日本語文書処理　21－3
（１９８５．５．８）

かなべた文の逐次単語分割アルゴリズムの一方式

藤田克彦、沼田泰之、山内佐敏
（株）リコー・技術本部

１　はじめに

ワード・プロセッサなどでの日本語入力方法は、最近では使いやすさの点から、「かな漢字変換」方式が多く使われている。

特に、最近は、作業者が文章入力中に文法的な判断を行なう必要が少なく、またそれに伴うキー操作が少なくてすむように、との観点から、入力したい文章の読みをそのままキー・インするだけでよい、「べた書きのかな文字列」を対象としたかな漢字変換方式の研究開発が盛んである。

一般に、文法的処理を組み込んだかな漢字変換システムでは、変換候補の抽出を単語辞書と文法辞書を用いて行なう。

単語辞書：入力されたかな文字列とマッチする単語を見出すための検索対象
文法辞書：単語辞書検索によって得られた単語と前後の単語との文法的接続性を記したテーブル

基本となる処理は、「辞書検索」、「接続検定」であり、この順に行なわれる。ただし、この処理を横型探索的に行なうか、縦型探索的に行なうかなどにより、種々の方式がある。

ここで方式によらず、問題となるのが、これら処理の結果、複数の変換候補が得られる場合である。

複数の候補から、最も適当と考えられる候補を選択する処理が必要である。

現実のかな漢字変換システムでは、この処理を発見的な手法によって行なっている。なかでも、べた書き文を対象とした方式の多くは、文節に着目している。多くのシステムでアルゴリズムの基本的部分に用いられている「文節数最小法」[1]が、その代表例である。

この方式は、かなり高い変換精度を有していることが報告されているが、次のような点に制約がある。

1）基本的に横型探索を行ない、その結果を利用せざるを得ないため、結果を記憶しておくメモリが多く必要になる。
2）方式の利点を生かすためには、ある程度長い入力に対して処理を行なった結果を必要とする。

文節という単位を重視する方式ではなく、入力の先頭から、評価値を利用しながら、順次単語を決定していく、単語単位に処理を行なう方式の一

例としては、すでに「自由入力形式のカナ漢字変換」[2]がある。

また「二文節最長一致法」[3]は両者の中間に位置づけられる方式といえよう。

筆者らは、使いやすいかな漢字変換方式の確立を目的に開発を行なってきた。

専任のオペレータでない、一般の作業者を対象として、入力している最中に、数文字前に入力した部分が次々と変換されて出力されることを目標にし、それを実現するため評価値を用いた縦型探索方式を採用した。

以下、その内容について説明する。

2　本方式の概要

筆者らが採用した方式の概要は、まとめると次のようになる。

1）単語を単位とした、縦型探索方式である。
2）探索経路は、単語の評価値により決定する。
3）評価値の算出には、次の３種の情報を用いる。
　ａ．単語の読みの長さ
　ｂ．単語の出現頻度のランク
　ｃ．単語の品詞と、その直前の単語の品詞との接続の重み
4）探索失敗時には、前に戻って別の経路を探索する（バック・トラック）。

注意：ここでいう単語とは、同じ読みの長さで同じ品詞である単語（同読み、同品詞単語）のグループのことである。

また、単語抽出処理の基本的な流れは、図１のようになる。

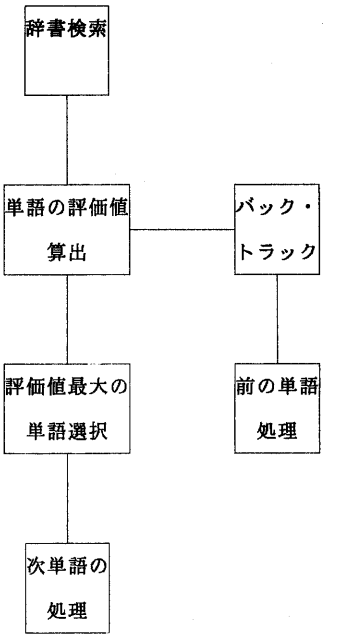

図１　処理の流れ

まず、ある位置からの文字列に対する辞書検索を行なう。次に、得られた単語それぞれに対し、評価値を計算し、評価値最大の単語を変換の結果とみなして、続く文字列の処理に進む。単語が見出されない場合は、前に戻って別の経路を探索する。

以下に、個々の事項について説明する。

3　単語の評価値算出

3．1　評価値の考え方

逐次処理を行なう場合、文節数最小法などと比較して、文字列から得られる情報が少ない。単語辞書、文法辞書などに従来より多くの情報を用意し、変換精度を確保することを目指した。

従来、単語の抽出において有効とされている単語の長さ、単語の頻度は、逐次処理における局所的な単語評価にも利用できる。それらは、単語そ

れぞれが固有にもっている情報である。

それに加えて本方式では、品詞間の接続の重みという情報をも単語の評価に採用した。

これは、次の点を考慮したためである。

1）直感的にも、連体詞の直後には体言が出現しやすい、などの傾向があるので、単語の現れる環境に依存した情報の効果があるのではないかと考えられる。

2）品詞間の接続の重みの設定は、実際の文例を対象とした実験によって行なう必要があるが、これは言い換えれば、現実の文章に適合した接続重みによる、精度の高い評価を可能にするはずである。

3）接続の重みの表現は、接続可否検定用の行列表を文節境界を越えた単語の接続も表現できるように拡張した上で、多段階化することで可能である。

3．2　評価用情報の特徴と数値化

評価に用いる情報は、それぞれ次に示すような特徴を有する。

読みの長さ：単語毎に固定的に定まる

出現頻度　：対象とする資料により絶対値が異なる
相対的な頻度順位はほぼ一定（ただし、専門用語は分野により大きく順位も異なる）

接続の重み：単語のクラス（品詞）同士の関係で決まる。

上の3種の情報のうち、単語の読みの長さは自動的に決まる。

出現頻度については、その実数は調査規模により大きく左右される。そこで実際の適用にあたっては、相対的な頻度のランクを利用することを考え、頻度を対数化して利用することとした。

また、接続の重みは、多くの文章例を対象として、かな漢字変換の実験を繰り返すしながら設定することとした。

その結果、各情報を次のように数値化した。

長さ	l	1から6 辞書中の単語のかな表記の長さ
頻度 ランク	F	0から7 $F=[\log_8 f]+1$ ただし$f=0$は$F=0$ f は国研調査[4]の数値を加工 [ ] はガウス記号
接続 重み	c	0から3 0－－文法的に接続不可 1から3　接続の重み

頻度のランク化は次の効果がある。

1）頻度の多少の揺れを捨象し、単語の読みの長さとほぼ同じオーダーで表現できる。

2）辞書中でのエリアが小さくてすむ。

接続重みを4段階にした理由は次の通りである。

1）重みは実験により設定するものなので、細かすぎると、作業が困難になる。

2）4段階であれば、2ビットに情報を収めることができ、文法辞書の巨大化を最小限に抑えることができる。

なお、このように設定すると、同じ読みで同じ品詞の単語のうち、常に頻度ランクが最高のもの

が最高の評価を得るということになる。つまり、同じ読み、同じ品詞の単語群は評価の際には一つと考えてよい。そして、その単語群の中での優先順位を別の情報として学習し、出力の際にはそれを利用すれば同音語選択とも整合がよい。

３．３　辞書中での情報の持ち方

評価用の情報をどのように保持しているかを説明する。

図２が単語辞書の模式図である。

読みの長さは、プログラム中でカウントするようにしたため、辞書には記述していない。

接続の重みは、接続行列表に記されている。

この行列表の検索は、直前の単語の後方への接続分類（かかりコード）と、当該の単語の直前の単語への接続分類（受けコード）により検索する。これらコードは、品詞コードに対応したテーブルを用意し、そこに記入してある。

品詞分類は従来のものより、やや細かくし、接続重みの設定を有効に生かせるようにしてある。

体言を例にとれば、大きく「名詞」と「さ変名詞」に分類した上で、それぞれを「漢語」と「和語」に分け、「漢語」はさらに、「２字の漢語」、「１字の漢語」などに分類している。これは、例えば、接頭辞や接尾辞は「２字の漢語」には接続するが、「１字の漢語」には接続しないという事実を表現するためである。

逆に、動詞の連用形とそれから転成した名詞（連用名詞）については、局所的には区別がつけにくいため、この二つを合わせた形で接続重みが検索できるよう一つの分類にまとめてある。

さらに、頻出する単語連続も単語とみなし品詞を与えてある。例えば、「について」や「において」などは、直前の単語に対しては格助詞の「に」と同様の振舞をし、直後に対しては接続助詞「て」と同様の振舞をするので、新しい品詞を設け、そこに分類した。このような、従来にない新しい品詞は、１００種以上に及んでいる。これら頻出単語連続の登録は、次の効果を狙ったものである。

読み	品詞	頻度	出力順位	表記
こうしょ	名詞	1	1	高所
こうしょう	名詞	0	1	校章
			2	鉱床
	さ変名詞	3	1	交渉
			3	考証
			2	公称
	形動語幹	1	1	高尚

図２　辞書の模式図

１）辞書検索回数を含め処理量が減少する。

２）登録された部分での誤変換が、ほとんど生じない。

３．４　評価式の設定

これらの情報をどのように組み合わせるかを明らかにするため、汎用機上で実験を行なった。

実験用の文例として、市販の文例集[5]から選んだ２１文例を用いた。

評価式は、長さ、頻度ランク、接続重みの１次式として、次のものを採用した。

評価値Ｖ＝ｐ＊ｌ＋ｑ＊Ｆ＋ｒ＊ｃ

この式は、頻出する単語には短い単語が多く、頻度を対数化すれば、ほぼ同じオーダーで評価に関与するという仮定のもとに設定した。接続重みは、その一般化に対して補正項として働くと考え

た。

実験に並行して、接続重みの設定も行なう必要があったため、ｑは１に固定し、ｐ，ｒも整数に限って実験を行なった。

表１は、その結果の一部（ｒが２の場合）である。これは２文例を対象とした結果である（単語区切り率については後述する）。

表１　評価式と単語区切率

評価式	単語区切率
１＋Ｆ＋２ｃ	９１．２％
２１＋Ｆ＋２ｃ	９８．０
３１＋Ｆ＋２ｃ	９８．７
４１＋Ｆ＋２ｃ	９５．４
３１＋Ｆ＋２ｃ ｃが１／０の場合	９４．７

その結果、厳密な意味での最適な係数を得ることはできなかったが、次の式がほぼ妥当であることがわかった。

評価値Ｖ＝３＊１＋Ｆ＋２＊ｃ

その後は、この式に限定して接続重みの設定に注力した。

４　バック・トラック

本方式では、直前の単語（品詞）に続く単語が見つからない場合、前に戻って解析をやりなおすバック・トラックを採用している。

ただし、無制限に許すと正しい解析結果にまで影響を及ぼす可能性もあり、また逐次結果を出力するという狙いも実現できない。

その制御のために、評価値を利用している。

１）探索経路に沿って、評価値の累積和を算出する。

２）累積和が一定の値を超えたとき、別に定めた一定の値に相当する分の単語を、未確定部分の先頭から確定したものとみなし、バック・トラックが及ばないようにする。同時に累積和からその分の値を引いておく。

３）１）に戻る。

単語の評価値は元来、その経路のもっともらしさを示すものであるから、評価値を用いた制御は有効であると考えられる。

５　本方式の性能

本方式の性能を測定するため、次の２つの数値を算出した。

単語区切率：単語の先頭と末尾が、原文の単語と一致した率

単語変換率：同義異表記を許して原文の単語が得られた率（母数は全単語）

単語区切率は、同音語の選択処理で対応できる割合を示していると考えられる。

２１文例（約５２００単語）における数値は、次の通りであった。

単語区切率　９７．１％

単語変換率　９２．２％

文節指定方式との比較のため、次の数値も測定した。

文節区切率：文節を構成する単語の区切りがすべて正しい文節の率

９６．１％

６　知見と対策

６．１　誤変換と対策

２１文例中での誤変換位置は約８０箇所であった。その約半数が、読みの長い単語が優先されたためである。

とくに、接辞と２字の漢語の間での評価が問題となった。

例　副社長様－－－＞複写調査魔

接頭辞の「副」より、さ変名詞の「複写」が優先された。

これらの多くは、辞書中に接辞の付いた形で単語（上の例では「副社長」）を登録することにより、解決できる。そのためには、読みの長い単語も自由に扱える辞書形態にする必要がある。

ビジネス文に頻出する、「お」と「御」については、評価をある程度優先しておき、それに後続する品詞が限られることを利用したバック・トラックを起動させるようにすることで、問題の大部分が解決できた。

次のようなタイプの誤変換もあった。

例　行使できる－－－＞講師で切る

名詞の「講師」の方が、さ変名詞の「行使」より優先されたため、その後ろで誤変換が発生した。

このタイプの誤りは、先読みをしなければ、解決が難しい。現時点では、先読み処理は行なっていない。

その他の誤りとしては、さ変名詞直後でさ変動詞が優先される、未登録語の影響などがある。

ただし、誤変換が生じたときの影響の範囲は、一般にあまり大きなものではなかった。１回の誤変換の影響範囲は、原文の単語に換算して平均２語であり、５、６単語にわたるような誤りは、ほとんど見られなかった。文節区切率がそれほど低下していないのは、そのためである。

６．２　バック・トラック

バック・トラックは、誤りのすべてを救うわけではないが、半数弱の評価誤りについて効果がある。範囲としては、少なくとも２語前まで遡れることが必要である。本方式ではこれを、評価値を利用した閾値の設定により実現できた。

６．３　品詞の変更

本方式では、文法を直接隣りあう品詞間の接続として、接続行列表で表現している。その場合には、たとえば「お」に後続する動詞連用形の接続性の変更が、表現できない。そこで、そのような場合には、プログラム中で直後の単語への接続の分類（かかりコード）を書き換え、それにより接続行列表を参照するようにした。

７　あとがき

本方式は、性能面では実用に十分耐えうるものと考える。

とくに、逐次変換結果が出力されるという点は他の多くの方式にない特徴であろう。

現在は、ヒューリスティックな処理につきまとう誤変換、とくに単語の区切りの誤りの解決に力を注いでいる。具体的には、誤って変換した場合の情報を学習し、以降の処理では同じ誤りを繰り返さない機能の組み込みである。これは、社内でテスト中であるが、１０文例を対象に学習を２回繰り返した後の単語正変換率として、９７．２％という結果を得ている。

参考文献

1　吉田ほか：日本語文の形態素解析における最長一致法と文節数最小法について
ＮＬ研資料３０－７　１９８２年

2　内田ほか：自由入力形式のカナ漢字変換
ＮＬ研資料２７－３　１９８１年

3　牧野ほか：べた書き文の仮名漢字変換システムとその同音語処理
情処学会論文誌２３－１

4　国立国語研究所報告３７　１９７０年

5　安田賀計編：「企業経営文例全書」
ぎょうせい　１９７８年