漢字文字列に対する読み仮名づけの自動化

1W－4

増田進二　納富一宏　加藤達矢　大野聡　内山明彦
早稲田大学理工学部

１．はじめに

我々は以前から日本語文書校正支援ツール（ＨＳＰ）の開発を行ってきた[1][2]。ＨＳＰで行っている主たるパージングは，字面からの文節単位の文法情報の取得，自立語辞書と付属語接続行列による自立語と付属語の検定とその接続検定である。しかしながら，ＨＳＰはパージングによって指摘された誤りに対して，なんら訂正する機能を持っていない。

日本語文書中の誤りには幾つか考えられるが，その中の一つに誤変換がある。最近のかな漢字変換には，ＡＩ変換など誤変換に対する様々な対策が取られているが，誤変換はまだ多く存在する。

誤変換に対する訂正を考えたとき，非常に有効な情報となり得るのは，誤変換文字列の変換以前の読みの情報である。誤変換の多くは文節の途中で変換してしまう場合で，それは読みの情報を得て正しい文節で再度変換することによって訂正できる。

本稿では，日本語文書中に存在する漢字列に対して自動的に読み仮名をつける手法について述べる。将来的には単独のツールとしても使用できるように，読み仮名をつける漢字列は誤変換文字列だけでなく，すべての漢字列を対象にしている。

２．誤変換の訂正と読み仮名

ワープロ入力で誤変換を起こしやすい操作例として，入力文を文節ごと入力せずに，付属語を先頭にして入力してしまうことが挙げられる。例として，「かれがはこをあけた」という文を入力するとき，「かれが」，「はこを」，「あけた」と文節ごと，または最近のかな漢字変換では一文まとめて入力すれば間違いは少ない。しかし，初心者によくみられる変換の仕方だが，「かれ」，「がはこを」，「あけた」のように文節の途中で変換をしてしまい，付属語を先頭にして変換すると次のような誤変換が生じることが多い。

（正しい変換）　彼が箱を開けた。
（誤変換）　　彼我箱を開けた。

An Automatic Kana Giving Method to Kanji Strings
Shinji MASUDA, Kazuhiro NOTOMI, Tatsuya KATO, Satoshi OHNO, Akihiko UCHIYAMA
School of Science and Engineering, WASEDA University

このような誤変換の文は，自立語辞書によるチェックによって発見できるが，もちろん完全な誤りとしてではなく，誤りの可能性の高い文として検出される。

このような誤りを訂正することを考えたとき，「彼我箱を」の部分の変換前の読み「かれがはこを」が分かれば，それを再度漢字に変換し直すことによって訂正することは容易である。

３．アルゴリズム

ここでは漢字列に対して読み仮名をつけるためのアルゴリズムを説明する。正しい漢字列に読み仮名をつける場合と，誤変換文字列の読み仮名をつける場合とに分けて説明する。

３．１　正しい漢字列の読み仮名

基本的なアルゴリズムは簡単である。ＨＳＰでは入力文の形態素解析を行う際に，文節単位で文字種別の情報も取得しているので，入力文から漢字列を抽出し，自立語辞書で検索し，対応する読みを取り出すだけである。

問題となるのは，一つの漢字（漢字列）に対して読みが幾つか存在する場合である。幾つか読みが存在する場合，その中から正しいものを一つ選択しなければならない。この問題は前後の文字を調べることによって，ある程度解決できる。

例１　行　読み：こう，ぎょう，い，おこな，おこ
　　①　行（い）く…カ行五段
　　②　行（おこな）う…ワ行五段
　　③　行（ぎょう）を…名詞

例２　①　時計（とけい）
　　②　古・時計（ふる・どけい）

例１の場合，「行」が動詞で使用されている場合（①，②）は，その活用を調べることによって読みを特定することができる。後に続く文字が「か」，「き」，「く」，「け」，「こ」ならば，活用が「カ行五段」と分かり，読みは「い」となる。後に続く文字が「わ」，「い」，「う」，「え」，「お」ならば，活用は「ワ行五段」で読みは「おこ

な」であることが分かる。ただし，後ろに続く文字が「っ」のときはどちらの活用かを特定するのは不可能であり，読みの特定も不可能である。また，活用の特定は，用言の活用表を用いることによって行っている。

例１の③は名詞として使用されている場合であるが，この場合，後に続く文字が助詞または名詞となるので，それを調べることによって名詞であることを特定できる。ただし，「こう」と「ぎょう」の判別はできない。

例２は「時計」の例であるが，時計の前に時計を修飾するような語（接頭語）がつくことによって読みが「どけい」と濁る場合である。これは，接頭語が存在するか否かで区別することができる。

３．２　誤変換文字列の読み仮名

誤変換文字列に対しては，その漢字列自体が誤りなのであるから正しい読み仮名などは存在しない。誤変換文字列の読み仮名とは，それを再度漢字に変換したときに，より確からしい文字列に変換できるものである。

誤変換文字列の読み仮名づけのアルゴリズムも，基本的なところは正しい漢字列に読み仮名をつけるアルゴリズムと同様であるが，相違のある部分は読みが幾つか存在する場合の正しい読みの選択方法である。まず誤変換文字列に対して当てはめられる読みをすべて抽出する。それらの読みを再度漢字に変換して解析しなおし，最も文として確からしいものを選択する。

例　彼我箱を

	（読み）	（変換）
①	かれがはこを	彼が箱を
②	かれわれはこを	彼我は子を
③	かれわれそうを	彼我層を

この他にも読みはまだ存在するが，この３つの文の中で文として最も確からしいものは①となる。

４．結果

入力文とそれに対する出力の例を幾つか示す。

４．１　正しい漢字列の読み仮名

例１

入力文

私は学校へ行く。

出力文

私(わたし)は学校(がっこう)へ行(い)く。

例２

入力文

私は学校へ行った。

出力文

私(わたし)は学校(がっこう)へ行(い・おこな)った。

例３

入力文

大きな古時計がある。

出力文

大(おお)きな古時計(ふるどけい)がある。

読みはその漢字列の後にかっこを付けて示してある。例２の「行った」は意味解析まで行わない限り，「いった」と「おこなった」の区別は無理である。候補を一つに絞り切れない場合は，複数表示するようにしている。

４．２　誤変換文字列の例

例１

入力：彼我箱を

出力：彼が箱を

例２

入力：本野表紙を

出力：本の表紙を

５．おわりに

本稿では誤変換の訂正に対して，読みを用いる手法と，漢字列に読み仮名を自動的につける手法について述べた。

問題点としては，候補を一つに絞り切れない場合が多いことが挙げられるが，この点に関してはルールの強化によって改善していく。

また，誤変換の訂正とはいっても特殊な誤変換についてだけであって，同音異義語の問題などには対処できない。今後はこの点に関しても検討していく。

参考文献

[1] 納富，白石，内山：マニュアル作成における日本語文書校正支援ツール，1992年電子情報通信学会秋季大会(1992.09).

[2] 納富，白石，内山：日本語文書校正支援ツールの開発ーマニュアル作成支援についてー，第45回情処全国大会(1992.10).