2K-7　連語解析を用いたべた書きかな漢字変換方式の評価

山階正樹　本間　茂　小橋史彦
(NTT　電気通信研究所)

1．まえがき

日本語ワードプロセッサ等の高機能化が進む中で、自然な入力操作を可能とする高精度なべた書きかな漢字変換法の実現が要望されている。従来の文節入力と異なり、べた書き入力では、同音語による曖昧さがあるばかりでなく、自動分かち書きによる曖昧さも生じ、文節内に閉じた文法解析(形態素解析)のみでは、十分な性能を得ることは難しいと考えられる。そこで、形態素解析に加え、文節間の関係を分析する連語解析を用いた変換法[1]の検討を進めてきた。ここでは、新聞社説をテスト・データに本手法を用いたべた書きかな漢字変換法の変換精度、誤変換原因、変換時間等について報告する。

2．処理の概要

(i)変換アルゴリズム

本処理の流れを図1に示す。キーボードから入力されたかな文字列で辞書を総当りで検索し、文法接続関係を満足する候補のみを蓄積する。句読点等の変換契機が入力されると、文節数が最小となる候補文を生成し、それらについて連語解析を行う。

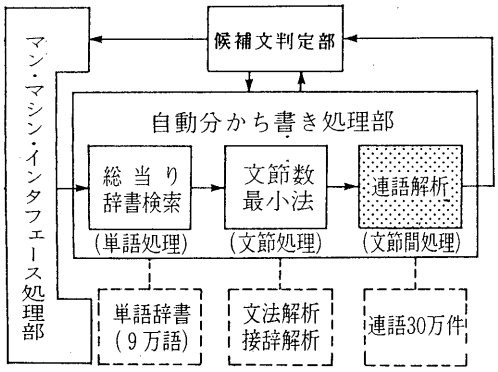

図1　処理の流れ

連語解析とは、「花－が－咲く」、「鳥－が－飛ぶ」等の強い結び付きを持つ単語間の関係を意味概念間の関係に拡張し、それらの関係を記述したマトリックスを参照して、文節間の関係を解析する処理である。なお、ここで用いるマトリックスは約30万件の連語情報[1]をベースとしたものである。

最後に、各候補文について、単語の使用頻度、語長、連語関係の有無を用いた評価値を計算し、最も評価値の高いものを最尤候補として出力する。

なお、辞書の検索はタイピングの空き時間を利用して行い、処理の高速化を図っている。

(ii)辞書構成

変換処理に用いる辞書の収録情報と収録語数を表1に示す。実験システムでは、自立語辞書は、磁気ディスクに、その他の辞書は、主記憶に格納して処理を行う。

表1　変換辞書

辞書の種類	収録情報	収録語数
一般語辞書	かな見出し、漢字表記、品詞、頻度、意味分類番号	約73800語(数詞を含む)
固有名詞辞書	姓、名、地名、企業名	15800語
付属語辞書	助詞、助動詞、形式名詞、補助用言等	422語
接辞辞書	接頭語、接尾語	236語
文法辞書	自立語－付属語接続規則 付属語－付属語接続規則	286×239のマトリックス
連語辞書	意味カテゴリー対応による連語関係	100×100のマトリックス

The Evaluation of Kana to Kanji Translation Method using Collocation
Masaki YAMASHINA, Shigeru HONMA, Fumihiko OBASHI
NTT Electorical Communication Laboratories

３．実験方法

本実験のテスト・データは、新聞社説１０編であり、総入力文数（句読点で区切られる単位）は、８２８文、総文字数は１３，７５９文字である。また、一文当たりの平均文節数は、５．３文節であり、平均かな文字長は２２．５文字である。なお、原文が片仮名の部分は、片仮名シフトで入力し、固有名詞部分については、その指示を行わずに入力した。

４．評価実験

（ｉ）変換精度

前記のテスト・データを変換した場合の変換率は９５．６％であり、変換率は下記の式で定義する。

$$\text{変換率} = \frac{\text{総文字数} - \text{誤り文字数}}{\text{総文字数}} \times 100$$

社説一編毎に見ると、それらの変換率は９７．２％から９４．３％までの差があり、これらの中で、変換率の低いものは、未登録語を多く含む社説や、社説の中で重複して使われる単語を誤変換した場合である。

（ii）候補順位と変換率の関係

表２は、文節区切りおよび同音語の候補順位と、各々の候補順位まで含めた場合の変換率の関係を示している。

表２　変換率と候補順位

文節区切り ＼ 同音語	1	2	3	4～
1	95.6	97.7	98.3	98.7
2	95.6	97.8	－	－
3	95.8	－	－	－
4～	95.8	98.1	－	－

（％）

文節区切りの場合、同音語順位が１位の候補のみでは文節区切り順位４位以下の候補まで含めても変換率の向上は０．２％である。一方、同音語の場合、文節区切り順位が１位の候補のみで同音語順位３位の候補まで含めると変換率は２．７％向上し、４位以下の候補まで含めると３．１％向上する。ここで、文節区切り順位の低い候補を含めてもその効果が同音語の場合と比べて少ないのは、それらの誤りの比率が２：７と文節区切り誤りが少なく、さらに、未登録語等の影響によって正解を得られない場合が多いためである。

（iii）誤変換の原因

誤変換には文節区切り誤りと同音語誤りがあり、まず、文節区切り誤りの主な原因と例を表３に示す。

表３　文節区切り誤りの原因

原　因	変換結果	原　文
未登録語	濃く／ほか／入社は、	国保／加入／者は、
文節数最小法	受け入れのと／上告の	受け入れの／途上／国の

原因の中では未登録語による誤りが約５０％を占めており、未登録語には略語（国公立、官民等）が多い。次に、文節数最小法による誤りが約２５％あり、この中では接辞を文節数１としたことによる誤りが大部分である。

同音語誤りの主な原因とその例を表４に示す。最も多いのが表記のゆれによる誤りであり、同音語誤りの約半数を占める。この中には、漢字表記をかな表記に誤る場合、その逆の場合、さらに、複合動詞間のかなの有無等がある。次に、評価値計算による誤りが約３５％あり、この中では、連語解析による評価が同じで頻度の高い候補を優先した場合等、頻度処理による誤りが９０％以上を占める。この他には、固有名詞を一般語に誤る場合などがある。

表４　同音語誤りの原因

原　因	変換結果	原　文
表記のゆれ	きびしい／政策	厳しい／政策
評価値計算	高級／自治体	高給／自治体
固有名詞	棚か／歯は	田中／派は

（iv）変換時間

実験システムでは、ＣＰＵに、ＭＣ６８０１０を使用している。この場合の平均変換時間は約２００msecと短く、マンマシン・インタフェイス上良好な応答が可能である。

５．むすび

本報告では、連語解析を用いたべた書きかな漢字変換法の変換精度、誤変換原因と変換時間について述べた。字単位での変換率は、９５．６％と高く、また、変換時間も平均２００msecと応答が速く、本方式は、十分、実用的であると言える。今後は本方法を、候補の曖昧さがより大きい音声入力処理等へ適用することを検討していく。

参考文献

１）本間他：連語解析を用いたべた書きかな漢字変換、情処学会、日本語文書処理研究会資料２１