5P-1

# 大語彙辞書を用いたかな漢字変換についての考察

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

現在、かな漢字変換用の単語辞書の語数は増大を続け、十万語を超える単語を備えたワープロも製品化されている。また、単語辞書以外に単語の用例を集めた辞書を用意して変換精度を上げる試みも行われている。本稿では、かな漢字変換を題材に大語彙辞書を用いることの意義と、大語彙化に伴う問題点について考察する。

## 2 現在の問題点

ここでは、従来のかな漢字変換の問題をいくつか挙げる。

### 2.1 未知語

偶然の一致を別にすると、単語辞書に登録されていない単語を含む文章を正しく変換することはできない。したがって、未知語の存在が理論上の変換率の上限を規定することになる。しかも、未知語があるとその周囲の変換結果に対しても悪影響を与えることがある。

以下に未知語による解析の失敗例を挙げる("ハーバード"が未知語の場合)。

× ハーバー土台が苦悩ら
○ ハーバード大学の裏

### 2.2 解析の曖昧さ

かな漢字変換においては同音語が頻出する。また、単語や文節の区切りも複数考えられる場合がほとんどである。曖昧さを解消するためにさまざまなヒューリスティックルールが用いられているが、解決できない部分が多く、システムを使用する人間の選択に任せているのが現状である。

以下に同音語と単語区切りの違いによる解析の曖昧さの例を挙げる。

同音語　× 実験対称を決める
　　○ 実験対象を決める
区切り　× 建築かに相談する
　　○ 建築家に相談する

### 2.3 一貫性の問題

現在のかな漢字変換システムでは、限られた単語数で変換精度を上げるために単語や文法の一貫性を犠牲にしている部分がある。

例えば、

- 変換誤りを起こしやすいとか同音語が増えるという理由で単語を登録しない。
- ある複合動詞は一つの単語として登録し、他のものは二つの単語として処理をする(接辞付きの名詞や複合名詞も同様)。

などの手段がとられることがある。こういった処置は、特定の語数や分野において最適な結果を得るためのもので、辞書や分野が変化する度に個別に対処していかなければならない。

## 3 辞書の大語彙化

2.1節で述べたように、より高精度の変換を実現するためには、必要な語彙を用意した上でその中から正しい変換結果を得ることが必要である。また、用例などのデータも広義の辞書であり、今後大量のデータが必要になることが予想される。そこで、今後のかな漢字変換システムの一つのあり方として、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、大語彙型のかな漢字変換が考えられる(図1)。

以下では、大語彙化のために登録する単語の種類や大語彙化による影響について、前章の問題点と関連させながら考察する。

---

A study of kana-kanji translation using large scale dictionary
Hiroshi YAMADA , Toshikazu FUKUSHIMA and Yutaka OHYAMA
C&C Systems Research Laboratories, NEC Corporation

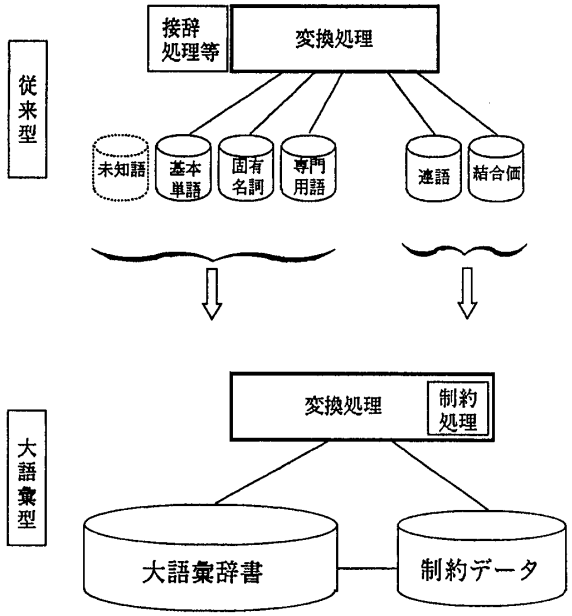


図 1　大語彙型かな漢字変換

### 3.1　未知語

未知語の問題は単語を辞書に登録することで直接に解消される。

日常使われる語は出来る限り辞書に登録する。また、固有名詞や専門用語も登録する。従来は、一部のみを登録し、残りは分野別の辞書として用意して使用者が辞書を選択するという場合が多かったが、大語彙の辞書を用いることで一般の単語とともに統一した枠組で扱うことが出来る。

そのほかの未知語としては、新出語 (専門用語の一部もここに含まれる)、流行語がある。こういった語は辞書の作成時点では予想できないので定期的に辞書の保守を行って補う必要がある。辞書を自動的に構築するための研究が各所で行われており [1][2]、その成果を利用することが出来る。

### 3.2　解析の曖昧さ

曖昧さについて、辞書の大語彙化は増加と減少の二つの方向に働く。

n 文節最長一致法や文節数最小法などの、長い単語や文節が優先されやすい変換方式においては、四字漢語や接尾語付き名詞などの長単位の語を増やすことで解析の曖昧さを減らすことが出来る。ただし、長単位の単語を利用すると単語を構成する個々の要素 (短単位の単語) に関する情報が失われるので、必要なら個々の単語の情報を得るための仕組みを与える。

一方、短単位語の増加は曖昧さ (例えば同音語数) を増加させる。長単位の単語についても、語数が非常に多くなった場合にどの程度の曖昧さが出てくるのか、検討の必要がある。

大語彙の辞書の使用を前提とすると、対象分野や登録単語を限定して曖昧さを減らすといった方法ではなく、それぞれの単語がどのような場合に使われるかを区別することが重要になる。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 [3][4]。今後は、こういった手法を大語彙の単語辞書に対して適用する場合に、必要なデータの量がどう変化するか、あるいは手法そのものの限界がどこにあるのかを検証する必要がある。

### 3.3　一貫性の問題

複合語も積極的に登録してゆく方が大語彙の辞書を有効に利用することが出来る。その場合に複合語同士が同音語誤りや分割誤りを起こすことが考えられるが、これは 3.2節の問題点と同じものであり、共通の枠組で対処することが出来る。

### 3.4　容量と速度

そのほかに語彙が増えることによる影響として、辞書容量・メモリ容量の増加とそれに伴う速度の低下が考えられる。

これらの問題はハードウェアの進歩に伴って、改善の方向に向かっている。また、辞書の圧縮や高速検索などの研究、並列処理、自然言語処理を対象としたハードウェアの研究 [5] が行われおり、考慮の必要がある。

## 4　おわりに

本稿では、大語彙の単語辞書を用いたかな漢字変換について考察した。今後は実際に数十万語規模の単語辞書を使って未知語数や解析の曖昧度の変化を調査し、大語彙辞書を有効に利用するための方式を追究してゆく。

### 参考文献

[1] 白井他, 自然言語処理研究会 51-3,1985
[2] 内田, 人工知能学会 3 全大 8-11,1989
[3] 本間他, 情処論文誌 Vol.27 No.11,1986
[4] 山田他, 情処 36 全大 2T-4,1988
[5] 福島他, 情処 39 全大 2F-5,1989