日本語文書処理　21－1
（１９８５．５．８）

ベター方式かな漢字変換入力システムの試作

（株）　東芝　総合研究所　　河田　勉　　武田　公人
情報通信システム技術研究所　　斎藤　裕美　　中里　茂美
楠元　達治

１．はじめに

文節分かち書きを必要としない、べた入力のかな漢字変換を日本語ワープロで実現するためには、高速かつ高精度なロジックが要求される。具体的には、専門オペレータの入力速度に充分追従できること、次候補や訂正回数が従来の文節式の時と同等かそれ以下であることなどである。これらを目指して開発したベター(BETTER)入力変換システムについて報告する。

２．目的

べた書き入力によるかな漢字変換方式を、実際の日本語ワードプロセッサに組込むための基礎実験を行った。ここでは実験で判明した問題点と解決方法について報告する。べた書きの入力方法としては、２文節最長法、文節数最少法などが知られている。また、現在複数の文節を区切らずに入力できる機種もあるが、未変換／誤変換、同音異義語の選択など操作性など改善の余地があると考えている。

実用的なワープロを実現するには、以下の事柄を解決する必要がある。

①　文書作成における読みがな列は、文節の単位を意識せず連続して何文字でも入力できること、

②　応答速度が現行のワープロ以上に速いこと、

③　同音異義語の選択が従来と同一で簡単にできること、

④　従来通りの方法と併用でき、同等の性能であること、

⑤　第１位に目的とする同音異義語が出力される率が高いこと、また、不必要な候補をできるだけ排除できることなどである。

３．システムの構成

本システムの動作は図１のように進行する。順次入力される文字列に対し、データ前処理部において自立語辞書の引当てと、解析区間の切出し（仮想ブロック切り）を行う。続いて単文節文法解析に基づく文節系列抽出を前方より多様に試みながら第２の切出しであ

る変換ブロックを決定する。ブロック内には多義の系列構造を含み、以下組合せ評価により、不自然な候補の排除と一位変換率の向上を図る。

4．辞書引きの高速化

応答速度を速くする方法として、文章の入力中に全ての自立語の情報をテーブル形式で準備する。このため、辞書は入力毎に読みを受けとって参照が容易なように、逆引き構造とする。また、読みが等しい単語はグループ化する。逆引き構造とすることで、順次入力される読みに対し、参照範囲が１つの範囲のみで可能となる。図２に例を示す。“・”の箇所は辞書に該当する単語があることを示す。行番号は読みの開始、列番号は終了を示す。じょうほう（１，５）は該当する単語“情報”があることを示している。

5．仮想ブロック切り

本実験では、変換の解析単位は４文節までを対象とする（実験の変換単位抽出アルゴリズムは、４文節が最長となる組み合せから、３文節の終点が最短となる地点を分割点とする）。このため、読みが連続的に入力された場合も、３～４文節入力された時点で変換を開始することが可能となる。この方法で、入力時に連続文字数の制限を設ける必要がなくなる。また、長い文をまとめて変換することが無くなるので変換速度も向上する。

ここで問題となるのは、３～４文節以上となった時点を検出する方法と、ワードプロセッサとして実現したとき、入力文と変換結果をどのように表示するかである。一定の文字数で変換を開始するのは、意味不明の所で切れてしまう。また、句読点などで変換を開始すると非常に長い文を変換する必要が生じる。

入力文字列と単語単位の照合がすでに終了している場合には、自立語がどの辺りにあるかの情報は用いることができる。但し付属語の解析を含めた自立語の認定は時間的に無理であるので、自立語を確実に判断する手段として、格助詞、係助詞の結びつきやひらがな以外の文字情報を利用する。

４文節以上又は確実な変換単位を決定するため、
自立語＋助詞＋自立語＋助詞のパターンマッチをする。この条件が満たされるときは原則として２番目の自立語の前を分割点とする。

ここで読みの長さが１～２の場合は、大部分の読みに自立語が存在するので自立語として扱わない。読みの長さが３以上で、音読み要素２以上を自立語とする。また、音読み要素をまたがる読みも排除する。

“しゅうちゅうてきに”では、

```
          ×    ×  ×  ×  ×    ×  ×
| 集中、周知、週、主、死 | 宇宙、内 | 注 | 的          ↓
  集中 | 的に
```

上記パターンマッチが必ずあるとは限らない。この場合は認定自立語数が４以上となれば、実際の変換で正しくブロック切りが行われるので、仮想ブロック切りでは４以上であれば、次の自立語の前で切断して問題ない。自立語の認定は単独で出現する場合は問題ないが、自立語が連続する場合の認定数は、音読要素の数から次のように求める。

認定数　＝　音読み要素の連続数／２　＋　１

例・すいしんりょくは　→　２

推　進

新　緑

・うんてんめんきょしょう　→　３

運　転

免　許

巨　匠

・こうつうかんきょうせいび　→　３

交　通　環　境　整　備

痛　感　強　制

６．変換ブロック決定

かな文字の長い列に対して自動分かちを行う。この単位は、多様な文節構造に解釈できる部分をなるべくまとめて変換候補に残すために、以下のような最長４文節内の構造解析により決定する。分かちの区間を変換ブロックと呼ぶ。

①　連続４文節で最長点に達する文節系列を探索する。

②　もし途中で入力終端に達することがあれば、入力系列全体をブロックとする。

③　①で得られる系列の中で次の順に文節端の共通点を探し、ブロックとする。

(1)３文節目　(2)２文節目　(3)１文節目

④　③で見つからないときは①で得られる系列の中で最短の３文節目とする。

以上の過程において、単文節解析が繰返し実行され、ブロック内でのあらゆる文節候補が抽出される。

７．　単文節解析

文節を以下のように定義する。

［自立語］(+［付属単語］ +［付属単語］+ …)　接辞語も単独の自立語とし、また形式名詞、補助用言は付属語に扱う。従って、複合語の処理も内部では文節列として扱う。単文節解析部の呼出し時には、自立語始点と外部条件が指示され、以下のように動作する。

ａ．始点が同一の自立語候補の各々に対し、可能な付属語接続を調べ、文節終端条件を満たすものすべてを出力する。
ｂ．直前が「お」「ご」の単語のときは敬語表現も試みる。
ｃ．直前が特殊な接頭「ふ“不”」「む“無”」などのときは、名詞を形容動詞としても処理する。
ｄ．入力の始点側では接尾、連濁語を禁止する。
ｅ．入力の終端側では接頭を禁止する。
ｆ．前の変換ブロックが非かな文字で終わったときはダミーの名詞を想定し、付属語だけの解析も試みる。

８．組合せ評価

変換ブロック内のとり得る文節系列を組立てるに際し、主に次のような棄却や出力優先度評価を行う。変換ブロック内では最大４文節の結合まで扱えるようにしている。

限定規則

a1. 文節数最小の優先に従い、最小構造のものと次点の構造（最小＋ 0.5）だけに限定する。各文節には以下のように点数をつけて数えた。

接辞以外の文節　－ 1.0点
接辞文節　　－ 0.5点

a2. 他の構造中の文節自立語部分を分割する単語の複合を認めない。

研究会　　×研究＋回

a3. １字読みの名詞どうしの複合をしない。これはa1,a2 の規則で除かれていない場合に、辞書未登録語やタイプミスによることが多く、無理に漢字列を作らないためである。

優先度規則

限定規則によって不要な候補データを落とし、次に主として文脈的にみた出現性の優劣を利用して、同音語系列構造の中での一位優先出力を決める。前後関係で順位を優先させるために評価点数を加点するものと、相対順位を下げるために、減点するものがある。以下にその適用例を示す。

付合い　出したのである　○　（○は加点）
付合いだ　したのである
“連用形＋動詞”を優先

増えてきたと　聞く

増えてきたとき　句　△　（△は減点）

１文字自立語を含むときは下げる

行政　改革　案　　○

行政か　威嚇　案

２漢字体言の複合を優先

真の　生物とは　○

芯の　生物とは

“連体詞＋体言”を優先

彼は　知っていると

彼　走っていると　　△

付属語のない体言に体言や接続詞以外の語が付く系列は下げる

９．同音異義語の選択

べた書き入力では目的とする同音異義語を速く表示させることは、優先規則や辞書に知識情報を持たすことで可能である。しかし、組み合せの数を考えると、文章によっては数多くの同音異義語を変換結果として出力せざるをえない。

実験では、同音異義語を組合せる順番と、カーソルの位置情報を利用する方法を用いた。

同音異義語の組み合せは、頻度を使わず、文節の優先点数で行い、頻度は単語単位の同音異義語の中に限定する。一定の規則で組合せる方法が、次候補での予測が可能で操作性が高かった。

例　“こうしょうけんを”に対しては、

①　名詞（交渉、厚相、考証、高尚、高承、……）　×　接尾（権、券、圏……）

②　名詞　×　名詞（剣、件、兼……）

③　接頭（高、広、……）　×　名詞（証券）

④　動詞（請う）　×　名詞（証券）

⑤　名詞（甲、項、……）　×　名詞（名詞）

の組み合せが同音異義語となる。

従来の方法では組合された候補を“次候補”キーで順番に入替えていたが、カーソルを変化させたい位置に合わせて行うことで、目的とする組み合せへ速く到達できる。また、

選択された同音異義語の間にある付属語を除いた組み合せで記憶し、次の変換では第１位に出力する方法で、べた書き入力でも従来通りの選択ができた。

図３に同音異義語の選択の例を示す。

１０学習機能

仮名漢字変換において単語の使用環境を記憶しておき、同じ語を再び使用したときに優先する学習機能は非常に効果が有ることはよく知られている。従来の機能は一語毎に単語を記憶してゆく方式であった。ここでは、べた入力の環境では単語の連続に対しても学習機能が働くことが望ましい。そこで、ベター方式では語の共起関係で記憶する方式とした。ず　に学習機能の例を示す。この学習機能を単語が連続して現れた場合には隣接する単語のペアで記憶する。出現する単語の順序が入れ替っても学習効果が発揮できるようにした。図４にその例を示す。

１１おわりに

以上べたかな漢字変換の概略について述べた。図５に実際の変換例を示す。文節入力に比べてはるかに使い易いこの方式をベター(BETTER)方式と呼んだ。現在、多種の例文を入力して評価実験を行っている。また、同音語選択操作や誤変換時の対処法などエディタ部まで含めたシステムの総合調整を行っている。

［参考文献］

1)斎藤他：かな漢字変換方式について、情報処理学会第25回全国大会論文集、pp1125(198

2)牧野他：べた書き文の分かち書きとかな漢字変換－二文節最長一致法による分かち書き－、情報処理学会論文誌、Vol.20.No.4(1979)

3)吉村他：文節数最小法を用いたべた書き日本語文の形態素解析、情報処理学会論文誌、Vol.24.No.1(1983)

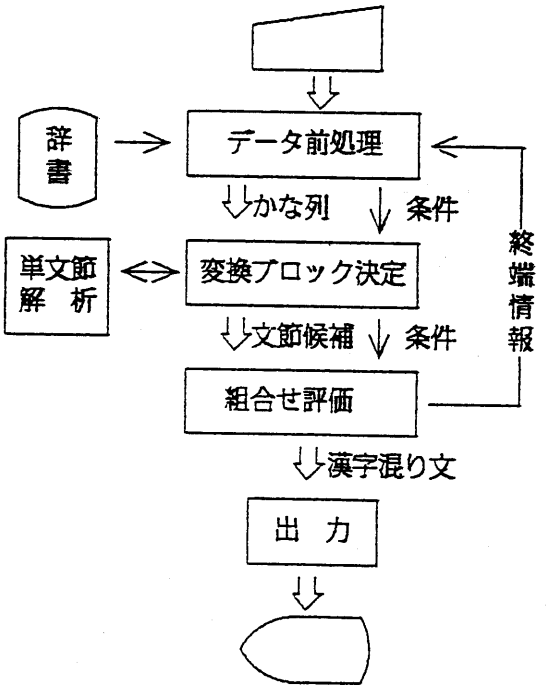

図1 ベター方式かな漢字変換の流れ

辞書先取りテーブル

	1 じ	2 ょ	3 う	4 ほ	5 う	6 し	7 ょ	8 り	9 し	10 ゃ	11 か	12 い	13 が	14 げ	15 ん	16 じ	17 つ	18 の	
1 じ	・	・	・	・	・	×	×	×	×	×	×	×	–	–	–	–	–	–	じ 字、序、錠、調歩、情報
2 ょ	×	・	×	×	×	×	×	×	×	×	×	×	×	–	–	–	–	–	ょ ×
3 う		×	・	×	×	×	×	×	×	×	×	×	×	×	–	–	–	–	う 右
4 ほ			×	・	・	・	×	×	×	×	×	×	×	×	×	–	–	–	ほ 帆、法、奉仕
5 う				×	・	・	×	×	×	×	×	×	×	×	×	×	–	–	う 右、牛
6 し					×	・	・	・	×	×	×	×	×	×	×	×	×	–	し 詩、幽、処理
7 ょ						×	×	×	×	×	×	×	×	×	×	×	×	×	ょ ×
8 り							×	・	・	×	×	×	×	×	×	×	×	×	り 裡
9 し								×	・	・	×	·	×	×	×	×	×	×	し 詩、社、社会
10 ゃ									×	×	×	×	×	×	×	×	×	×	ゃ ×
11 か										×	・	・	・	×	×	×	×	×	か 蚊、解、絵画
12 い											×	・	×	×	×	×	×	×	い 胃
13 が												×	・	×	×	×	×	×	が 蛾
14 げ													×	・	・	×	・	×	げ 下、弦、現実
15 ん														×	×	×	×	×	ん ×
16 じ															×	・	・	×	じ 字、実
17 つ																×	・	・	つ 角
18 の																		・	の 野

図2 単語辞書のアクセスマトリックス

政府は交通環境整備の本年度予算を上期に

集中的に投入すると発表しました。

↓カーソル移動

政府は交通環境整備の本年度予算を上期に

集中的に投入すると発表しました。

交通－間－強制－日の

↓次候補

高－痛感－教－整備の

↓次候補

高－痛感－今日－整備の

図3　同音語の選択とカーサの形状

交通環境整備の

交通－環境－整備

↓学習単位

① 交通－環境

② 環境－整備

↓学習効果

交通の環境を整備する

自然環境の整備が

図4　共起関係による単語学習

せいふはこうつうかんきょうせいびのほんねんどよさんをかみきにしゅうちゅうてきにとうにゅうするとはっぴょうしました。

↓ 仮想ブロック切り

せいふは

こうつうかんきょうせいびの

ほんねんどよさんを

かみきにしゅうちゅうてきに

とうにゅうするとはっぴょうしました。

↓

政府は／交通環境整備の／本年度予算を／上期に集中的に／投入すると発表しました。

図5　変換例