自 然 言 語 処 理 87－5
（1992．1．16）

大語彙かな漢字変換方式の変換率評価

山田洋志　大山裕

日本電気(株) C&Cシステム研究所

本稿では、文書の分野によらず高変換率を実現する大語彙かな漢字変換方式を提案し、大量のテキストを用いたシステム評価・改良の必要性を述べる。また試作システムの評価結果を報告する。

本方式は、大語彙の単語辞書、単語の組み合わせのデータ(連語や格構文など)を用い、未登録語や同音語誤りを減らして高変換率を実現する。特長として、(a)高い変換率、(b)分野に依存しない性能、(c)ユーザの負担の軽減、が挙げられる。

試作システムは、77万語の単語辞書、27万件の連語辞書を使用する。評価用テキスト45万文字を用意し、また、変換率の計算、誤変換箇所の抽出、変換結果の差分の抽出を自動化した。評価の結果、拡張文節単位の変換率86.9%(従来比+11.9%)、未登録語率1%(従来比 −5.7%)という高性能を実現できた。

A Kana-Kanji translation method using a large-scale dictionary

Evaluation of translation precision rate

Hiroshi Yamada　Yutaka Ohyama

C&C Systems Research Laboratories, NEC Corporation

The authors propose a new Kana-Kanji translation method. The authors also argue that a large text corpus is necessary for the evaluation of natural language processing systems.

The proposed method uses a large scale word dictionary and a large scale collocation dictionary, so that unknown-word-errors and homonym-errors are reduced. The method offers: (1) highly accurate translation rate, (2) corresponding to the sentences of various fields and (3) reduction of user's effort such as new word entry.

The authors made a experimental system based on the proposed method. The system uses 770,000 words and 270,000 collocations. The authors also made a corpus (98,000 phrases) and tools for the evaluation of translation results. The translation precision rate was 86.9% (11.9% higher than a conventional system). The unknown word rate was 1.0% (decrease of 5.7%).

1 はじめに

日本語の入力手段として、現在最も一般的なのはかな漢字変換である。かな漢字変換による入力では変換の誤りが生じ、変換率が問題となる。

現在、変換率の点でメーカーや機種間に大差はなく、画期的な変換率改善も行われていない。しかし、使用者が現在の変換率に満足しているわけではなく、一層の変換率向上を望む声も多い[7]。ここ数年の論文等に見るかな漢字変換システムの変換率は、文節単位[1]で85%前後、文字単位で90%前後である[3][5][6]。筆者らが実際にパソコン用の製品を調べた結果でも拡張文節単位で80%前後(文節単位85%前後に相当)という結果を得ている(2.1節)。

変換率の停滞の理由は、基本的には言語知識の不足である。正しい変換のための最低限の条件として単語が辞書に登録されていることが必要であるが、多くのシステムが使用している10万語程度の単語辞書では多彩な文書をカバーできない。また、同音語は前後関係への依存度が大きいが、そのような単語相互の関係に関する知識は少ない。

さらに、変換率の信頼性にも問題がある。信頼の出来る評価結果を得るには、多くの種類のテキストを充分な量用いて、評価を行う必要がある。しかし多くの論文では、数千～数万文字程度のテキストで評価を行っている。しかも、新聞記事、論文など特定分野のテキストを使っている場合が多い。そのため変換率がテキストの量や種類で変化しやすい。

充分な量のテキストによる評価が行いにくい理由としては、テキストが増えると評価のための工数が多くなることが挙げられる。大規模なテキストによる評価作業を実現するためには、機械的作業の自動化や見やすい形での情報の提示など、評価作業を支援するための環境が必要である。

以上の問題点の背景として、従来手法が研究・開発された時期には、計算機の速度や記憶容量が不足していたため、大量のデータを使うシステムは実用にならなかったという事実がある。しかし現在では、速度も記憶容量も各段に進歩し、自然言語処理用の専用LSIの開発も行われている[14]。

筆者らは、従来方式の限界を超えた高変換率を実現するために、大語彙の単語辞書と、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ている。さらに、本方式に基づくシステムを作成し、また、大量のテキストによる評価作業を円滑に行うための評価環境を構築してきた[9]。評価環境を利用して、試作したシステムの評価・改良を行い、今回拡張文節単位の変換率86.9%(文節単位90%以上に相当)、未登録語率1.0%という高い性能を実現した[10][11][12]。

本稿では、2章で従来のかな漢字変換システムの問題点、3章で大語彙かな漢字変換方式の特長を述べる。4章で実際のアルゴリズム、使用した辞書、評価環境について述べ、5章で変換率を中心とした評価結果を述べる。

[1]本稿では自立語を1語だけ含む場合を文節、複合語等をまとめて扱って複数の自立語を含む場合を拡張文節と呼ぶ。

2 従来のかな漢字変換方式

従来、かな漢字変換の変換率を上げるためのアプローチは、アルゴリズムの改良を中心として行われてきた。すなわち、単語辞書の語数は、一定(10万語前後が多い)で固定してあり、アルゴリズムの工夫によって変換率を上げようとする。基本的なアルゴリズムは、最長一致法、n文節最長一致法、文節数最小法[1]などといった改良が行われてきた。さらにそれに加えて、単語の意味分類を用いた連語[2]、結合価あるいは格フレーム[4]などの補助的な方式を付け加える。こういった方式の改良によって、変換率を上げてきたが、前章でも述べたとおり最近は大きな進展がなくなってきている。

本章では、従来のかな漢字変換方式の変換率と問題点について述べる。

2.1 従来方式による変換率の一例

現在の製品での変換率の一例として、パソコン用の日本語入力システムで拡張文節単位の変換率を求めた結果を以下に示す(求め方は4.4節参照)。使用したのは評価用テキストA(表3)の約半数である。

文書の種類	FEP1	FEP2
変換率	80.7%	76.8%

80%の変換率では、5文節に1つは変換を誤る。使用したテキストは1文節あたり4～5文字であるので、20～25文字に1ヵ所の誤りが生じている。

2.2 従来方式の問題点

本節では従来方式での問題点について使用者、システム開発者のそれぞれの観点から考察する。

2.2.1 ユーザの負担

前節の結果で分かるとおり、現在のかな漢字変換では、かなりの変換誤りが生じる。

ユーザ側からの対策としては、入力した文章の修正の他、単語登録や入力の工夫(通常の読み方では変換できない単語を本来と違う読みで入力)などがある。いずれにしても、文書作成とは別の思考が必要になり、円滑な文書作成の妨げになる。

間接的な手段としては、学習機能の利用がある。ユーザが特別な操作をする必要はないが、効果が出てくるまでに時間がかかる、文書の傾向が変わると効果が落ちる、必ずしも希望の学習を行わないなどの欠点がある。

2.2.2 開発者の負担

従来の方式では、基本となるアルゴリズムだけでは充分な変換率が得られない場合に、プログラム中に多くの規則を組み込んで対処してきた。しかし、より高い変換率を規則の追加で達成しようとすると、次第に規則が複雑になり、しかも効果がだんだん少なくなる。また、全体の関係が分かりにくくなり、新たな改良を加えようとすると、他の部分への影響が予測できなくなる。

2.3 変換率を下げる要因

変換率を下げる要因として、未登録語と同音異義語がある。筆者の以前の調査でもこの二つが誤り原因の3分の2を占めていた[10]。以下ではその二つについて考察する。

2.3.1 未登録語

かな漢字変換は、単純に言うと単語辞書の単語を並べて文章を作成している。したがって、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。結果として、未登録語の出現比率が理論上の変換率の上限を規定することになる。しかも、未登録語があるとその周囲にも悪影響を与え、意味のない単語の並びが現れることがあり、変換率が下がるのみならず、使用者に対する印象も非常に悪くなる。

2.3.2 同音語

日本語では非常に多くの単語に同音語があり、かな漢字変換システムは、複数の同音語のうちから適切なものを選択することが要求される。また、単語や文節の区切り方も複数考えられる場合がほとんどである。

従来、適切な同音語、語区切り選択のための手段として、単語の頻度、単語長や文節数、品詞に基づく経験的規則などが用いられてきた。しかし、前後関係や意味内容を考慮しなければ決定できない場合も多い。前後関係を見る例として、連語や動詞構文を使っているものもあるが、現状ではデータ量が少なく、補助的な役割を果たすに留まっている。

3 大語彙かな漢字変換方式

単語を始めとする言語知識が不足している場合には、変換率がある程度以上良くならないことを前章で述べた。かつては、メモリやCPU性能などのハードウェア上の制約から、システムを出来るだけコンパクトにする必要があり、多くのデータを備えることは出来なかった。しかし、近年、コストの低下や性能の向上によってハードウェア的な制約は少なくなってきている。そこで筆者らは、大語彙の単語辞書をはじめとする多量のデータを活用する"大語彙かな漢字変換方式"によるシステムを開発するという方針をとった。

大語彙かな漢字変換方式の構成上の特徴を以下に挙げる。

1. 大語彙の単語辞書
2. 単語の組み合わせを記述する制約データ
3. 大量のテキストによる評価作業を支援する評価環境

本方式の特長を以下に挙げる。

1. 従来以上の高変換率を実現できる。
   前記の1.、2.によって未登録語や同音語誤りが減り変換率が上がる。また、評価作業の効率が上がることで試行錯誤が容易になる。
2. 幅広い分野の文章を効率よく変換できる。
   単語や制約データ、評価テキストを分野に依存せずに収集すれば、苦手分野を作らず高変換率を維持できる。
3. ユーザの負担が減少する。
   変換率の向上は、ユーザによる修正作業や登録作業の減少につながる。
4. 少数の経験規則でシステムを構成できる。
   単語や連語を増やすことでシステムの複雑化を招く経験規則を減らすことができる。

単語の組み合わせを制限するためのデータ(格フレームや単語の共起、連語など。本稿ではまとめて制約データと呼ぶ)は従来も使用されていた。大語彙かな漢字変換方式では、単語数が増えて組み合わせの数も増えている。同音語順位などでは対処できない組み合わせの曖昧さに対しては、制約データを使用し前後の単語との関連で最適な単語を選ぶ。

本方式を用いて、実用的な性能を実現するためには評価環境の整備も重要である。

本方式の基本—単語や制約データを多く使い変換率を上げる—は特に新しい考え方ではない。それが今までに実際に行われなかった理由のひとつは、実際にシステムを作成してもその評価が難しいということにある。

実用レベルの性能を持つシステムの開発には、システムの試作—大量のテキストによる評価—変換システムの改良という作業を繰り返すことが必要である。しかし、評価作業の全てを人間に頼ったのでは、現実的な期間・人員で評価や改良を行うことができない。したがって、大語彙かな漢字変換方式に基づくシステムの作成には、大量のテキストによる評価を効率的に行うための評価環境の構築が必要である。

4 大語彙かな漢字変換システム

本方式の実効性を示すためシステムを作成した。システムはUNIXワークステーション上にC言語で作成した。変換プログラムの大きさは約82Kbyteである。

本システムの概要を図1に示す。

4.1 大語彙単語辞書

77万語の単語辞書(以下では単語辞書Lと呼ぶ)を作成した。従来使用していた辞書(語数約10万。単語辞書Sと呼ぶ)に、他の辞書からの語彙を加え、語数約31万の中規模の単語辞書(単語辞書Mと呼ぶ)を作成した。さらに、固有名詞を中心として語彙の追加を行った。各辞書の語数を表1に示す。辞書容量は、S、M、Lの順に3.2Mbyte、14.1Mbyte、30.9Mbyteである。ただし、データ圧縮の技術は用いていない。

表 1: 単語辞書の品詞別内訳

	辞書S	辞書M	辞書L
名詞	55,257	215,176	271,986
固有名詞	18,186	32,454	414,344
用言	21,907	58,041	71,104
副詞等	4,109	7,143	9,552
付属語等	498	652	1,083
合 計	99,957	313,466	768,069

4.2 制約データ

単語間の関係を制限するための制約データとして、連語と結合価の2種類を用いている。

4.2.1 連語(単語の共起関係)

連語は、意味的に関連があり、同一の文や文章中に共起しやすい単語の組み合わせである。

現在作成したシステムでは、隣り合った2単語に限定して使用している。また、連語適用の精度を増すために2単語間の位置関係に限定を設けている。関係は、格助詞6種類など全部で9種類である[11]。現在使用している単語の位置関係を表2に示す。

表 2: 連語の種類(単語の位置関係)

種類	例
AB	「情報—処理」
AをB	「情報–を–処理」
AのB	「情報–の–処理」
AがB	「事故–が–起こる」
AにB	「口–に–くわえる」
AでB	「あご–で–使う」
AとB	「需要–と–供給」
AがBされる	「効果–が–期待される」
AするB	「始まる–時間」

連語辞書は、姫路短大(現・愛知淑徳大)の田中教授による「語と語の関係データ」[8]を中心にして作成した。これは、朝日新聞の記事およびJICSTの科学文献抄録から単語の組み合わせを抽出したものである。約50万件の原データから、見出しが単語のもの26万件を選んだ。さらに、テキストから人手によって抽出した連語1万件を追加し、計27万件を含む連語辞書を作成した。

4.2.2 結合価(用言の構文)

名詞の意味分類と格助詞を組み合わせたパターンを登録している。名詞の意味分類は、基本語約8万語に与えてある。結合価は約1,800用言分を登録してある。辞書の一部を図2に挙げる。

4.3 変換アルゴリズム

基本となる変換アルゴリズムとしては、コスト最小法を用いた。すなわち、単語や文節の数、連語の適用数を数値(コスト)に置き換え、最もコストの

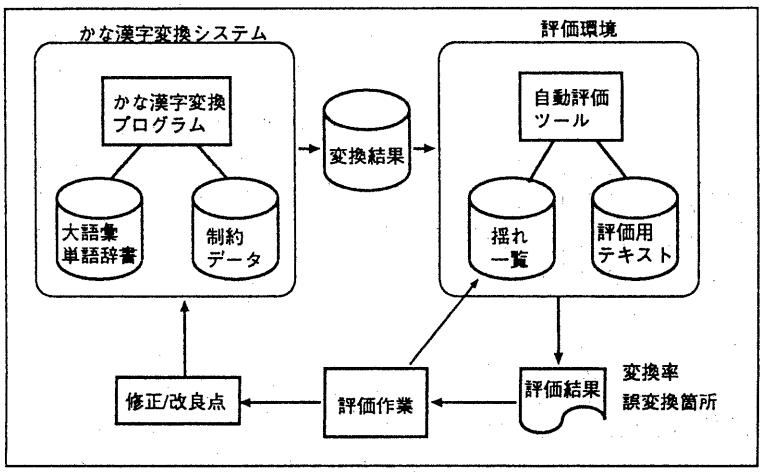

図 1: 大語彙かな漢字変換システムの概要

上がる	: [人,もの]が [場所]から [場所]に
	: [人,もの]が [場所]から [場所]へ
	: [抽象]が
揚がる	: [もの]が [場所]から [場所]に
	: [もの]が [場所]から [場所]へ
騰る	: [抽象]が
挙がる	: [抽象,人]が

図 2: 結合価辞書の例

小さくなる組み合わせを変換結果とする。以下に変換の手順を簡単に示す。

1. 文節形成　単語辞書の検索、接続検定を行い、文節を作成する。
2. 文節のコスト計算　文節に含まれる単語の数と品詞によって決定する。品詞の種類に応じて7種類の数値を用いている。
3. 連語・経験規則によるコスト計算　経験的規則として、連体形＋名詞、助詞の連続など6項目のみを利用している。
4. 変換結果候補の作成　2.、3.の計算結果を用いて、コストの小さい順に一定個数の変換結果を作成する。
5. 結合価によるコスト計算　作成した各結果に対して結合価によるコスト計算を行い、最終的にコストが最小のものを変換結果とする。

4.4 評価環境

評価に用いるためのテキストデータベースと、評価作業を支援するツール群からなる。

4.4.1 評価用テキスト

評価用のテキストは、目的によって以下の2つに大別することが出来る。

- **アルゴリズム検証用**　動作が正常であるかどうかを確認するためのテキストで、導入した手法が効果を発揮する典型的な例や、エラーを起こす恐れのある特殊な構文の文章などを用いる。一通りの動作が確認できれば良く、一般的には少量である。
- **性能評価用**　変換率や誤りの傾向を調べるためのテキスト。実際にシステムを使用する対象に近い文章を使用する。できるだけ多い方が評価結果の信頼性が増す。

アルゴリズム検証用テキストは、アルゴリズムの変更時にその都度用意している。性能評価用のテキストは以下の5分野に分けている。

- **教科書** 国語、理科、歴史など。
- **新聞記事** 政治・経済など。
- **事務文書** 業務上の手紙、法律文書など。
- **日常文書** 手紙、広告、小説など。
- **専門分野** 技術論文、哲学書など。

また、評価結果に作為が入りにくいようにテキストを2組に分割し、一方をシステム改良用、他方

を変換率計算用としている。表3に分野別の文字数と文節数を示す。

表 3: 評価用テキスト

	テキストA		テキストB	
	文字数	文節数	文字数	文節数
教科書	50,250	11,886	48,346	12,334
新聞	70,060	12,025	55,787	13,773
日常	27,182	5,087	35,405	8,454
専門	50,090	8,700	35,199	8,196
事務	48,453	7,777	43,777	10,491
合計	246,035	45,475	218,514	53,248

評価用のテキストは、文節区切り("／")の入った漢字かな混じり文と、読みがなの組からなる(図3)。図3で、'h'で始まっているのが漢字かな混じり文、'y'で始まっているのが読みがなである。

```
h六カ国が／それを／承認すれば／会議は／閉幕する。／
yろっかこくがそれをしょうにんすればかいぎはへいまくする。
```

図 3: 評価用テキスト

4.4.2 評価作業支援

大量テキストによる評価作業を可能にするために、作業の一部自動化を中心とする評価環境が必要である[6][9]。今回は、2つの作業を完全に自動化し、3つの作業を一部自動化した。

完全に自動化した作業は以下のとおりである。

かな漢字変換 読みがなファイルを指定するだけで自動的に行える。

変換率の計算 かな漢字変換結果を評価用テキストと比較し、一致している箇所の数を計算する。変換率は

正変換文節数 / 原文書の総文節数 ×100(%)

で計算し、文節中の一部でも間違っている場合は誤りとする。ただし、表記の揺れは正解に含めている。変換率は拡張文節単位で求めている。

一部を自動化した作業は以下のとおりである。

表記の揺れ 表記の揺れを正解に含めるため、表記揺れを登録したファイルを人手で作成する。揺れファイルに基づいて変換率を再計算する作業は自動的に行う。

誤変換の検討作業 変換結果から誤変換箇所を抽出する作業は自動化した。抽出結果の例を図4に示す。変換誤りの箇所は"【"と"】"で、表記の揺れは"《"と"》"でくくって示している。

```
----------<69>
お得意様へ／ご迷惑の／ないよう、／
お得意様へご迷惑の【内容、】
----------<131>
改めて／お支払いの／可能な／時日を、／
改めて《お支払の》可能な【事実を、】
```

図 4: 誤変換箇所の抽出結果

抽出結果を見て、具体的な対策を考える作業は人間が行う必要がある。そのため、正解テキストを併記して違いがわかりやすいようにしている。また、出力されるのは誤りを含む箇所だけなので、人間が見なければならない量はかなり減っている。

修正による影響の検討 アルゴリズムやデータに変更を加えた場合、変換結果にどのような変化が生じるかの差分を図4と同形式で出力する。それによって、修正による改善(改悪)を確認できる。

5 大語彙かな漢字変換システムの評価

5.1 変換率の評価

単語辞書Lと27万件の連語辞書を用い、評価用テキストB(表3)を対象にして、変換率を測定した。また、比較用に単語辞書Sを使用した。

評価は、辞書を4通りに組み合わせて行った。

1. 単語辞書L＋連語辞書＋結合価辞書
2. 単語辞書L＋結合価辞書
3. 単語辞書S＋連語辞書＋結合価辞書
4. 単語辞書S＋結合価辞書

各条件での変換率を表4に示す。

単語辞書の大語彙化による変換率の向上を表5(左)に示す。これは、単語辞書Sと単語辞書Lでの変換率の差である。

連語の使用による変換率の向上を表5(右)に示す。これは、連語辞書を使用するかどうかによる変

表 4: 正変換率

分野	評価1	評価2	評価3	評価4
事務文書	88.6	87.0	78.6	75.7
新聞記事	84.6	83.2	75.1	72.6
教科書	87.5	85.7	79.2	76.6
日常文書	85.7	83.9	76.5	74.7
専門分野	88.7	87.3	77.9	75.8
平均	86.9%	85.3%	77.3%	75.0%

表 5: 大語彙化・連語の影響

	単語増の影響		連語の影響	
分野	連語有	連語無	辞書L	辞書S
事務文書	10.0	11.4	1.5	2.9
新聞記事	9.6	10.6	1.4	2.5
教科書	8.3	9.1	1.8	2.6
日常文書	9.2	9.2	1.8	1.8
専門分野	10.9	11.5	1.4	2.1
平均	9.5%	10.3%	1.6%	2.4%

換率の変化である。

評価テキストの変換時に使用された連語数を表6に示す。この中には、変換結果に影響を与えなかった(連語なしでも正しく変換した)連語の数も含まれている。

表 6: 実際に使用された連語数

分野	延べ個数	種類	平均頻度
事務文書	772	409	1.77
新聞記事	992	938	1.06
教科書	888	744	1.19
日常文書	467	432	1.08
専門分野	676	577	1.17
合計	3,745	3,100	1.20

5.2 未登録語率の評価

単語辞書の大語彙化による未登録語の減少を確認するために、未登録語を原因とする誤変換率を評価した。未登録語の出現率は下表のとおりである。

	単語辞書S	単語辞書M
未登録語率	6.7 %	1.0 %

評価用テキストA(表3)の教科書を使用し、単語辞書Sと単語辞書Mでの比較を行い、連語辞書、結合価辞書は使用していない。

単語辞書Mの使用で、未登録語による誤変換率が5.7%減少して1.0%となった。単語辞書Lによる未登録語率調査は行っていないが、(語数が増えているので)1%以下であり、辞書サイズ拡張の優先度は低いと考えている。

5.3 連語の効果の評価

5.1節の結果では連語による変換率向上の効果が2%程度と少なかったため、理想的な状態でどのくらいの変換率向上が望めるかを調べた。

実験に用いた連語辞書は、連語を用いない場合の変換結果から誤変換個所を抜きだし、誤変換箇所を参照しながら手作業で連語を辞書に登録して作成した。登録した連語は1,468件である。従って、この実験で用いた連語辞書は、実験に使ったテキストを変換するための連語はほとんど収録している、という理想的な状態になっている。

上記の連語辞書を用いて変換率の測定を行った。単語辞書Mを使用し、変換対象として評価用テキストA(図3)の教科書の一部を用いた。

連語を使用しない場合の変換率と、使用した場合の変換率を下表に示す。

文書	連語無	連語有	差
平均	86.1%	91.4%	5.3%

連語を用いることで平均5.3%の変換率向上が得られた。これは、変換に必要な連語が辞書にすべて蓄えられている場合の向上分である。

実験に使ったテキストが少ないが、良質の連語データを充分な量用意すると、5%以上の変換率向上の可能性があることが分かる。

5.4 考察

5.2節の結果によると単語辞書の大語彙化で未登録語による誤りが5.7%減少した。5.1節の結果から見ても大語彙化の効果が大きいことが分かる。このように、高変換率を実現するには単語辞書の大語彙化が有効であることがはっきりした。

連語による効果は約2%であったが、5.3節の結果から分かるように、連語の追加で変換率がさらに上がる可能性がある。また、事務文書では各連語の使

用頻度が大きく、新聞記事では小さい。これは、事務文書では表現の繰り返しが多いため連語の効果が出やすく、新聞記事は内容が多岐にわたるため、多くの連語を必要とするためと考えられる。

以下にその他の誤りについて幾つか挙げる。

- **未登録語**　カタカナ語(地名・人名など)の不足が目立つ。これが新聞記事の変換率が平均より低い原因になっている。
- **会話文の変換**　"～しちゃった"、"～するんです"、など砕けた会話表現が変換できない。これが日常文書の変換率が平均より低い原因になっている。
- **連語による誤り**　一字漢字語を含む連語による単語分割誤りが生じた。以下に例を挙げる(下線部が連語)。

> 豆を挽く → 真<u>目を引く</u>
> 耕運機を使う → 幸運<u>気を使う</u>

連語の利用によって生じた誤りのほとんどがこのパターンであるため、適当な対策を講じることで連語による誤りを無くすことが出来る。対策としては、一字漢字語を含む連語のコストを変えることを考えている。

6　おわりに

文章の分野によらず高変換率を実現する大語彙かな漢字変換方式を提案した。また、実用的な性能を持つシステムを開発する場合に、大量のテキストを用いたシステム評価・改良が必要であることを述べた。さらに、本方式に基づくシステムを試作し、その性能の評価結果を報告した。

本方式は、大語彙の単語辞書、大量の制約データ(連語や格構文)を用い、従来方式による誤りの中で大きな割合を占める未登録語・同音語誤りを減らして高変換率を実現する。本方式の特長として、(a)高い変換率、(b)分野に依存しない性能、(c)ユーザの負担の軽減、(d)経験規則が少数、がある。

試作システムには、77万語の単語辞書、27万件の連語辞書を使用した。評価用に45万文字のテキストを用意し、また、変換率の計算、誤変換箇所の抽出、変換結果の差分の抽出作業を自動化した。評価の結果、拡張文節単位の変換率86.9%(従来比+11.9%)、未登録語率1%(従来比 −5.7%)という高い性能を実現できた。この変換率は、文節単位では90%以上に相当する。またパソコン用の製品と比較しても6～10%の向上である。

今後の課題として以下のものがある。

- **連語の追加**　5.3節の結果から分かるように、連語の効果が充分出ていない。今後、より多くの分野から連語を収集することで連語数を増やし、慣用的な表現など多くの文書で共通に使われる連語も増やしたい。
- **単語辞書の整備**　未登録語率が1%程度と低いため緊急に追加する予定はない。ただし、何らかの方法でまとまった数の単語が入手・作成できれば登録する予定である。また、会話文対応は、早期に行いたい。
- **評価環境の整備**　未登録語の判別、誤りの分類・データベース化など自動化・半自動化できる作業が残っている。また、大規模な辞書を管理するための環境が必要である。
  現在の環境をかな漢字変換以外へ適用することも試みたい。
- **実用化のための技術開発**　本方式は、辞書容量が従来よりかなり大きくなっている。実用化のためには、容量や速度も重要であり、辞書の圧縮技術を開発する[13]。形態素解析用LSI[14]の利用も検討したい。

謝辞　辞書作成、評価環境の構築に協力していただいた日本電気オフィスシステム(株)浅川氏、本研究のために辞書・データを提供していただいた社内外の関連部門の方々に深謝致します。

参考文献

[1] 吉村他,情処論Vol.24,No.1,pp40-46,1983
[2] 本間他,情処論Vol.27,No.11,pp1062-1068,1986
[3] 山階他,情処33全大2K-7,1986
[4] 大島他,情処論Vol.27,No.7,pp679-687,1986
[5] 隈井他,電情通学会89秋季全大D-26,1989
[6] 恒川他,情処HI研究会34-4,1991
[7] 日経バイト1989年5月号,pp161-162
[8] 田中他,情処40全大5F-1,1990
[9] 浅川他,情処41全大3J-2,1990
[10] 山田他,情処41全大3J-1,1990
[11] 山田他,情処42全大5Q-3,1991
[12] 山田他,情処43全大1F-1,1991
[13] 福島,情処43全大3H-9,1991
[14] 福島,情処論Vol.32,No.10,pp.1259-1268,1991