情報処理学会第37回(昭和63年後期)全国大会

自立語辞書依存の少ない日本語文の形態素解析

5C-7

諸橋　正幸　　梅田　茂樹

日本アイ・ビー・エム（株）　東京基礎研究所

1．はじめに

日本語の形態素解析は、漢字かな交りで書かれた日本語文書を単語に分割すると同時に、分割された単語や形態素にそれらの品詞や形態素の役割情報を付与することである。この処理は、文献検索システムの自動キーワード抽出のような分割した単語を直接利用するタイプの応用だけでなく、日本語構文解析プログラムの前処理としても重要な働きを果し［1］、その精度の向上は日本語解析に大きく寄与する。

しかしながら、精度を高めるには従来の品詞分類に基づく語や形態素の接続条件による解析だけでは限界があるため、その他のヒューリスティックな規則を利用する必要があり、こうした規則を経験によりどれだけ集めるかが精度に大きく影響する。従来の研究［2,3,4］では、こうした規則はアルゴリズムと一体になってシステム中に組み込まれたり、辞書中に個別の規則として記述されたりしており、試行錯誤による規則の積上げが難しい。

本発表ではこれらヒューリスティックな規則をその性質により５つの範ちゅうに分け、それぞれの規則に対応するアルゴリズムを用意することで、柔軟なシステムを実現した。

2．システムの特徴

システムは次のような特徴を持つ。

1）辞書の語い数に依存しない語分割処理

多くの実用的な語分割システムは単語辞書を頼りに分割を行うが、この場合分割された結果の精度は使用する辞書に依存する。すなわち、辞書にない語（未知語）が入力文に現れた場合、その近傍およびそれ以後の部分について正しい分割ができなくなる。それを避けるために字種の並びの規則を導入し、未知語の推定を行う。

2）語分割の整合性

辞書中に短単位語（意味を持つ最小の長さの単語）だけでなく、長単位語、複合語を登録することは辞書を頼りに分割を行う際に分割の基準を乱すことになる。しかしながら、辞書に複合語を登録することは、非常に短かい語を登録することで起こる誤分割を防ぐ上で不可避なことである。ところが、これにより「神経系統」は２語で「神経衰弱」は１語（たまたま辞書に登録されていた）という事態も起こりうる。これを避けるために、本システムは辞書中に語の分割に関する情報をもつ。

3）文法、経験則のプログラムからの独立と規則の分類

前節で述べたようにアルゴリズムから文法や経験則を分離することは、システムを柔軟にする上で重要であるが、特に、自然言語を対象とするシステムにおいては、規則が１００％正しいという保証はなく、実際、しばしば規則を作り直す必要が生じるため、この発想は有効である。ところが、使用する規則は構文規則のように統一的に整理されたものだけでは記述が困難であるから、その性質によってあらかじめ分類し、各範ちゅうごとに順序よく適用する必要がある。本システムがもつ規則の範ちゅうは、

a.字種による文節分割規則
b.付属語列による文節分割規則
c.自立語辞書と付属語表、および品詞間の接続関係
d.字種の並びによる未知語推定規則
e.品詞の並びによる複合語合成規則

3．処理の概要

システムは、規則の分類に従い次の５つの処理部から構成される（このほかに入出力部がある）。

1）字種による分割処理

字種の変り目が文節の切れ目の候補となること（ひらがなから漢字へ変わるとき、句読点の前後等で語分割が起きる、など）ことはよく知られているが、その性質を適用するのが、ここでの処理である。

また、domain (application)に依存する規則の多くが

Japanese morphological analysis less dependent to a word dictionary

Masayuki Morohashi & Shigeki Umeda

IBM Research, Tokyo Research Laboratory

ここで処理できる（改行を語の切れ目とする、フォーマット用の制御文字を単語としてあつかう、等）。

分割は図１のシフト・レジスタを使って行われる。入力文字列は、各漢字に与えられた字種の定義により字種コードがふられ、シフト・レジスタに順次送られる。レジスタの右側２つの字種の組合せにより呼ばれるマクロ（ユーザー定義可能）がレジスタ内に保存されている文字の分割を決定する。

２）付属語列による分割処理

ひらがな書き付属語列のうち、文節の切れ目を確実に決定できるもののリストを用意し、分割に用いる。現代語においては「を」がその代表例である。

システムには、付属語列のリストのなかからその右側部分がひらがな書き自立語の一部となる可能性のあるものを指摘するユーティリティが用意され、リストの管理を容易にしている。

３）自立語辞書と付属語表による分割処理

自立語辞書と付属語表を用いて最長一致法による語分割が行われる。その際、１）、２）で見つかった分割位置は尊重される。

最長一致の原則は自立語と付属語列について適用される。自立語と付属語の接続検定については、大河内のアルゴリズム［５］を用いるが、検定は左から右の方向へ行われる。この際、多品詞語の品詞が、後に続く付属語によって接続可能なものだけに絞られる（文脈による品詞決定）。

４）未知語推定規則を加味した再分割処理

最長一致法によって分割できなかった（未知語があった）クローズに対し、未知語推定規則を用いて、再度分割を試みる。

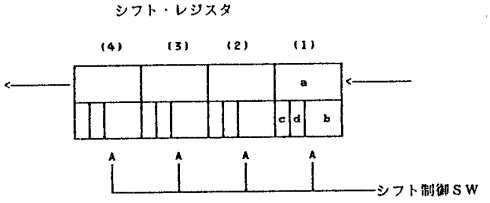
未知語推定は、漢語を想定した字種に基づく規則で現在のところ、以下の語が推定できることを狙ったものである。

a.音読２字漢語とそのうちの１字がひらがな書きされたもの

b.４文字成句、慣用句

c.人名等を想定した３漢字語

処理部３）と異なり、ここで扱う語の候補には未知語推定規則で作り出された不確かな語が含まれる。そこで、分割にあたって最長一致法は用いず、全ての語の候補による可能な分割の中から最適なものを評価式を使って選ぶ方式をとる。評価式は以下に示すもので、文節数最小、未知語長最小を優先する。



a.文字
b.字種コード
c.分割フラッグ（オン：この後で分割せよ）
d.廃棄フラッグ（オン：この文字は分割実行部へ渡さずに捨てよ）

図１．シフト・レジスタの構造

Ｖ＝（最大文節数－当該分割における文節数）
×（句の全文字数－未知語の文字数の合計）

５）複合語合成規則による調整

上記処理部で得られた分割を、まとめあげて複合名詞、複合動詞を認定する処理で、品詞の並びに基づく合成規則が適用される。また、特徴２）で述べた辞書中の分割情報も利用する。

４．実験と評価

新聞記事とＪＩＣＳＴ文献データベースに登録された論文抄録の２種類のデータを用いて、解析精度の検証を行った。新聞記事データは、朝日新聞´８４年６月１日朝刊より任意にサンプリングを行い、各紙面より２～３記事ずつ抜粋したもの（２６，８１４文字、７，２０５文節）であり、ＪＩＣＳＴデータベースは´８３年度電気工学編第１巻より６８０抄録、約１７万文字分任意抽出したものである。

その結果、文節単位による語分割の精度は新聞記事で９９．４%、ＪＩＣＳＴ抄録で９９．１%であった。

参考文献

[1]N.Maruyama,M.Morohashi,et.al.,"A Japanese sentence analyzer",IBM J.of R&D (March,1988)

[2]長尾他「国語辞書の記憶と日本語文の自動分割」情報処理(1978.1)

[3]野村、森「漢字かな変換システムの試作」信学会誌Ｄ(1983.7)

[4]吉村、日高、吉田「文節数最小法を用いたべた書き日本語文の形態素解析」情処論文誌(1983.1)

[5]大河内「分かち書き方式仮名漢字変換のためのバックトラックを必要としない文法解析」情処論文誌(1983.7)