同音語誤りリストからのかな漢字変換辞書作成

2V－3

鳥原信一　野崎広志
日本アイ・ビー・エム（株）　東京基礎研究所

１．はじめに

かな漢字変換において同音語誤りを減少させることは大きな課題である．大規模なテキストを漢字かな変換を用いてかな漢字変換用テスト例文を作成し，かな漢字変換を評価するシステムの構成法が報告されている[1)]．本論文では，このようにテキストから作成したテスト例文でかな漢字変換を行い，そこから同音語誤りリストを得る．ある入力文字列に対してある程度一意的に決定されるような出力文字列となる複合語・連語を自動抽出する．最後に人手により読みを確認してシステム辞書に登録する．堅実な方式ではあるが着実に同音語誤りを減少させることが可能である．

２．テキストからシステム辞書作成までの処理過程

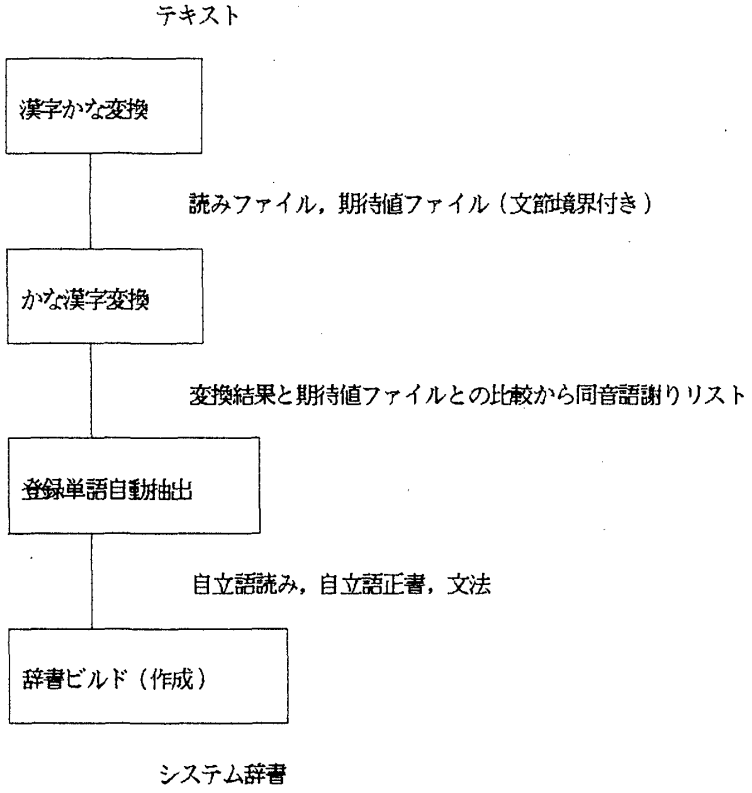

図１．テキストからシステム辞書作成までの流れ

KKC dictionary-building from homonym error list.

Shinichi Toribara and Hiroshi Nozaki

Tokyo Research Laboratory, IBM Japan Ltd.

1623-14, Shimotsuruma Yamato-shi, Kanagawa-ken 242 Japan

3．同音語謝りリストからの登録単語自動抽出

同音語謝りリストからつぎにあげるような単語を消去して，複合語・連語を自動抽出する．

（1）単文節単語

語基が一つの単語は消去する．

（2）自立語にひらがなを含む単語

（例）おかいあげひん　お買い上げ品

複合語であっても，送り仮名のゆれを含む単語は消去する．

（3）異表記のある単語

（例）たばこや　タバコ屋

辞書に「タバコ」の異表記（意味が同じで表記が異なる単語）があるので消去する．

（4）同音語と共起情報をもっている単語

（例）ぶんしょうこうせい　文章構成

「文章」は「校正」と共起するので，ここでの期待値「文章構成」は消去する．

4．かな漢字変換率測定

同音語誤りリストから作成した辞書を使用（従来辞書に追加使用）して変換率を測定した．同一テキストであるが，同音語誤りが約３０％減少し，変換率が約６％向上している．

文節数	5373		
	文節境界誤り	同音語誤り	変換率
未使用	348(6.5%)	941(17.5%)	76.0%
使用	327(6.1%)	662(12.3%)	81.6%

表1．同音語誤りリストから作成した辞書による変換率

5．おわりに

テキストを指定すると，漢字かな変換およびかな漢字変換を用いて，かな漢字変換の同音語誤りリストから複合語・連語を抽出して辞書を作成する方式について述べた．蓄積されたテキストでこの方式を用いて辞書を作成するならば，その分野での同音語誤りは減少する．この方式での問題点は，漢字かな変換の誤りである．効率のよい改良が必要である．この同音語誤りリストは文章校正支援システムにおいても利用可能であると思われる．

参考文献

1)清水，橋本ほか：漢字仮名変換を用いた仮名漢字変換率評価
情報処理学会，自然言語研究会，Vol.1，No.95，pp.1-8(1993)