4C−2　かな漢字変換と漢字かな変換を共に用いる同音語誤りの検出方式

野崎　広志　　鳥原　信一

日本アイ・ビー・エム(株) 東京基礎研究所

1．はじめに

日本語入力においてかな漢字変換入力が普及するにつれて、かな漢字変換結果を過信したり、また、かな漢字変換結果の誤りをうっかり見過ごしてしまうことがあるせいで、同じ読みを持つが意味の異なる単語(いわゆる同音異義語)の間違った使い方(同音語誤り)をした文書が増えている。例えば、「危機一髪」を「危機一発」と間違えたり,「鳥が鳴く」を「鳥が泣く」と間違えたり。

本稿では、日本語入力されて出来上がった文書中に現れる、これらの同音語誤りを検出し、かつ訂正候補を提示するために，漢字かな変換とかな漢字変換と共起関係処理を組み合わせる方式を提案する。

2．同音語誤り検出の従来方式

同音語誤りの検出は，多くが校正支援システムの中の一機能として実現されているが，そのほとんどが以下のような誤用辞書方式またはその変種の方式が採られている([1],[2],[4])。誤用辞書方式では、誤り易い単語または形態素の並びとその書き換えるべき正しい単語の並び(訂正候補)を誤用辞書として持ち、そこに登録されている単語または形態素の並びが文書中に現れることを検出することで、同音語の誤りを検出する。たとえば、誤用辞書に「危機一発」,これの正解として「危機一髪」を登録しておけば、文書中に現れた「危機一発」を検出し、訂正候補として「危機一髪」を提示する。この方式の変種としては、訂正候補を持たないで単に警告を出すだけのものもある。

しかし、この方法では誤用辞書に登録されている単語列と一致しないと、同音語の誤りを検出できない。検出するためには、同音語の組み合わせを網羅的に持つ必要が生じる。例えば、「機器一発」や「既記一発」を同音語誤りとして検出するためには、これらを登録しておく必要がある。さらに、この方式を広義同音語(品詞の異なる同音語)にまで拡張しようとすると「聞き一発」,「利き一発」や「利き一髪」 なども誤用辞書に登録する必要があり、付属語部分まで含めて登録するのは実際には無理がある。

3．本方式の特長

本稿で提案する方式は、検出のための誤りの単語列を辞書に持つ必要は全くなく、文節間の正しい接続しやすさを表す共起関係の情報を持った共起辞書による共起処理と、単語の切り出しと読みを求めるための漢字かな変換と、広義同音異義語を求めるためのかな漢字変換とを組み合わせることで同音語誤りを検出する方式となっている。これにより、本方式では誤用辞書方式に比べ、さらに以下の2つのケースの同音語誤りも検出し訂正候補を提示でき、また、正しい文節間のつながり部分(正しい使い分け部分)も確認することができる。

1)品詞の異なるような同音語誤り(広義同音語誤り)も検出及び訂正候補の提示ができる。

2)係り受け関係にある文節の両方が同音語誤りを起こしている場合にも、同音語誤りの検出及び訂正候補の提示ができる(例えば、「暖かい気候」に対して、「温かい機構」の誤りを検出可能)。

また、その他に訂正候補が複数ある場合に，訂正候補を提示する順番をもっともらしい順番にすることもできる。例えば、「権限を移乗」に対して訂正候補「権限を委譲」と「権限を移譲」の順番を逆にする。

4．本方式の構成

本方式では，大きく以下の4つの処理部から成る。

漢字かな変換部：

日本語漢字かな交じり文を入力として形態素解析を行い、さらに入力文に対するもっともらしい読みを得る。(学校へ行った=>○がっこう・へ／い・った)

かな漢字変換部：

文節の読みに対して、かな漢字変換の全候補を得る。(とって=>取っ手,取って,採って,捕って) かな漢字変換部は、広義同音語を得るためのブラックボックスとなる。

共起処理部：

２つの文節間にどのレベルの共起関係があるかどうかをチェックする。本方式での共起関係は、ある単語と共起しやすいか、あるシソーラスコードを持った単語と共起しやすいかの２種類の関係を持っており、それら関係の組み合わせに応じた共起関係の強さを表1に示す。この共起関係のレベルは、複数の訂正候補があるときの提示の順番決定に使用する。

表1　共起関係のレベルの例

共起関係	共起関係の強さﾚﾍﾞﾙ(C)	例文
単語＋単語	C1	息を呑む
ｼｿｰﾗｽ＋単語	C2	雀が鳴く
(単語＋ｼｿｰﾗｽ		
ｼｿｰﾗｽ＋ｼｿｰﾗｽ	C3	助詞と助動詞

訂正候補決定部：

訂正候補が複数ある場合、入力文との一致度を計算し、共起関係のレベルも考慮して、入力文に対する訂正候補のパスの総合得点で最終的な訂正候補の順番を決定する。入力文との一致度は、共起関係にあったペアで、入力文の自立語と一致した文字数÷入力文の自立語の文字数(Ｊ)、である。訂正候補パスの得点は、訂正候補の得点(＝Ｃ×α＋Ｊ,αは重み)を加算した得点であり、高い順が訂正候補の順番となる。

Homonym Error detection method with both Kana-to-Kanji and Kanji-to-Kana Conversion

Hiroshi NOZAKI , Shinich TORIHARA
IBM Japan , Tokyo Research Laboratory

表2　入力文「権限を移乗」に対する順位の例
（Ｃ１が7、αを１０点とする）

訂正候補	一致度	得点	順位
権限を委譲	2/4	70.5(＝7×10＋2/4)	2
権限を移譲	3/4	70.8(＝7×10＋3/4)	1

５．本方式の処理の流れ

本方式による同音語誤りの検出および訂正候補の提示のステップを、「無視を取っ手から」の同音語誤りを例として示す。なお、係り受け関係とは、本稿では、前後関係も含める。

1.入力された日本語文から単語を切り出す。この結果、文を構成している各文節の自立語と付属語を分けることができる。(無視・を／取っ手・から)

2.各文節の自立語の読みを、漢字かな変換辞書を検索して求める。
　上の1,2のステップは、漢字かな変換(形態素解析)として1つのステップにすることが可能。(無視(むし)・を／取っ手(とって)・から)

3.各文節の読み(自立語の読みと付属語の読みを結合した読み)で、かな漢字変換を行い、文節の読みに対する全ての候補を求める。(むしを=>無視を,虫を,無私を,無死を,蒸しを
　とってから=>取っ手から,取ってから,採ってから,捕ってから)

4.かな漢字変換で得た全ての候補に対して、各候補に関する共起情報を共起関係辞書を検索して求める。
　上の4のステップは、3のステップで、同時に行うことも可能。その場合には、かな漢字変換辞書が共起関係辞書も兼ねることになる。また、漢字かな変換辞書が兼ねることも可能。(共起情報として、　虫 ->「(を)捕る」の前に、
無視 ->「人(を)」の後ろに、取っ手 ->「ドア(の)」の後ろに、捕(る) ->「虫(を)」の後ろに接続しやすい、を得る)

4.入力文の中で、係り受け関係にある文節を係り受けペアとして取り出す。(無視を＋取っ手から)

5.全ての係り受けペアに対して、その係り受けペアに対する同音語ペアの共起情報を用いて、共起関係を計算する。共起関係の計算は、同音語ペアを構成している両方の単語の共起情報に、互いに他方の単語と接続しやすいかどうかがあるかどうかをチェックすることで行う。共起関係にあれば、共起関係に応じた得点をその同音語ペアに与える。(共起関係が「虫を＋捕ってから」で成立する)

6.入力された文に現れる係り受けペアの自立語と同音語ペアの自立語と比較し、入力文との一致度を計算する。(「虫を捕ってから」の入力文との一致度は0)

7. ステップ5とステップ6での得点を足し合わせ、入力された係り受けペアの得点より高い同音語ペアを訂正候補ペアとする。

8. 訂正候補ペアの得点の高い順に、訂正候補として提示する(「無視を取っ手から」の訂正候補を「虫を捕ってから」とする)

６．おわりに

本稿では、同音語誤り検出のために漢字かな変換とかな漢字変換および共起処理を用いる方式を提案した。この方式での問題点として、時間がかかること、漢字かな変換で正しい読みに変換する必要があることが挙げられる。現在、この方式でのプロトタイプがPC上で動作しており、一括して文書を流し、出力結果を得てから、他のツールで一括して文書を訂正するという使い方をしている。一般に同音語誤りは、文書全体の誤りから見れば約14%ほどとかなり少なく([3])、なおかつ広義同音語誤りを起こしている箇所はさらに少ないと思われる。しかしながら、文書中に堂々と現れたときの悪い印象には強いものがある。単体で実用的な同音語誤り検出プログラム(同音語チェッカーまたは、ホモニムチェッカー(Homonym Checker)と呼ぼう)とするためには、高速だが浅い解析と低速だが深い解析を組み合わせることのできる手法が必要であろう。今後さらに研究を進めたい。

参考文献

[1]「次世代ワープロの決め手となるか校正支援／可読性評価ツール」、日経バイト 1988-3
[2]奥村、他：日本語校正支援システム「FleCS」、IPSJ 自然言語処理87-11(1992)
[3]奥村、他：日本語校正支援システムFleCS-新聞社における実用化報告、IPSJ第45回全国大会3F-5(1992-10)
[4]福島、他：日本語文章作成支援システムCOMET、IPSJ 文書処理とヒューマンインターフェース20-2(1988)

付録:

入力:同音語誤りを検出する文書例

台風一家のこの時期に
妹はペットが真でがっかりしていた。
病院からもらった薬には発刊左様があるらしかった。
本当にこの薬は効くのだろうか?
...

出力:一括処理での出力結果例
　　(訂正候補は、[#] 行番号 置換コマンド、
　　"#"はコメント行を表す。)

```
1  S/台風一家の/台風一過の/g
# 1 S/台風一家の/大風一過の/g
2  S/ペットが真で/ペットが死んで/g
3  S/発刊左様/発汗作用/g
# 4 =/薬は効くのだろうか/
```