自然言語処理 113－4
音声言語情報処理 11－13
（1996．5．28）

最長共通部分文字列探索を用いたテキストからの仮名漢字変換候補単語の抽出方法

安達久博†

本論文は，与えられた漢字仮名混じり文とその読み仮名文字列を対象とし，この2つの文字列相互間の対応関係を同定する文字列照合方法を提案し，日本語のテキストから仮名漢字変換候補単語を抽出する方法について検討する．すなわち，形態素解析処理を行わずにテキストから漢字表記とその読みのペアを抽出する方法について述べる．この抽出されたペアを仮名漢字変換候補単語と呼ぶ．文字列照合における問題点は，漢字表記の読みの中に平仮名で表記される語と同じ読みが存在する場合に対応関係が多対多になり，一意に決定できずに誤った対応関係が抽出されてしまうという点である．この問題に対する従来の方法は，前方と後方から文字列照合を行い，対応関係が双方で一致したペアだけを抽出する双方向解析を導入し，抽出されなかったペアについては，曖昧さの生じない他の例文から抽出されるとするものであった．しかし，他の例文から抽出される可能性は否定できないが，抽出されない可能性も否定できず，他の例文に依存せずに抽出する方法を検討する必要性は大きい．本提案手法は，2つの文字列間の最長共通部分文字列探索の手法を導入し，対象文字列間の表層的な特徴に基づく制約条件と単漢字変換辞書を用いて，決定論的に対応関係を同定し，仮名漢字変換候補単語を抽出する．新聞記事の見出し300文を対象に実験を行った結果，100％の抽出成功率を得た．

A Word Extraction Method for Kana-Kanji Conversion from Japanese Texts Based on Longest Common Subsequence

HISAHIRO ADACHI†

This paper describes a information extraction method from Japanese texts using string matching for finding the corresponding relations between Japanese sentences and their kana-character strings. This method is based on the longest common subsequences (LCS) using several constraints which are derived from concatenations of Japanese characters. To evaluate the performance of the proposed method, we have applied it to extract pairs of kanji- and kana- character strings from texts which are 300 headlines in newspaper articles and obtained 100 % of extraction accuracy.

1. はじめに

日本語ワープロの仮名漢字変換技術に関する問題点の一つは，同音語選択の問題である[2)]．例えば，「あつい」という読みに対しては，「熱い，暑い，厚い」などの同音語が存在する．最近では，単語と単語の共起関係を利用して，「あつい夏」には「暑い」を，「あつい本」には「厚い」を，というように正しく同音語の選択を行う技術が開発されている．さらには，意味情報，文脈情報までをも取り込む本格的な自然言語処理技術を駆使した大規模なシステムへと改良が行われつつある．しかし，同時に，現在の日本語ワープロが抱える問題点として，仮名漢字変換の精度を向上させるための辞書が複数存在し，かつ巨大化したために複数の辞書の適用順序などを制御する規則が複雑化し，意図しない変換結果を生ずる副作用などの問題と，巨大化した辞書の保守，作成に多大な労力が必要であるという問題がある．

一方，仮名漢字変換技術の実用化に大きく貢献した技術の1つとして，単語の使用頻度に基づく学習機能があげられる．この学習機能は，主に同音語選択に利用され，例えば，「あつい」に対する同音語を「厚い」と選択すると，次回から「あつい」が入力されると，「厚い」が優先的に表示される同音語候補の優先順位を変更する短期学習と，使用した単語の頻度を計測し，辞書内の単語の使用頻度に偏りを生じさせ，辞書を使用者へ適合化させる長期頻度学習に分けられる．すなわち，「使えば使うほど賢くなる」をキャッチフレーズとする学習戦略であった．これは，国語辞典等に収録

† 宇都宮大学工学部情報工学科
Dept. of Information Science, Faculty of Engineering, Utsunomiya University

されている単語の総数に比べ，個人が使用する単語の数は限られており，個々の使用者の単語の使用頻度にはバラツキがあるという考えに基づいている．しかし，この使用頻度に基づく辞書の適合化処理は，仮名漢字変換処理の確定操作を行った場合に起動され頻度が計測され，適合化は徐々に行われる．すなわち，新しい機種や別のメーカの機種に移行した（したい）場合には，以前の機種で学習された単語の使用頻度に関する情報や使用者が登録した単語の情報を新しい機種の辞書情報へ反映させることは一般に困難であった．しかし，学習効果が変換精度の向上に大きな影響を与えることは，明らかである．

ところで，使用者が作成，蓄積した文書の文字情報に関しては，例えば，MS-DOS 形式に変換することで異なる機種間でも利用できる．従って，この使用者が蓄積した文書の文字情報から漢字表記とその読み仮名をペアとした情報を容易に抽出できれば，抽出された情報の頻度を計測し，その結果と既存の単語辞書との融合を行えば，単語の使用頻度の履歴情報を復元し，利用できる可能性は大きい．すなわち，漢字仮名混じり文とその読み仮名文字列が与えられた場合に，漢字表記とその読み仮名文字列との対応関係を同定する文字列照合により上記の問題点が解消できると考える．

2. 背　　景

与えられた漢字仮名混じり文とその読み仮名文字列から仮名漢字変換候補単語，すなわち，漢字表記とその読みのペアを文字列照合を用いて獲得する場合，荒木ら[1]も指摘しているように，漢字の読みの中に平仮名で表記される語と同じ読みが存在する場合，誤った対応関係が導出されるという問題がある．例えば，表1に示したように，漢字仮名混じり文を「宇都宮の街」，その読み仮名文字列を「うつのみやのまち」とし，文字の出現順序に従って，前方から文字列照合を行うと，「宇都宮 = うつ，街 = みやのまち」とうい誤った対応関係が導出される．一方，後方から文字列照合を行うと正しい対応関係「宇都宮 = うつのみや，街 = まち」が導出される．

そのため，荒木らは双方向解析と呼ばれる，前方からと後方からの文字列照合を行い，導出された対応関係が一致したものだけを抽出し，抽出されなかった対応関係については，曖昧さがない他の例文から抽出されるとしている．すなわち，上記の例文からは，何も情報は得られないことになる．確かに，他の例文から獲得できる可能性は否定できないが，そのような例文が得られない場合には，問題点として残る．

本論文では，この問題点を解決するための文字列間の照合方法について提案する．この照合方法は，基本的に2つの文字列間の最長共通部分文字列の探索手法を利用し，対象文字列間の特徴に基づく制約条件を導入し，文字列照合の探索空間を絞り込む方法により，この問題の解決を試みる．

3. 最長共通部分文字列探索

最長共通部分文字列（ LCS ）探索とは，2つの文字列 $A = a_1, a_2, ..., a_m$ と $B = b_1, b_2, ..., b_n$ に対して，両者に共通でかつ最長の部分文字列を導出することである．ここで，部分文字列とは元の文字列から文字の出現順序を変えずに取り出したもので，連続している必要はない．LCS の探索問題に対する再帰的解法は，行列の乗法と同様に最適解のコストを得るための再帰性を立証することであり，p[i,j] を A の部分文字列 $A_i = a_1, \ldots, a_i$ と $B_j = b_1, \ldots, b_j$ の LCS の長さと定義する．すなわち，$LCS(A,B)$ の長さは，p[m,n] を求めることである．この LCS の最適部分構造は，ダイナミックプログラミングを利用して以下のアルゴリズムで計算できる[3]．

```
p[0,0]:=0;
for i:=1 to m do
  p[i,0]:=p[i-1,0];
for j:=1 to n do
  p[0,j]:=p[0,j-1];
for i:=1 to m do
for j:=1 to n do
{ 文字列照合処理 }
 if a[i] = b[j] then
   p[i,j] := p[i-1,j-1]+1;
 else
   p[i,j] := max(p[i-1,j],p[i,j-1]);
```

表1の例文に対する LCS の計算過程を図1に示す．

また，議論を明確にするため，以後，本論文では図2に示すような簡略図を使用する．また，漢字仮名混じ

表1　誤った対応関係が導出される例

宇都宮の街	うつのみやのまち
前方からの文字列照合	
宇都宮	うつ
街	みやのまち
後方からの文字列照合	
宇都宮	うつのみや
街	まち

	うつのみやのまち
宇	0 0 0 0 0 0 0 0
都	0 0 0 0 0 0 0 0
宮	0 0 0 0 0 0 0 0
の	0 0 1 1 1 1 1 1
街	0 0 1 1 1 1 1 1

図1　LCSの計算例

	うつのみやのまち
宇	○○○○○○○○
都	○○○○○○○○
宮	○○○○○○○○
の	○○1○○1○○
街	○○○○○○○○

図2　簡略化した例

り文を $A = a_1, \ldots, a_m$ と，その読みに対応する平仮名文字列を $B = b_1, \ldots, b_n$ とみなし，文字列照合アルゴリズムを説明する．ここで，図中の数字は $a_i = b_j$ の場合の p[i,j] の値とし，それ以外は○で示す．

4. 文字列間の照合方法

4.1 対象文字列の特徴

本論文で対象とする2つの文字列は，漢字仮名混じり文とその読み仮名文字列（平仮名文）である．明らかに，この2つの文字列間の最長共通部分文字列は，漢字仮名混じり文の中に出現する平仮名文字をその出現順序に従って，配列した文字列であり，最長共通部分文字列の長さは，平仮名文字の総数に対応する．

従って，本来の最長共通部分文字列探索問題に対する答えは，漢字仮名混じり文の平仮名文字だけに着目することで，容易に得られる．しかし，本論文の目的は，平仮名文字列との対応関係を同定することである．すなわち，共通文字の照合位置に焦点をあてた導出過程が問題となる．このため，次の制約条件を設定して探索空間の絞り込みによる照合位置の曖昧さの除去について検討する．この制約条件は，漢字仮名混じり文とその読み仮名文字列の文字数に関するヒューリステイックスである．

- 制約条件
  漢字1文字に対する読みは，最低でも平仮名1文字が対応する．すなわち，漢字仮名混じり日本語文の文字数は，対応する読み仮名文字列の文字数以下である．形式的には，漢字仮名混じり文の文字数を m とし，平仮名文字列の文字数を n とすると，$m \leq n$ の関係が成り立つ．

この制約条件を利用すると探索アルゴリズムは，以下のように変更される．

```
for i:= 1 to m do
for j:= i to n do
```

	うつのみやのまち
宇	○○○○○○○○
都	○○○○○○○
宮	○○○○○○
の	○○1○○
街	○○○○

図3　探索範囲の絞り込みにより曖昧さが解消された例

	にほんのうつのみや
日	○○○○○○○○○○
本	○○○○○○○○
の	○1○○1○○
宇	○○○○○○
都	○○○○○
宮	○○○○

図4　曖昧さが残る例

	にほんのうつのみや
日	○○○○
本	○○○○
の	○1○○
宇	○○○○
都	○○○○
宮	○○○○

図5　探索範囲の絞り込みで曖昧さが解消された例

これにより，図2は，探索範囲が絞り込まれ，**図3** に示すように曖昧さが解消される．

一方，図4は，この探索範囲の絞り込みでも，まだ曖昧さが残る例を示す．

これは，一致した文字以降の文字列の文字数について考慮していないためであり，探索アルゴリズムを以下のように変更する．

```
for i:= 1 to m do
for j:= i to i+n-m do
```

これにより，**図5**に示したように探索範囲がさらに絞り込まれ，曖昧さは解消される．

図6に示した例は，参考文献1) で対応関係を誤る例として引用されている例文である．すなわち，双方向解析による文字列照合では，「微細化=びさいか」が抽出されない例である．

明らかに，これまで議論した探索範囲の絞り込みでは曖昧さは解消されない．そこで，探索アルゴリズムを以下のように変更する．

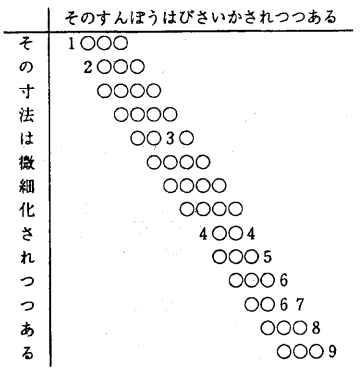

図6　対応関係を誤る例

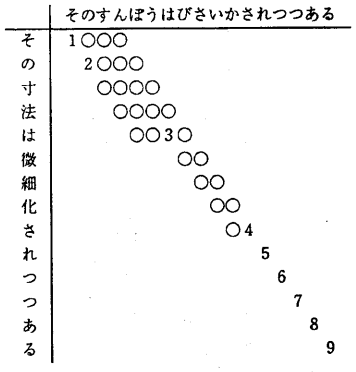

図7　探索範囲の絞り込みにより曖昧さが解消された例

```
k:=0;
for i:=1 to m do
begin
 count:=0;
 for j:=i+k to i+n-m do
 if a[i] = b[j] then
   begin
    count:=count+1;
    p[i,j] := p[i-1,j-1]+1;
    if count = 1 then
         k:=j-i;
   end;
 else
   p[i,j] := max(p[i-1,j],p[i,j-1]);
end;
```

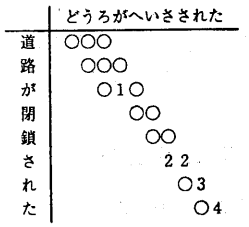

図8　制約条件の適用後も曖昧さが残る例1

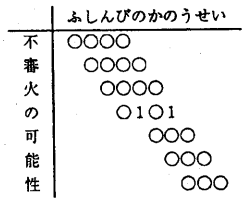

図9　制約条件の適用後も曖昧さが残る例2

この変更は，漢字仮名混じり文の照合対象文字の位置に対応する平仮名文字列の照合開始位置としてきた部分を，実際に文字が一致した照合位置と照合開始位置の差分を加算したものを次の文字の照合開始位置とする．図7にこの探索範囲の絞り込みにより曖昧さが解消された例を示す．

次に，図8と図9に示した例は，これまで議論してきた文字数に関する制約条件による探索範囲の絞り込みを利用しても曖昧さが解消されない例である．

この曖昧さを除去するために，注目している平仮名文字の前後の文字の連接可能性を検査する文脈処理が必要である．すなわち，平仮名文字との連接判定処理と，漢字の読みとの連接判定処理である．特に，後者については漢字表記とその読み仮名に対する先頭の文字と末尾の文字の3項表現からなる連接候補表を常用漢字の1945語について作成した☆．表2にその一部を示す☆☆．

この連接候補表を利用すると，漢字仮名混じり文中の「の」に対する前接文字「火」の読み仮名の末尾候補は，「か，ひ，び」であり，後接文字「可」の読み仮名の先頭候補は，「か，べ」であるから，平仮名文字列の7文字目は接続条件を満たさず，5文字目は連接条件を満たしているため，曖昧さを除去できる．また，図10に示すように片側の文脈文字が片仮名の場合にも対処させるため，便宜上，連接候補表に片仮名文字

☆ 漢字の読み仮名が1文字の場合には，便宜上，先頭，末尾の両方に登録した

☆☆ 常用漢字の表記と読み仮名については Monash 大学の Jim Breen 氏の漢字辞書 kanjidic を利用した．この辞書には，通常の音読み，訓読み以外に接頭語，接尾語として使用される場合の読みも記述されている．

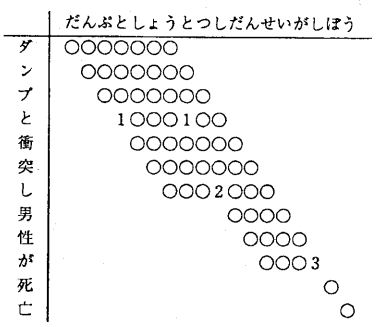

図10 制約条件の適用後も曖昧さが残る例3

表2 連接候補表

漢字	先頭文字	末尾文字
亜	あ	あ
哀	あ, か	い, な, わ
...	...	...
可	か	か, べ
...	...	...
火	か, ひ, ほ	か, ひ, び
...	...	...
湾	わ	ん
腕	う, わ	で, ん

とその読みも追加した.

ところで, 漢字仮名混じり文中の非平仮名文字（漢字, 片仮名文字☆）と平仮名文字列との照合は効率的でないため, 漢字仮名混じり文中の平仮名文字だけを照合対象とする. その結果, 照合アルゴリズムは以下のようになる. 最終的には, 得られた p[i,0] と p[0,j] に格納された値が仮名漢字候補単語の抽出に利用される.

```
{ 前処理 : 初期値設定 }
lcs:=1;
for i:=1 to m do
  if a[i] = 平仮名文字 then
     p[i,0]:=lcs+1;
  else p[i,0]:=0;
for j:=1 to n do
  p[0,j]:=0;

{ 文字列照合処理 }
k:=0;
for i:=1 to m do
begin
 count:=0;
 if p[i,0] > 0 then
    for j:=i+k to i+n-m do
     if a[i] = b[j] then
       begin
        count:=count+1;
        if count = 1 then
           k:=j-i;
        候補 [count]:=j;
        p[i,j]:=p[i,0];
        p[0,j]:=p[i,j];
       end;

{ 複数候補の曖昧さの除去 }
 if count > 1 then
 for l:=1 to count do
  if 候補 [kk] = 不適格判定
     then
       p[0, 候補 [kk]]:=0;
end
```

☆ 一般に長音記号の文字コードは平仮名文字のコードと共通のため, 非平仮名文字に含める.

上記のアルゴリズムの候補の曖昧さの除去は以下に示す適用順序の (2), (3) の処理を意味する.

文字列照合における曖昧さ解消処理の適用順序

(1) 制約条件に基づく探索範囲の絞り込み
(2) 平仮名連続文字に基づく絞り込み
(3) 連接候補表に基づく絞り込み

5. 実験と検討

本論文で提案した文字列間の照合方法を評価するため, 朝日新聞栃木版の事件簿欄1年分から任意に抽出した300文の見出し文を対象として実験を行った. 特に, この中の95文は, 従来の双方向解析で正しく抽出できない例文である. 評価尺度は, 抽出成功率 = 正しく抽出された対応関係の数/正しい対応関係の数 を用いた. 実験の結果, 本手法により対象文のすべてから100 %の抽出成功率を得た. **表3**に, 上記の95文に対する曖昧さの解消に使用された判定条件の内容を示す. なお, 複数の判定条件を使用した例文が1文あり, その内容は平仮名連鎖と両側とも漢字の文脈判定である.

5.1 検　　討

以下では, 図11に示す実験対象の例文を用いて, 両文字列間の共通文字候補が複数存在する場合の候補の決定に利用した連接候補表に関する問題点と今後の検

表3　曖昧さの解消判定条件の適用内容

候補の曖昧さが解消された判定条件	例文数
文字数に関する制約条件	74 文
平仮名連鎖	13 文
漢字と片仮名の文脈判定	2 文
両側とも漢字の文脈判定	7 文

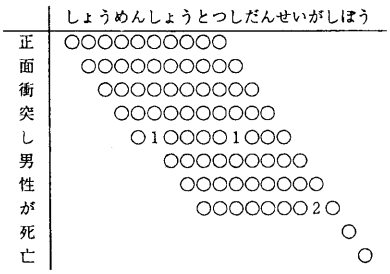

図11　検討例

討課題について議論する．

- **接続表の積極的な利用**
  本論文で提案した照合方法では，文字数に関する制約条件により探索範囲を絞り込み，不適格な照合候補を排除する処理を最優先としてきた．一方，前後の文字を含めた3文字を探索対象として，連接候補表を積極的に利用し，文字列照合を行う戦略が考えられる．そこで，以下の例文を用いて連接候補表を積極的に利用した場合の問題点について考えてみる．
  大洗で小山の会社員が水死（おおあらいでおやまのかいしゃいんがすいし）
  ここで，漢字仮名混じり文中の部分文字列「洗で小」に対する連接候補表から得られる読み仮名候補は，単漢字「洗」の末尾文字が「ら，ん」であり，単漢字「小」の先頭文字が「お，こ」であるから,「らでこ」,「らでお」,「んでこ」,「んでお」となる．しかし，平仮名文字列中には該当する部分文字列が存在しないことになる．そこで，漢字の送り仮名のゆれや仮名固有名詞などの特殊な読み仮名を連接候補表へ追加する対応策が考えられるが，特殊な読みの候補表への追加は，候補表に大きく依存する戦略では「候補の衝突」という問題を増幅させる．現状の連接候補表でも，この「候補の衝突」は避けられない問題であり以下ではこの候補表を利用しない方向で照合方法の改良について検討する．
- **文字数の制約条件**
  図11の例文の場合,「が」は探索範囲内に他の候補が無く確定できる．一方,「し」には曖昧さがある．しかし,「し」と「が」の間が「男性」という2文字漢字ということを考慮すると,「漢字1文字の読みの平均最大値は3文字とする」という制約条件を新たに設定し候補を絞り込む．すなわち，漢字2文字の読みの長さの最大値は6文字となり最初の候補は9文字であり不適切となる．
- **促音，拗音，滲音に関する制約条件**
  また，促音，拗音，滲音等は単語の読みの先頭文字とはならないという制約条件を利用することで，同様に最初の候補は不適切と判定できる．

これまでの議論から，単漢字の読み仮名に基づく連接可能性を判定する連接候補表の使用は，対象文字列間の制約条件を用いても候補を絞り込めない場合にのみ，適用する必要がある．また,「候補の衝突」という連接候補表を用いても候補を確定できない場合の処理については，今後の検討課題である．

6. おわりに

本論文では，形態素解析を用いないで，すなわち常用漢字表から抽出した漢字1文字の読みに基づく連接候補表と文字数の制約条件のみを用いて，与えられた漢字仮名混じり文とその読み仮名文字列から仮名漢字変換候補単語を抽出するための文字列照合方法について述べた．新聞記事の見出し300文を対象として実験を行ったところ，漢字とその読みのペアを100％正しく抽出できることを確認した．特に，抽出候補に曖昧さがある場合にも他の例文に依存せずに1回の照合でもらさず抽出することができる．今後は，対象文の種類と規模を拡大し，大量の文書データに適用して本手法の有効性を検証する必要がある．また，得られた仮名漢字変換候補単語を利用した日本語ワープロの自動学習機能への応用の可能性を検討する予定である．

参考文献

1) 荒木 健治, 高橋 祐治, 桃内 佳雄, 栃内 香次：帰納的学習を用いたべた書き文のかな漢字変換, 電子情報通信学会論文誌 D-II, Vol.J79-D-II, No.3, pp.391–402, (1996).
2) 斉藤 裕美, 野上 宏康：日本語ワードプロセッサにおける自然言語処理, 情報処理, Vol.34, No.10, pp.1241–1248 (1993).
3) Thomas, H., Charles, E. and Ronald, L.: *Introduction to Algorithms*, MIT Press, pp.314–320 (1991).