自然言語処理 97−16
（1993. 9. 17）

カタカナ表記の統一方式
予備分類とグラフ比較によるカタカナ表記のゆらぎ検出法

久保田淳市 † 庄田幸恵 † 河合眞宏 † 玉川博文†† 杉村領一 †

† 松下電器産業株式会社 情報システム研究所
†† 松下電器産業株式会社 ワープロ事業部

大阪府門真市大字門真１００６ 松下電器産業情報機器本部構内

文章中のカタカナ表現を統一する効率的なカタカナ表記不統一検出方式を開発した。従来、辞書の見出しと文章中のカタカナ表記をゆらぎを許容して比較し、不統一箇所を検出する方式が多く提案されたが、カタカナ語は外来語を中心に新語が多く、予め辞書を整備しておくことは困難である。本方式では、基本的には辞書を使わず、カタカナ文字列同士のゆらぎ規則を文章中のカタカナ語に適用し、カタカナ表記のゆらぎを検出する。事前分類したカタカナ表記を有向グラフ形式の中間形式に変換し、効率良く比較検出を行う。試作と実験を行った結果、再現率97.4% 適合率86.7% という結果を得た。

A Method of Detecting KATAKANA Variants in a Document

Jun'ich Kubota † Yukie Shoda † Masahiro Kawai † Hirofumi Tamagawa ††
Ryoichi Sugimura †

† Information System Res. Lab. Matsushita Electric Industrial Co.,Ltd
†† Word Processor Division Matsushita Electric Industrial Co.,Ltd
1006,Kadoma,Kadoma-shi,Osaka,571,JAPAN

A method to detect equivalents for approximate KATAKANA expressions from a Japanese document is proposed. As a Japanese phonetic symbol system (KATAKANA) cannot express precise pronunciation, a KATAKANA loan word can have many expressions. It is also difficult to maintain a dictionary with all KATAKANA expression items. To detect them without a dictionary, this algorithm transforms KATAKANA strings to directed graphs based on rewrite rules, then checks whether they have the same labeled path or not. This method can recall alternative KATAKANA expressions with an accuracy of 97.4%.

1．はじめに

推敲作業を計算機で支援する試みを目的別に大別すると誤り検出、表現展開、表現統一の３つがある。

表 1.1　推敲支援の目的別分類

目　的	実現機能例
誤り検出	綴り誤り検出、同音語ミス検出 文法誤り検出、…
表現展開	類義語辞書、言い換え、国語辞書支援、…
表現統一	文体統一、送りがな統一、カタカナ統一、…

このうち「表現統一」の重要課題にカタカナ表記の統一がある。通常カタカナ語は文章中に5~25% 含まれる[1]が、主要なカタカナ語である外来語は発音表現が不安定なため原語で同じ単語（例えば show ）が異なるカタカナ表現（ショーやショウ）に表される「表記のゆらぎ」という現象が発生する。すべての表記を辞書の見出しとして登録し異表記同士をグループ化すれば良いが、外来語は新語も多くこれらをすべて辞書の見出しとして登録することは事実上不可能である。

従来もこの問題に着目した研究は行われてきたが、主として未登録語として扱われるカタカナ異表記を減らすため辞書のカタカナ見出しと「ゆらぎを許容したマッチング」を行うもの[4][5]であった。

文章の推敲のため文章内のカタカナ表記のゆらぎを検出するものでは、カタカナ列同士の距離を計算するもの[3]や、カタカナ読み列を書き換え規則を用いて「標準読み列」へ変換した上で相互に比較しカタカナ語の同一性を判定する方式[2]がある。これらの手法で、基本的なカタカナ表記のゆらぎを検出することは可能だが、平均的な類似度を基にする「距離方式」では適合率（正しく検出したゆらぎ／検出したゆらぎ候補の数）や再現率（正しく検出したゆらぎの数／検出すべきゆらぎの数）等の精度を上げ難く、また、「標準読み列方式」では、ルールの適用条件を文字列の出現位置、前後のコンテキスト等で調整したりルール間の相互関係、優先順位の関係を調整したりする必要がありルールの保守が課題になる。

本稿では、文章内のカタカナ表記のゆらぎを解消するため、辞書を使わず、標準読み列を設定せず独立のルールを複数個同時に適用し、カタカナ列をグラフ型の中間形式に変換し、それら同士の類似性に基づいてゆらぎの検出を行うカタカナ表記統一方式を紹介する。

本方式により、カタカナ語を延べ6.8 万語含む技術文書を対象に評価を行い、再現率97.4% 、適合率86.7% という結果を得た。また、30% のカタカナ含有量の900 字の文書を処理した時レスポンス時間は0.8 秒（Sun3　ワークステーション）であった。また独立ルール数は58である。

2．ゆらぎの実態

カタカナ表記のゆらぎは種類が多く、規則も不安定である。国語審議会による１９９１年２月の「外来語の表記」に関する答申[7]では、外来語の表記に用いるカナ符号表１と２を定め、原則留意事項１（６項目）と細則留意事項２（１８項目）を併記した。基本的には従来『バと表記するのが望ましい』としていた『ヴァ』なども表記として認め、規制を緩やかにする方針である。一方、国立国語研究所刊「現代表記のゆれ」[6]では、新聞などに現れた「表記のゆらぎ」の実例を挙げているが、「カーテンウォール／カーテンオール」「キャー／キャアー」のように必ずしも答申に現れていないものがある。また、カタカナ語が必ずしも外来語とは限らず「ドッジボール／ドッヂボール」のようなかな遣いのゆれも現実には出現していることを指摘している。

これらに対して実態を把握するため、２つの実文例中における文章中のカタカナ表記のゆら

ぎを調査すると「・」（中点）と「ー」（長音）の有無にかかわるものが多いことが分った。

第一に、数名で作成した計算機マニュアル原稿中のカタカナ語の抽出を行ったところ、全体で24,000語(1,171種)のカタカナ語が存在し164種類のカタカナ表記のゆれが発見されたが、その大半（およそ96%）が中点と長音の有無によるものであった。（表2.1 参照）予め用語統一を図って執筆するマニュアルでも中点、長音のゆらぎは多く見落されている。

表 2.1 計算機マニュアル中の実態

ゆれの種類	種　類	例
中点の有無	145	ウィンドウ・システム ウィンドウシステム
長音の有無	13	インターフェース インターフェス
その他	6	フロッピィ フロッピー
合計	164	

延べカタカナ語数 24,000語　異なりカタカナ語 1,171種類

また、第二の例として、多数の人が独立して作成した技術文書の集合を対象として調査した結果を表2.2 に示す。延べカタカナ語数68,447語（異なり数 7,605種）に対し、496 種のゆらぎがあり、その内、中点の有無 221件、長音の有無 134件であり、この二種で全体の72% を占める。表2.1 と比べると予め用語を統制していないので、中点と長音以外のカタカナ表記のゆらぎが増えている。

表 2.2 技術文書中の実態

ゆれの種類	種　類	例
中点の有無	221	メモリ・アクセス メモリアクセス
長音の有無	134	パートナ パートナー
その他	141	シベリア シベリヤ
合計	496	

延べカタカナ語数 68,447語　異なりカタカナ語 7,605種類

以上の結果から、中点、長音を十分にカバーした上で、多様なゆらぎを扱え、漸増するルールを簡単に整理できる拡張性の高い検出方式が必要であることが分かる。

なお、「ゆらぎ」の定義は「現代表記のゆれ」[6]では「スペシャル／スペシァル」のような表記のゆれと「クロス／クロース」のような語形のゆれに分けてあるが、本稿では基本的には類似性の高いものは語形のゆれも含めて対象とした。また、実文例に生じている「ー（長音）／－（マイナス）」などのような誤用も検出対象に含めた。ただし、以下のような語形が明らかに異なると判断したものは対象外とした。

「プリン／プディング」
「ヘボン／ヘップバーン」
「エルゴノミクス／アーゴノミクス」

3．ルール記法と中間形式

従来のカタカナ表記ゆらぎ検出方式で用いられている規則は、次のようにカタカナ元読み列を標準読み列に変換する書き換え規則であった。

元読み列→標準読み列（適用条件）

【例】クェ→クエ（語頭）

この表現では、複数のルールを記述する場合適用規則の衝突の問題が生じる。例えば、以下、のように左辺に同一の記号が現れるルールがあると、一つのカタカナ列が複数のカタカナ列に書き換えられ合流性が保証できなくなる。

【規則例】イェ→エ
ェ→エ
ェ→削除

【適用例】A：イェルサレム→エルサレム。
B：イェルサレム→イエルサレム
C：イェルサレム→イルサレム

そこで、前後の文脈や、出現位置等を条件に適用条件、適用順序、さらに優先度を考慮し規則の適用対象を制限することになる。このため許容が変化しつつある外来語表記のゆらぎのルールを副作用を抑制しながら増加・調整することは簡単ではない。

そこで、本方式では標準読み列を設定せず、コンパクトに複数の表記を表現できるグラフ型の中間形式を定義する。　このグラフは始点と終点が各々一つのループのない有向グラフである。規則は、部分カタカナ列同士の書き換え規則ではなく、以下のように同一音を表すカタカナ列同士を組にしてゆらぎグループとして記載するため、現実の現象を表現しやすい。また、各ルールは独立であり左辺に同一文字列が現れても構わない。

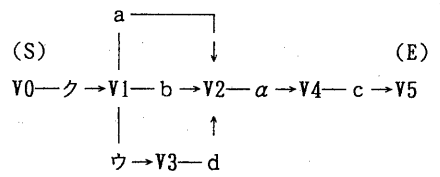
部分列1,部分列2,…部分列N →ゆらぎグループ識別子

【例】ウェ，エ　　→a
　　ウェ，ウエ　→b
　　トゥ，ト，ツ→c
　　ー（長音）　→α（削除可能記号）
　　ェ，エ　　→d

入力読み列に各規則を適用すると中間形式は各辺がカタカナ文字ないし、ゆらぎグループ識別子に対応する有向グラフで表現される。例えば入力カタカナ列「クウェート」に以上の規則が適用され図3.1 のような３つのパスを持つ中間形式を得る。グラフは各頂点がカタカナ列の解釈位置、各辺が解釈記号（ゆらぎグループ識別子もしくはカタカナ文字）に対応している。

２つのカタカナ列に対応したグラフ形式の中間形式G_1 とG_2 が得られると、これらのグラフのＳからＥに至るパスの中で共通のパスがあるかどうかで、これらの二つのグラフが同一のカタカナ語の異表記であるかどうか判定する。

ク［ウェ］[ー][ト]　⇨　クａαｃ
ク［ウェ］[ー][ト]　⇨　クｂαｃ
クウ［ェ］[ー][ト]⇨　クウｄαｃ



S:スタート　E:終了

図３．１　グラフ形式の中間形式

つまり、従来と異なり中間結果がグラフであるため、中間結果の同一性は問わずパスの共通性を判定基準とする。

グラフの比較は縦型優先に行う。すなわち、ＳからＥまでの共通パスが最初に見つかった時点で計算を打ち切り、また、共通の探索パスがなくなった時点で「不一致」と判定し計算を打ち切ることで、平均の探索時間を削減するが、明らかに、最悪の場合はすべてのパスを比較することもありえる。つまり、グラフ形の中間形式を用いることでルールの保守性、一貫性は向上するが、処理時間は増加するおそれがある。

そこで、本方式では次節で述べるように処理時間を短縮するための予備分類、予備比較という処理を行う。

４．カタカナ表記のゆらぎ検出手順

基本的には、文章中のカタカナ表記の組を抽出し中間形式に書き換えた後、中間形式同士の共通パスを探索する。しかし、単純にこれを実施するとカタカナ表記の数をｎとすると比較回数は ${}_nC_2$ となりｎが増すと、組合せの数が増加する。（ｎ＝１０のとき４５、ｎ＝２０のとき１９０、ｎ＝３０のとき４３５）　そこで、比較回数を低減するために、予めカタカナ語を

予備分類して、nの数を減らす手段を設けた。図4.1 に予備分類を含めた検出手順を示す。CとDが予備分類に相当する。

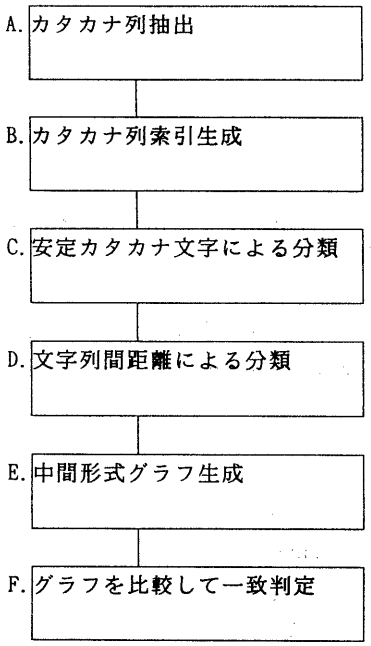


図４．１　検出手順

Ａ．カタカナ列抽出

テキスト中からカタカナ文字列を抽出する。ただし、長音「ー」とハイフン「－」のゆらぎなども存在することから、中点や等号記号などの一部の記号文字を含めて抽出する。

Ｂ．カタカナ列索引生成

文章中から抽出したカタカナ列をソートして異なり表記毎に文章中の格納位置情報を記憶する。表1.2 のように大規模な文書の場合は、カタカナ異なり数は延べ語数の１０分の１程度に低減する。

Ｃ．安定カタカナ文字による分類

ゆらぎ規則の左辺に出現しないカタカナ文字を安定カタカナ文字と呼ぶ。カタカナ文字の中には殆どゆれが生じないものがある。少なくともゆらぎ規則に出現しないカタカナ文字は他の記号に変わらないので、この安定カタカナ文字をインデックスとして異なりカタカナ語を分類する。例えば、「ス」「ル」「セ」がゆらぎ規則に現れていないとすると、「テキスチャ、テクスチャ、ペンシルバニア、アクセント、ペンシルヴァニア、テスト、パセリ、ベルト」の８種類のカタカナ列は以下のように３つに分類される。

「ス」……テキスチャ、テクスチャ、テスト
「ル」……ペンシルバニア、ベルト、
　　ペンシルヴァニア
「セ」……アクセント、パセリ

Ｄ．距離による分類

Ｃで分類されたカタカナ列同士の類似度を距離関数で定義し、距離がある閾値以下のもの同士だけを比較対象とする。距離関数は、中点等の一文字の異なりを許容するように、以下のように定義する。

> ２つの文字列の出現順が同じ一致文字を数え（ｋ）、さらにその文字が文字列の中で出現する際に、間に他の文字が存在するとき（間の文字数－１）×０．５をペナルティーとして除くこの値の、短い文字列長さに対する割合（百分率）を二つの文字列の距離とする。

距離関数をＤ、一致得点を計算する関数をmatch は以下のように定義する。

文字列 $S=s_1, s_2, s_3 \cdots s_n$
文字列 $T=t_1, t_2, t_3 \cdots t_m \quad (n \geq m)$
とすると、

$$D(S,T)=match(S,T)/m$$

$$m=\min\{|S|,|T|\}$$

s_i t_j 以降の一致得点を計算するmatch( )は以下のように定義する。

```
match( s_i ,t_j )=
 {
  s_i =t_j の場合  1+match( s_i+1 ,t_j+1 )
   その他の場合
        (s_i+y =t_j+x ) かつ min(x) かつ
        min(y)  であるx,y に対して、
        1+match( s_i+y+1 ,t_j+x+1 )
              -0.5*(x-j-1)-0.5*(y-i-1)
 }
```

E．中間形式生成

中間形式同士の比較はステップDまでの処理で限定された同一分類中の２つのカタカナ列を対象に行う。まず、２つのカタカナ列を各々グラフ形中間形式 G_1 G_2 に展開する。

３．で説明したように、グラフは各頂点がカタカナ列の解釈位置に、各辺が解釈記号（ゆらぎグループ識別子もしくは安定カタカナ文字）に対応している。まず、カタカナ列は先頭からゆらぎ規則の左辺と比較され、合致しない場合は、安定カタカナ文字を対応させた辺を生成する。合致する場合は、合致する部分をゆらぎグループ識別子記号に書き換え、識別子と対応させた辺を生成する。各辺の先の頂点はカタカナ列の解釈位置に対応しているので、各頂点から同様に辺を伸ばす。この際、解釈位置が共通な場合は、頂点を共有する。

図4.2 に中間形式例を示すが、図でｖn （ｎは自然数）が頂点、ｅn が辺を表す。また、スタート位置の頂点は特にＳ、最終位置の頂点はＥと表現する。

F．グラフ比較

二つのグラフの頂点Ｓから頂点Ｅに達するパスの上にある辺の解釈記号（安定カタカナ文字もしくはゆらぎ識別子）の列同士の中に、一つでも等しいものがあれば、二つのグラフで表現される中間形式は「相互にゆらぎ関係にある」と判定する。逆に、共通のパスがないことが判明すると「相互にゆらぎの関係にない」と判定する。

基本的なグラフ比較手順は以下のような縦方向優先の探索である。

頂点ｕからｖへ記号ｅで繋がっているとき、これをv=next(u,e) と表現する。また、頂点ｖから出ている辺の記号の集合をP=out(v)と表現する。すると、G1上のある頂点ｘとG2上の頂点ｙが対応しているとき、次に探索するべき有効辺上の記号の集合Peは

Pe=out(x)∩out(y)である。

このとき、ｘとｙ以降が一致しているかどうかを判定する関数をyure(x,y) とすると

```
yure(x,y)=
 {
  x=y=E    の場合  一致
  x=E y≠E の場合  不一致
  x≠E y=E の場合  不一致
  x≠E y≠E の場合
   {
     Pe=out(x)∩out(y)
     Pe=φ  の場合  不一致
     その他の場合
     repeat(Pe の各要素e について)
      {
        yure(next(x,e),next(y,e))の値が一
          つでも一致なら一致
      }
   }
 }
```

例えば、図4.2 に２つのグラフの例を示すが、G1におけるパスv1-v2-v4-v7 とG2におけるパス、v1-v2-v5-v6 はその辺に対応する記号がe1,e6,e8と共通なので、これら二つのグラフで表現されるカタカナ列は表記のゆらぎの関係にあると判定する。

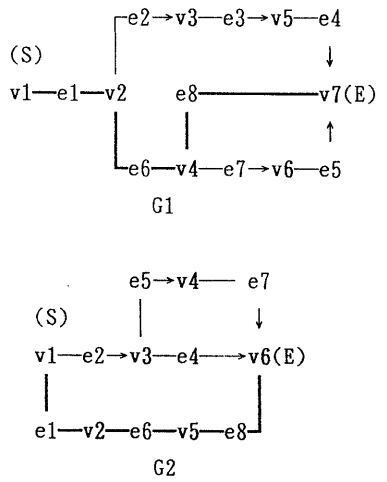


図４．２　中間形式例

５．評価／結果

表2.2 のデータ（複数著者）を評価対象として検出実験を行った。評価指標である再現率と適合率は片方が高いときに片方が低くなるという関係にあるので、距離の閾値をパラメータとして調整した結果、距離閾値34% という条件で、再現率97.4% 、適合率86.7% という結果である。また、30% のカタカナ含有量の900 字の文書を処理した時レスポンス時間は0.8 秒（Sun3ワークステーション）であった。

６．むすび

本方式では、基本的には辞書を使わず、カタカナ文字列同士のゆらぎ規則を文章中のカタカナ語に適用し、カタカナ表記のゆらぎを検出する。事前分類したカタカナ表記を有向グラフ形式の中間形式に変換し、効率良く比較検出を行う。試作と実験を行った結果、再現率97.4%
適合率86.7% という結果を得た。

曖昧性を許容して一致を判定する方式としては、基本的に一文字誤りの検出を対象とするスペルチェックでは「編集距離」(Edit Distance) 等が使われている[8] が、カタカナ表記のゆらぎのように数箇所以上の異なる、不一致の度合いが高い場合には、粗い基準で予備分類した後、グラフ比較等で最終判定する方式が再現率を上げるのには適していると考える。

今後の課題としては、高速化および「マンマシンインターフェース」と「インタフェイス」のようなカタカナ複合語中の表記が他のカタカナ語とゆらぐケースの対策が残っている。

参考文献

(1)黒田他：日本語文におけるカタカナ英語の研究，自然言語処理研究会，68-3,1988,9
(2)福島他：日本語文章作成支援システムＣＯＭＥＴ，信学技報,OS86-21,pp15-22,,1986
(3)奥村他：日本語校正支援システム「Ｆｌｅｘ Ｃ Ｓ」，自然言語処理研究会,87-11,1992,1,pp83-90
(4)島津他：カタカナ異表記・誤記修正機能の開発・評価，情処第４４回全国大会，3Q-4,1992,3,pp3-249-250
(5)青江：カタカナ異表記の生成および統一手法，自然言語処理研究会,94-5,19　93,3,pp33-40
(6)国立国語研究所「現代表記のゆれ」秀英出版　1983年
(7)国語審議会：「外来語の表記」答申,1991.2
(8)Hall,P.A.V.and Dowling,G.R.:Approximate String Matching,Comput. Surv.,Vol.12,No.4,pp.381-402(1980)