3J−1

大語彙かな漢字変換
—未登録語と区切り誤りの減少—

山田洋志　福島俊一　大山裕
(日本電気株式会社 C&Cシステム研究所)

1 はじめに

従来、かな漢字変換システムは、限定された大きさの辞書と、無数のヒューリスティックルールという組み合わせを基本として進歩してきた。現在の主流である、10万語前後の単語辞書とヒューリスティックルールからなる方式において、文節単位の変換率は80〜85%程度であり、ほぼ限界に達している。しかし、まだユーザの修正作業が多く、一層の変換率向上が望まれている。

筆者らは、変換率向上のための新しい方式として、数十万語以上の規模の単語辞書と、単語間の関係を限定する大規模な制約データを中心とする大語彙かな漢字変換方式を提案し、従来方式と比較しながら考察した[1]。

本稿では、大語彙かな漢字変換方式の概要を説明し、また、本方式が未登録語および変換率に与える効果を確認するために、大語彙(31万語)の単語辞書を試作して実験を行い、未登録語の出現率と、単語区切り正解率を調べたので報告する。

2 大語彙かな漢字変換

大語彙かな漢字変換方式は、大語彙の単語辞書と単語間の関係や制約を記述したデータベースとを用いて高い変換精度を目指す方式である。

図1は、大語彙方式を従来の方式と比較した図である。従来のかな漢字変換方式では、基本単語辞書のほかに、多くの辞書を用いる。大語彙かな漢字変換方式では、それらを大語彙の単語辞書としてまとめて扱う。単語間の関係を記述する方法として、これまでに、格フレームや単語の共起情報や用例などを同音語の決定などに利用して変換精度を上げる試みがなされている[2][3]がこれらは、単語間の関係を規定した制約データとしてまとめられる。

大語彙かな漢字変換方式には、以下に挙げる特長がある。

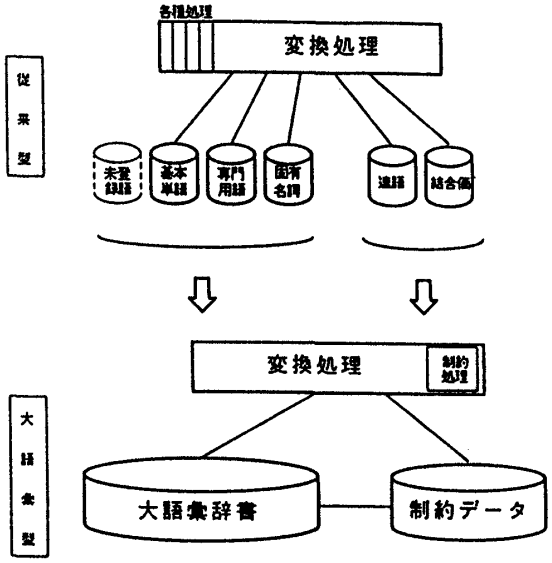

図1 大語彙方式と従来方式

1. 変換率向上のための壁となっている未登録語を減少させる。
2. 複雑に絡み合ったヒューリスティックルールを整理し、自然な枠組みでシステムを構成できる。
3. 使用初期のユーザ単語登録など、ユーザに適応する期間が短縮できる。

3 実験

前述の大語彙かな漢字変換方式の特長を確認するために、(1) 単語区切り正解率、(2) 未登録語を原因とする誤変換の減少、について実験を行った。

3.1 実験条件

研究用のかな漢字変換用辞書(語数約8万。以下、8万語辞書と呼ぶ)に、他の辞書からの語彙を加え、語

Kana-kanji translation method using large scale dictionary
— decreasing unknown words and word-segmentation errors —
Hiroshi YAMADA , Toshikazu FUKUSHIMA and Yutaka OHYAMA
C&C Systems Research Laboratories, NEC Corporation

数約31万の単語辞書(以下、31万語辞書と呼ぶ)を作成した。

語彙の増大の効果を調べるために、単純な変換アルゴリズムを用いている(文節数最小法を基本とし、数種類のヒューリスティックルールを組み込んだ)。また、今回の実験では、同音異義語を選別するための処理は行っていない。

実験のためのテキストは、高校生用の教科書など6つの文献から、1,095文(11,886文節、50,250文字)を使用した[4]。

3.2 実験結果

実験結果を表1に示す。

表1で、変換率は文節単位で計算した。文節は長単位にとっており、四字漢語などは1文節として扱っている。また、表記の揺れは正解として扱った。

表1: 実験結果

	8万語辞書	31万語辞書
区切り正解率(%)	**77.5**	**85.6**
正変換率(%)	64.4	67.2
(表記揺れ)	5.2	10.1
同音語誤り率(%)	13.1	18.4
未登録語率(%)	**6.7**	**1.0**

3.2.1 区切り正解率

この実験では、同音語選択の処理を行っていないので、正変換率の評価をせず、単語区切り正解率を評価する。単語区切り正解率は、正変換率に同音語誤り率を加えたもの、つまり同音語の誤りを許容した変換率と定義した。

単語区切り正解率は、8万語辞書使用時に77.5%、31万語辞書使用時に85.6%であり、8.1%向上している。

3.2.2 未登録語

変換誤りのうちで、未登録語が原因であるものの率を調べた。結果を表1に示す。

未登録語による誤変換率は、8万語辞書使用時に6.7%、31万語辞書使用時に1.0%であり、5.7%減少している。

31万語辞書を使用することで未登録語でなくなった語を、語構成で単語, 接辞付き単語, 複合語に分類し、それぞれを一般語, 地名, 人名, 専門語に分類した。分類ごとの語数を表2に示す(括弧内の数値はのべ語数)。

表2: 未登録語の内訳

	一般	地名	人名	専門	計
単語	52	9	28	55	144
	(81)	(12)	(48)	(129)	(270)
接辞付	68	24	0	30	122
	(88)	(30)		(49)	(167)
複合語	73	11	0	44	128
	(123)	(15)		(73)	(211)
計	193	44	28	129	394
	(292)	(57)	(48)	(251)	(648)

3.3 実験のまとめと考察

以下に実験の結果についてまとめる。

- 区切り正解率が8.1%向上した。この数値は未登録語誤りの減少分5.7%よりも大きい。これは、未登録語が減ることでその周囲の語の区切り誤りも減少したことが原因であり、辞書の大語彙化の効果が単純に未登録語の救済に留まらないことを示している。
- 辞書の語数を増すことで、未登録語による誤りが5.7%減少した。これによって、辞書の大語彙化の効果が直接確認できた。
- 増加した語彙のうち、未登録語減少に寄与したのは、一般語と専門語が多かった。また、語構成では、単語, 接辞つき単語, 複合語が大体同じ割合だった2。この結果は、他の分野のテキストについても調べた上で、語彙拡充の参考にしていく。
- 未登録語以外の誤りには、制約データの導入によって解決するものの他、助詞や付属語の不備によるものがあり、実験システムを細かい点で整備することも必要である。

4 おわりに

本稿では、大語彙かな漢字変換方式について説明した。大語彙かな漢字変換方式には、未登録語の減少によって変換率の限界値が上がる、ヒューリスティックルールが整理される、ユーザ単語登録の負担が減少するなどの利点がある。

今回は、31万語の単語辞書を使って、未登録語による誤りの減少、単語区切り正解率の向上を確認した。

今後は、辞書の大語彙化、制約データの充実を進めながら、正変換率の向上、複数の分野での有効性、アルゴリズムの簡略化などを検証していく。

参考文献

[1] 山田他, 情処40全大 5P−1,1990
[2] 本間他, 情処論文誌 Vol.27 No.11,1986
[3] 山田他, 情処36全大 2T−4,1988
[4] 浅川他, 情処41全大,1990