2K-3

べた書き文の単語分割における エラーの抽出法と自動訂正

内田幸司　山田洋志　荒木健治　栃内香次　永田邦一
北海道大学工学部

1．はじめに

我々はこれまで，べた書き文のかな漢字変換について研究を行ってきた[1]．その経験から，どのような変換手法であっても，そのアルゴリズムにはなんらかの限界があり，ある量の誤変換の発生は避けがたいと思われる．一方，誤変換となる部分には，一般になんらかの特徴があり，常に同一の誤変換を起すと考えられる．したがって，誤変換部分を抽出してこれをその訂正結果とともに記録しておき，以後の変換結果と比較することにより，同一の誤変換を検出し，自動的に訂正することが可能となる．

本稿では，このうち誤変換の抽出法について報告する．

2．エラーの分析

誤変換抽出の対象としたかな漢字変換システムは，我々が開発中のキーワード方式によるものである．この方式は，べた書き文字列中から一意に識別可能な特定のキーワードによってべた書き文をまず分割し，さらに分割されたブロックごとに単語への変換を行う多段階の変換手法を用いている．現在開発中のシステムは，べた書き文を一旦字種の指定された単語列に変換した後，既に開発済みのＫＫＨシステム[2]によって同音語選択処理を行い，漢字かな混じり文を得ている．本報告では，前半のべた書き文から字種指定文への変換部分を対象とした，したがって，誤変換は字種の誤りと漢字語区切りの誤りのみで，同音語選択の誤りについては考慮していない．

誤変換には以下の２つのパターンがある．

ａ）語の境界が正しくない．

ｂ）語の境界は正しいが字種が異なる．

このほかに，単語をあてはめることができなかったために起った誤変換（未変換）がある．

3．エラーの抽出

誤変換の訂正は，その部分の字種指定をしなおすことによって行う．そこで，訂正の際に同一字種が連続して指定される部分を，誤変換の１単位とする．

したがって，誤変換部分は以下のように抽出される．

ａ）の場合には，訂正を含む単語すべてを取り出す．

ｂ）の場合には，境界が同一であるから，対応する部分の単語を取り出す．

（抽出される単語には，変換された段階についての情報が付いている．）

未変換の場合には，訂正された部分を１単語とみなして取り出す．

このようにして取り出された誤変換を自動訂正するためには，同じ形に変換された部分が出現したときに，それが誤変換であるか否かを判断しなければならない．これに対し，誤変換部分の前後の単語を誤変換と一緒に取り出して，前後の単語との連接パターンの一致を利用するという手法を用いている．（ある単語とそれに連接する文字，単語との間に

A Method for Error Detection and Correction in the Word segmentation of Non-segmentation Japanese Kana Sentences

Kouji Uchida, Hiroshi Yamada, Kenji Araki, Koji Tochinai, Kuniiti Nagata
Hokkaido University

は，一定の関係があることは既に確認されている[3]．）

誤変換切り出しの例を図１に示す．

４．自動訂正への応用

上記の抽出法を用いて，自動訂正を試みた．対象は情報処理関係の論文１篇（斉藤他；北大工学部研究報告，Ｎｏ．１０８，ｐｐ．４３－５２（１９８２））である．変換用辞書には，同系統の論文５篇に現われる単語を収録したものを用いた．実験の対象とした論文は，３３２９個所の字種指定単位からなっている．これを実験システムで処理した結果，誤変換（誤変換を含む）は６２４個所（約１９％）発生し，このうち訂正が可能である同一の形の２回目以降の誤変換は２７９個所であった．なお，この実験では辞書の収録語数が少ないので，未変換が極めて多数になっている．これに対し，前後どちらかの連接語が一致した場合に自動訂正を行ったところ，訂正可能な誤変換のうち約６１％（１７１個所）を自動訂正することができた．

５．今後の課題

今回の実験は辞書の収録語数が少ない状態（約１４００語）で行われた．その結果未変換が多く発生したなどの問題があった．今後は，さらに多くの文献を使って，実験を重ね，また，同音語の誤選択に対しても，同様の手法を用いた解決の可能性について，検討する予定である．

参考文献

１）荒木，内田，山田，栃内，永田
情報処理学会第３２回（昭和６１年前期）全国大会，４Ｔ－１

２）栃内，伊藤，鈴木
情報処理学会論文誌Ｖｏｌ．２７，Ｎｏ．３，ｐｐ．３１３－３２０（１９８６）

３）鈴木
「日本語情報処理における語の接続関係とその応用に関する研究」
北海道大学工学部修士論文（１９８６）

a）の場合

① についてのべる。
② に｜ついて｜の｜べる｜。
③ に ついて｜述べ｜る｜。

① じしょのこがたかが
② 辞書｜残｜が｜高｜が
③ 辞書｜の｜小型｜化｜が

b）の場合

① をへんかんようじしょに
② を｜変換｜よう｜辞書｜に
③ を｜変換｜用｜辞書｜に

① するとかんがえられる。
② すると｜考｜得ら｜れる｜。
③ すると｜考｜えら｜れる｜。

未変換の場合

① にほんごしょりしすてむを
② に｜ほん｜語｜しょりしすてむ｜を
③ 日本語｜処理｜システム｜を

①：べた書き ②：変換結果 ③：目的文
｜：語境界 □：誤変換部分 ＿：連接語

図１ 誤変換の抽出例