日本語形態素解析とその応用

Japanese-Language Morphological Analysis and Its Applications

山中　紀子　N. Yamanaka
矢島　眞人　M. Yajima
福永　幸弘　Y. Fukunaga

日本語の文章をコンピュータ上で扱う際の基本的な機能である日本語形態素解析を開発し，さまざまな分野に応用している。日本語を扱う応用分野においては，単語と単語の切れ目を見いだすなどの日本語形態素解析がその処理の基礎となり，その結果を用いてすべての処理を行うことになる。

当社では，この日本語形態素解析を，文音声合成システム，日本語ドキュメントリーダ文章処理，校閲チエックシステムなどのシステム商品に応用している。

Japanese-language morphological analysis is a basic component of Japanese-language processing. It is important for applications using Japanese, and its results comprise basic information for such applications.

Among the various Toshiba products in which Japanese-language morphological analysis is a component are text-to-speech systems, Japanese document readers, the automatic text revision guidance system, and others.

1 まえがき

日本語ワードプロセッサの普及やOCR（光学式文字読取り装置）技術の発達により，日本語の文書が電子化されるようになった。それに応じて，日本語を機械的に処理できる日本語処理技術への要望も高まっている。

日本語形態素解析は，日本語を機械的に扱う際の基本となる技術である。漢字かな混じりの日本語で書かれた文章からユーザの役にたつ情報を提供しようとするには，文字コードとしての日本語を扱うだけでは不十分である。日本語形態素解析は，日本語を扱うときの基礎となる情報，つまりその文章を構成する単語の情報を提供するものである。

日本語処理を行う場合には，日本語形態素解析が必要不可欠な要素となる。日本語形態素解析の結果を用いてさまざまな分野に応用することで，ユーザに役にたつ機能を提供することができる。

ここでは，日本語形態素解析の機能概要と，それを利用したさまざまな分野の応用システムを紹介する。

2 日本語形態素解析

2.1 概要

英語などのように，スペースなどではっきりと単語が区切られている言語とは違い，日本語は単語と単語の切れ目が特定の文字で示されているわけではない。日本語形態素解析を行ううえでは，明示的に示されていない単語の切れ目を見つけることが重要となる。

日本語形態素解析は，まず，入力された日本語の文章を正しく単語に分ける必要がある。さらに，分けられた単語にさまざまな情報を加えることで，日本語形態素解析の本来の目的を達成したことになる。

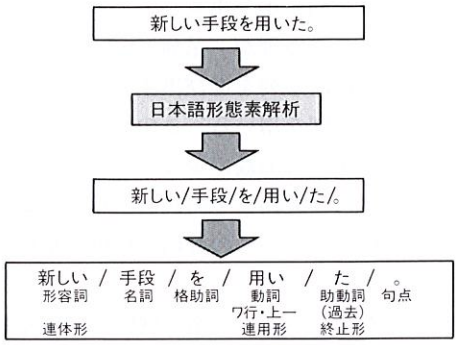

図1．日本語形態素解析の概要　日本語の文章を単語単位に分割し，単語ごとの情報をつける。
Outline of Japanese-language morphological analysis

日本語形態素解析の基本的な概念を図1に示す。ここでは，入力された日本語の文章を単語単位に分割し，それに単語ごとの情報を付加している。

2.2 構成

図2に，日本語形態素解析の構成を示す。

日本語形態素解析は，日本語の辞書を用いて，入力文の各文字から辞書に収められている単語とのマッチングを行い，さらにその単語について活用処理を行う。この活用処理は，例えば動詞や形容詞など，用法によって語尾変化を行う単語のための処理である。

単語は，品詞および活用形ごとに，接続の情報をもって

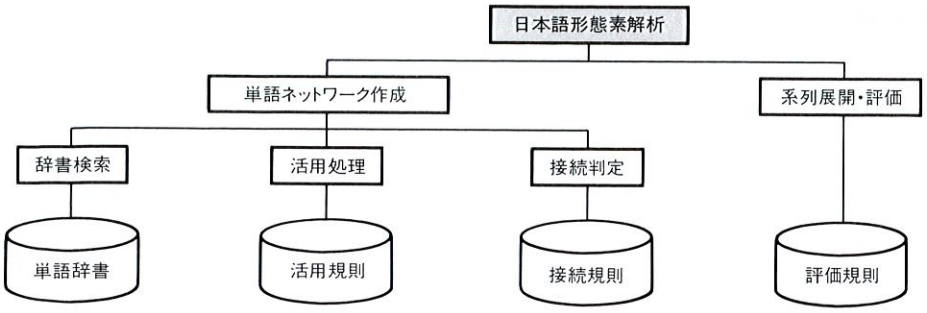

図2．日本語形態素解析の構成　辞書や各種規則を用いて単語ネットワークを作成し，系列を選び出す。

Structure of Japanese-language morphological analysis

いる。隣り合った位置にある単語については，それぞれの単語どうしが接続するかどうかのテーブルを用いて，接続判定を行い，妥当性のある単語だけを候補として残すことになる。

そのようにして得られた単語と単語のつながりを示す単語ネットワークは，図3のようになる。この段階では，形態的に接続可能な単語がすべて候補として現われている。

単語ネットワークの状態では，そのネットワークのどの系列をとっても，形態的には誤りではない。この中から一番もっともらしい系列を選び出すことで，日本語形態素解析の結果とする。この処理が，系列展開・評価である。

系列展開・評価は，系列を選び出すのに，日本語のヒューリスティックな規則を用いる。このヒューリスティックな規則は，例えばより文節数の少ない系列のほうが正しい場合が多い，などの一般的なものから，単語に依存する例外的なものまで種々あり，この評価の精度が形態素解析結果の精度に大きく影響する。

図3の網掛けをした系列は，系列展開・評価が選び出した結果である。以上のように，日本語形態素解析は，入力された文を単語単位に分割する処理を行う。

3 日本語形態素解析の応用

3.1 文音声合成システム

文音声合成システムとは，文字で書かれた文章を，音声で読み上げるためのシステムである。その構成は，大きく分けて二つの部分がある。まず，入力された文章を解析して音韻列を生成する部分，そして生成された音韻列に従って発声させる部分である。文章を解析して音韻列を生成する際に，日本語処理が必要となる。

文音声合成システムにおける日本語解析部分の役割は，入力された日本語の文章に読みを付与し，アクセントパターンを生成することである。ここでは，日本語形態素解析による解析結果を用いて，単語分割されたそれぞれの語の読みとアクセントパターンを，規則にそって生成する（図4）。

このようにして生成された音韻列は，音声合成モジュールにより発声される。

このシステムは“東芝音声システム”として，当社発売のWindows®(注1) 95プレインストールマシンに搭載されている。図5は，“おしゃべりテキスト”の表示画面である。

3.2 日本語ドキュメントリーダ文章処理

日本語ドキュメントリーダは，活字を対象としたOCRである。帳票などの決まったフォーマットのものではなく，日本語の文章を含む印刷物の読取りを行う。当社製品である日本語ドキュメントリーダExpressReader™ Proでは，日本語解析機能を用いた文章処理を搭載している。

OCRでは文字ごとに認識を行う。各文字ごとに類似度などにより，候補文字を決定していく。日本語ドキュメントリーダで日本語の文章を読み取る際には，単に文字単位の

(注1)　Windowsは，Microsoft社の商標。

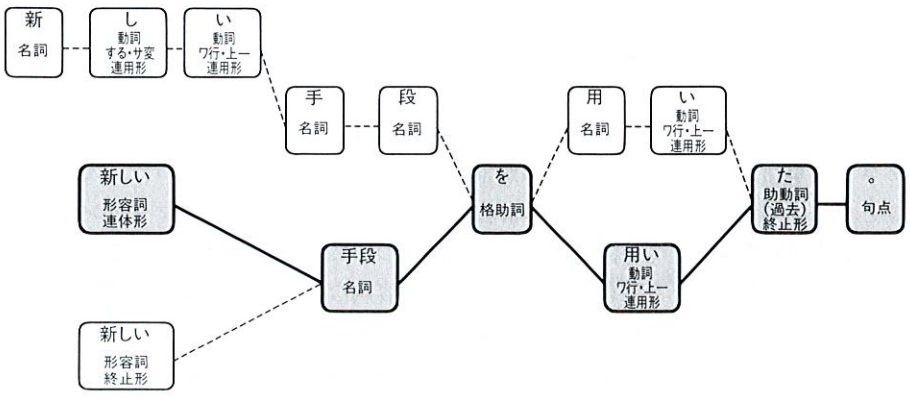

図3．単語ネットワーク　形態的に接続することができる単語で構成される。

Network consisting of words and their information

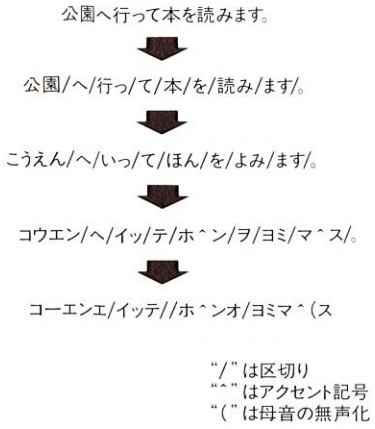

図4．文章の音韻列化　日本語形態素解析の結果を用いて，単語に読みとアクセント情報を与え，音韻列を生成する。
Accent pattern of sentence

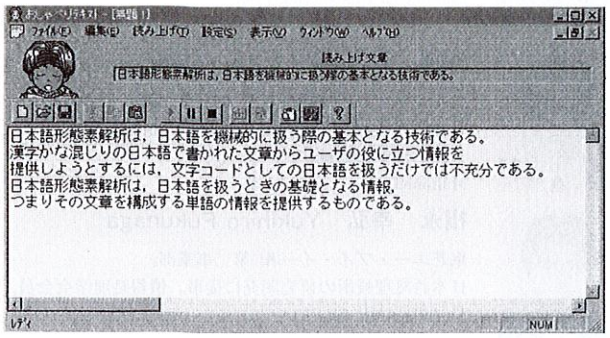

図5．東芝おしゃべりテキスト　当社のWindows®95プレインストール版には，標準で文章を読み上げるシステムが搭載されている。
Example of text-to-speech

認識ではなく，前後の文字を続けて文章として処理することによって，誤認識の可能性を検出できる。

文字認識の結果は，各文字ごとに複数の候補をもっている。日本語形態素解析は，このような複数候補も入力として認めている。日本語ドキュメントリーダではこれを利用して，文字認識した後に日本語形態素解析を用いて候補文字の妥当性のチェックを行っている。

図6は，複数候補をもった入力データを処理した結果である。辞書と各種規則を参照することで，第一候補でなくても正しい候補を選び出すことができる。

文字認識の結果を，候補文字も含めて形態素解析し，第二候補以下に第一候補よりも正しいと思われる解析結果の文字があれば，候補文字を入れ替えて，その文字を第一候補とする。候補文字のどれにも正しいと思われる文字がない，入れ替える候補文字の類似度が低いなどの場合には，警告をつけることでユーザの注意を喚起する。

日本語形態素解析を用いて認識文字候補の入替えを行う

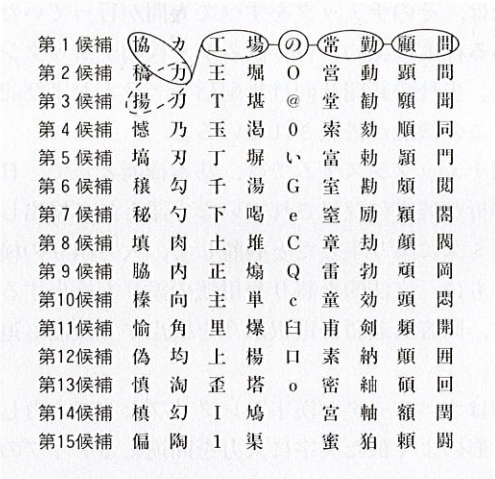

図6．複数候補文字をもつ入力文の解析　複数の文字候補のそれぞれについて，組合せで辞書を検索し，形態素解析を行う。
Analysis of ambiguous input

ことにより，文字単位では発見できなかった認識誤りを発見したり，候補文字を入れ替えて正しい認識候補に置き換えたりできるようになった。

図7は，文章処理をかけた後の日本語ドキュメントリーダの修正画面である。文章処理によって指摘された誤りや候補文字の入替えが起こった部分がわかるようになっている。

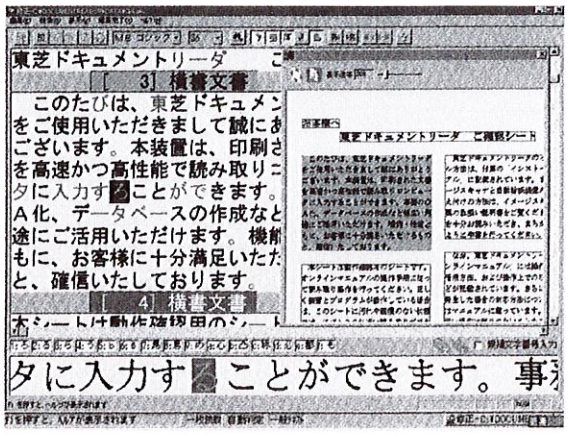

図7．日本語ドキュメントリーダ文章処理　ExpressReader$_{TM}$ Proには，候補文字の中から正しい候補を選び出す機能が搭載されている。
Editing phase of Japanese document reader

3.3　校閲チェックシステム

新聞社や通信社では，毎日大量の記事電文が集配信システムを通過していく。記事電文の入力は，記者もしくはオペレータが行うため，入力ミスや思い違いなどによる間違いが多い。

従来は，そのチェックをすべて人間が行っていた。それを，ある程度機械的に行う試みが，校閲チェックシステムである。当社の新聞社向け集配信システムおよび記者端末には，この機能が搭載されている。

校閲チェックシステムでは，基本機能として，日本語形態素解析が辞書に登録されていない語として検出した部分が入力ミスにより生じたと判断して，入力誤りの検出を行うとともに，文法的な誤りや用法の誤りも検出する。

また，同音異義語の選択誤りを検出する機能も追加されている。

以前はオペレータが漢字テレタイプなどで入力していたので，形のよく似た文字に入力を間違えるタイプの誤りが多かった。そのため，入力誤りは辞書に登録されていない語となる場合が多く，ほとんどが形態的な解析で検出することができた。

最近は記者自身がかな漢字変換を使って入力することが多いので，同音異義語の選択誤りが多くなっている。校閲チェックシステムでは，形態的な誤りだけでなく，同音異義語選択誤りも検出することで，校正担当者の負担を軽くすることができる。

日本語形態素解析はこのシステムに，辞書に未登録であると認められた部分や，日本語として正しくない部分，同音異義語を見つけるための単語の情報などを提供する。

4 あとがき

従来は，一部の専用システムを除いては，個人がそれぞれの環境で利用するには，日本語処理は実用的な能力を発揮できなかった。しかし，パーソナルコンピュータの性能が上がり，大容量メモリで高速に処理ができるようになるにつれ，日本語処理を積極的にユーザインタフェースとして採用するケースが増えている。

今後，日本語処理を用いた応用システムが増えていくに従って，その基本となる情報を提供する日本語形態素解析の役割は大きくなる。

そのためにも日本語形態素解析は，精度および速度の向上を図るとともに，応用システム側の要求に柔軟にこたえられるような機能拡張を実現していく。

山中　紀子　Noriko Yamanaka

マルチメディア技術研究所 開発第六部開発主務。
日本語処理技術の研究開発に従事。情報処理学会会員。
Multimedia Engineering Lab.

矢島　眞人　Makoto Yajima

マルチメディア技術研究所 開発第六部。
日本語処理技術の研究開発に従事。情報処理学会，計測自動制御学会会員。
Multimedia Engineering Lab.

福永　幸弘　Yukihiro Fukunaga

東芝エー・ブイ・イー㈱ 第二事業部。
日本語処理技術の研究開発に従事。情報処理学会会員。
Toshiba AVE Co., Ltd.