値から文節に区切るか区切らないかを決定する。この手法ではひらがな列のみを対象とし、漢字については個々の漢字としての意味情報をもたず、字種のみの情報を用いた結果、文節区切りの精度は約９７％であった。

アプローチの第2番目は形態素解析の解析結果の曖昧さを統計的な手法により減少させるものである。前節にも述べたように形態素解析における曖昧さはそのままその後の解析に影響を与える。具体的には多品詞語と呼ばれる、1つで複数の品詞をもつ語や、特定の語の並び(前節における「である」のような例)があたかも1つの語のように解析され得る句が文中に存在すると解析結果は曖昧性をもち、その後の解析結果における妥当性が低下する。例えば、得られた形態素が多品詞の場合、「の」には、助詞の「の」、名詞の「野」が考えられる。「は」では、副助詞の「は」、名詞の「歯」、「波」、「派」、「葉」、「羽」、「刃」など多くの候補がある。形態素解析処理では、これらの順列組み合わせの数だけ解析の可能性があるため、文が長くなればなるほど、解析結果の数は膨大になる。本論文で提案する手法は、形態素解析の段階でこの解析曖昧さを減らし、その後の解析結果の精度を向上させようとするものである。ここでは形態素解析の曖昧性を増やす原因となる特定の語と句に注目し、これらの複数存在する解析結果の曖昧さを統計的な手法を用いて減らそうとするものである。この結果、最近の形態素解析の動向である、より長い単位での形態素の認定が、形態素解析はもとより、構文解析における曖昧さの解消に効果的であることを実験により確認した。また、多品詞語についても、前後の語の並びの条件を設定することによりいくつかの場合分けを行って品詞の曖昧さの数を減らせることを確認した。この曖昧さ解消方式は汎用的にも有効であると考える。

第3のアプローチとしては、文法情報を含む大規模な辞書の代わりに見出し語のみからなる小規模なテーブルを用いることにより、辞書構築の手間と辞書引きにかかる時間を削減し、かつ、文節区切りのルールを知識ベース化してアルゴリズムから独立させることによって分かち書きに適応した文節区切りを得る方法について述べる。計算機を用いた点字翻訳のための漢字かな変換手法[39]や漢字文字列のかな付与の方法[71]については、従来より研究が行われてきている。分かち書きについては、点字翻訳のための分かち書きとしてではないが、上述した形態素解析の研究から容易に分かち書き可能であるように考えられる。しかし、これらの形態素解析技術は「点字翻訳」を目的とする分かち書きのためには、

1. 現在用いられている一般の形態素の単位が、点字のための分かち書きには大きすぎること、
2. 他の自然言語の応用問題(例えば機械翻訳)のように、後からの処理がフィードバックしてきて形態素解析における誤りを修復可能でないこと、
3. 点字翻訳のための分かち書きでは連濁や複合語化によって区切り方が変わるにも拘らず、従来の形態素解析ではそのような複合情報を出力しないこと、

といった点から、点字翻訳のための分かち書きには不十分であることが判明した。ここでは点字翻訳のための分かち書きに特化した文節区切りルールのルールベースと見出し語テーブルを構築したが、テーブルは点字翻訳を行う分野に応じて専門用語の辞書から見出し語のみを抽出するこ

とで自動的に構築可能であり、従来の文節区切り手法のようにユーザが複雑な辞書情報を入力したりする必要はない。また、このルールベースのルールは簡単なIf-thenルールで表現されているため、変更は容易であり、点字翻訳以外の通常の文節区切りにも容易に適応可能である。

2.3 まとめ

計算機を利用して日本語処理を行うために、従来より、大規模な辞書の整備と発見的手法による文節区切りの検討が行われてきた。大規模な辞書を使っての日本語処理は年々計算機のハードウェアの性能の向上とソフトウェアの進歩を享受して徐々に実用性を得つつあるが、より簡便な方法を用いてかな漢字変換の誤りや、ミスタイプのようなノイズを含む文書を処理する、実用的なアプローチが望まれている。

本論文ではこのような状況を踏まえて計算機の得意とする大容量のデータ処理を行い、従来よりコンパクトなテーブルやルールベース、あるいは統計情報を用いて日本語を文節に区切り、各種応用問題に利用する手法を提案する。