自然言語処理 122－16
（1997．11．21）

かな漢字変換における誤入力の訂正

山本 喜大† 久保田 淳市† 庄田 幸恵† 白井 豊‡

† 松下電器産業株式会社 マルチメディア開発センター
‡ 松下電器産業株式会社 ＡＶＣ商品開発研究所

日本語の入力操作全体での効率向上のために、かな漢字変換上に誤入力を訂正する機能を試作した。本かな漢字変換は、置換、挿入、脱落誤りのいずれかを一つ含む文節を訂正する。ユーザの訂正指示操作に応じて複数の誤り訂正候補を提示しユーザが選択する方法（選択型訂正）と、変換操作に応じて自動的に誤りを訂正する方法（自動訂正）の２つのインターフェースを想定し性能を評価した。実験の結果、選択型訂正の再現率は第１位が39%、上位６位までで59%であった。また自動訂正の再現率は44%、誤り率は3%であった。

Input Error Correction on Kana-Kanji Conversion

Yoshio YAMAMOTO† Junichi KUBOTA† Yukie SHODA† Yutaka SHIRAI‡

† Multimedia Development Center, Matsushita Electric Industrial Co., LTD.
‡ AVC Products Development Laboratory, Matsushita Electric Industrial Co., LTD.

In order to improve the efficiency of a Japanese input interface, we developed functions for correcting input errors. It corrects a substituted, inserted or deleted error in a clause. We implemented two interfaces, one is a system that outputs corrected candidates on demand by a user and the user selects one (correction by selection), and the other is a system that automatically converts a Kana string to a corrected string on kana-kanji conversion (automatic correction). The recall rate of "correction by selection" is 39% for the first order, and 59% in total for six orders. The recall rate of "automatic correction" is 44%, and its error rate is 3%.

1．はじめに

従来、日本語入力の効率を向上するために、かな漢字変換の変換精度の向上に重点が置かれてきた。しかし、入力操作全体では、本来入力すべき文字キーや変換キーの他に、誤入力に起因する無駄なキー操作が多数含まれている。誤入力を容易に訂正する機能の実現により、実質的な入力効率の向上が期待できる。

誤入力の分析としては、英語のスペルの誤りを置換、挿入、脱落、交換誤りに分類し、これらが誤りを含む単語の8割を占めることがみいだされた[1]。日本語の入力では、置換、挿入、脱落が誤入力の9割を占めるとの報告がある[2]。文章中の検出・訂正方式としては、校正支援システムの一機能として提案されているが、これらは漢字かな混在文字列中の誤りを検出・訂正するものである。一方、入力時の誤りの訂正は、キーコードまたはかな文字列を対象とする。近年ではパーソナルコンピュータのＩＭＥ(Input Method Editor)上に、誤入力の訂正機能が実現され商用化されているが、誤入力のごく一部だけを訂正するにとどまっている。関連する研究に、かな文字列の文節を対象にマルコフモデルに基づき、置換、挿入、脱落誤りを検出、訂正する方法[3]や、音節マトリクスを対象にマルコフモデルと辞書引きを併用する方法[4]が提案されているが、入力時の誤り訂正はあまり報告されていない。

我々は、かな漢字変換上にかな入力による誤入力（置換、挿入、脱落誤り）を訂正する機能を試作した。ユーザの訂正指示操作に応じて誤り訂正候補を提示し選択する方法（選択型訂正）と、ユーザの変換操作に応じて自動的に誤りを訂正して変換結果を出力する（自動訂正）方法について実装し、その性能を評価した。

2. 誤入力の分析

2.1. 調査方法

キー操作履歴を採取・整理して319文節の誤入力を抽出した。

・ 被験者：初級者（1ヵ月に2～3回の頻度でWPを使用）６８名
・ 入力文章：ワープロ技能検定試験3級の問題（延べ文字数　約8500文字）
・ 機種：ＷＰ（panasonic U1pro）
・ 入力方式：かな入力

2.2. 誤入力の分類1（キーストロークに基づく）

ユーザが入力したキーストロークに基づき誤入力を次のように分類する。

(1) 文字誤り：正しい綴りを知っていて、運動能的調整がずれたことに起因する誤り。
  (a) 置換　(正：ぶんしょう、誤：ぶんしよう)
  (b) 挿入　(正：にほんご、誤：にほこんご)
  (c) 脱落　(正：はっきさせる、誤：はきさせる)
  (d) 交換　(正：そうちである、誤：うそちである)
  (e) 複数誤り (正：こゆうの　誤：こうのの)
  (a)～(d)を単一誤りといい、これらを複数含む場合(e)を複数誤りという。(e)の例では、「ゆ」の脱落誤りと、「の」の挿入誤りが同時に起こっている。
(2) 認知誤り：思い違いによる誤り。
  (f) 読み違い (正：しゅもくてき、誤：しゅたい)
  (g) 仮名遣い (正：むずかしい、誤：むづかしい)
(3) また、文字誤りと認知誤りの複合的な場合にモード誤りがある。(正：4つの　誤：うつの)

上記分類に基づき集計した結果を図1に示す。なお、交換誤りは1例ありその他に分類した。

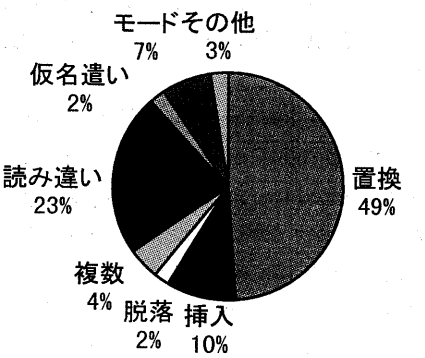

図1　誤入力の分類1(キーストロークに基づく)

2.3. 誤入力の分類2（文字単位）

誤入力の分類1ではキーストロークに着目して誤入力を分類したのに対し、分類2ではキーストロークから得られた文字に着目し、単一誤りのいずれかに分類する。

(1) 誤入力の分類1の認知誤りは、形式的には単一誤りとみなせる場合が多い。例えば、「ぶんしょう」を「ぶんしょ」と誤る読み違いは、脱落誤りとみなせる。また、仮名遣いの誤りはすべて置換誤りとみなせる。

(2) 通常、ストローク数と入力文字数は同じであるが、かな入力モードでは、濁音と半濁音は2ストロークで1文字が入力される。このため入力されたかな文字列の各文字に着目すると、分類1と一致しない場合がある。例えば、「どうようである」を入力しようとして、濁点の入力を誤って隣のキー「せ」をタイプすると「どうようてせある」となる。ストロークに着目する分類1では置換誤りであるが、得られた文字に着目すると複数誤りとみなされる。この他、分類1の挿入誤りが置換誤りとみなされる例（「なるほど」を「なるぼど」と誤入力）、分類1の脱落誤りが、置換誤りとみなされる例（「てんじょうから」を「てんしょうから」と誤入力）等がある。

そこで、得られた文字に着目して、誤入力を再分類し集計した。この結果を図2に示す。

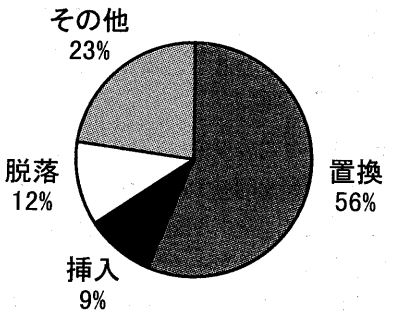

図2　誤入力の分類2（文字単位）

分類2の調査から次の点が明らかとなった。

（傾向1）単一誤りは誤入力の8割弱を占める。

（傾向2）置換誤りは単一誤りの7割を占める。

また、置換誤りについては次の傾向がみられた。

（傾向3）置換誤り中、隣接キーを押下する誤りが多く、シフトキーの押し忘れによる誤りやキーの位置を覚えていないことに起因する誤りも比較的多い。

3. 誤入力訂正方式

3.1. 概要

誤入力の(傾向1)に基づき、本方式では次の誤入力を訂正対象とする。

- かな入力モードの入力
- 分類2における置換、挿入、脱落誤り
- 文節内に誤りを一つ含む文節（単一誤り）

本誤入力訂正方式は、誤りを許容した文節を生成するように、かな漢字変換の文節生成処理を拡張したものである。

本かな漢字変換の全体の構成を図3に示す。

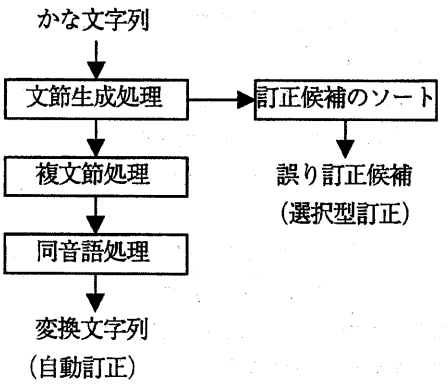

図3 全体の構成

図3に示すように、本かな漢字変換は、文節生成処理で得られた複数の文節候補を優先順に出力する選択型訂正と、複文節処理と同音語処理の後に出力する自動訂正の2つのインターフェースを実装する。

3.2. 文節の生成

文節を次のように定義する。

(接頭語) 自立語 (接尾語) (語尾) 付属語*

() は省略される場合があることを示す。

* は省略または1つ以上の列を示す。

文節生成処理では、上記を満たし、かつ単一誤りを許容する文節を生成する。以下、手順を示す。

(1) 誤り許容検索：形態素内に一つの誤りを許容して辞書を検索する。

(2) 文法処理：文節内文法に基づき、形態素間の接続を検定する。

(3) 単一誤りチェック：文節内の誤り数を1以下（単一誤り）に制御する。したがって、例えば自立語と付属語の双方を訂正した文節は生成しない。

3.3. 誤りを許容する辞書検索

辞書検索の方法は、逆編集最小距離法による。すなわち、まず、入力かな列から訂正文字列を生成し、次に、生成されたかな文字列が単語として正しいか否かを辞書で調べる。

文字列C1, C2, .., Ci,..において文字位置Ciに誤りを含むと想定したときの文字列の訂正操作は次のようになる。

(1) 置換誤り：誤入力文字Ciを正しい文字Csと置き換える。（置換操作）

(2) 挿入誤り：挿入文字Ciを削除する。（削除操作）

(3) 脱落誤り：脱落文字CdをCi-1とCiの間に挿入する。（挿入操作）

以上の操作を行い、長さnの文字列に対する検索すべき見出し語の数は以下となる。

(1) 置換操作　Ns(C1) + Ns(C2) + ... + Ns(Cn)

(2) 削除操作　n

(3) 挿入操作　Ni × (n+1)

Ns(Ci)は、i番目の文字Ciに置換誤りがあると想定した場合の置換文字の数である。また、Niは、脱落誤りを想定したときの挿入文字の数である。置換文字は訂正操作する文字毎の文字集合として保持し、挿入文字は共通の文字集合を保持する。前者の文字集合を置換テーブルと呼び、後者の文字集合を挿入テーブルと呼ぶ。

なお、キー操作履歴より誤入力を採取したが、その絶対数が少ないため十分な傾向をつかめていない。そこで、採取データに特化されるのを避けるために、置換、挿入テーブルにほぼ全ての対象文字を登録している。対象文字87文字に対し、Ns(Ci)は平均31文字、Niは80文字である。Niに比べてNs(Ci)が少ないのは、1ストロークの誤りに限定すると、次の場合に置換文字数が少なくなるためである。

- シフトキーとともに入力された文字が置換誤りとなるのは、不要なシフト、またはシフトキーとともに誤った文字キーを押下したかのいずれかである。前者は1文字、後者の場合、文字として入力されるのは「ぁぃぅぇぉゃゅょっを」のみである。例えば「ぁ」に対する置換パターンは「あぃぅぇぉゃゅょっを」の10文字である。
- 濁音（半濁音）が置換誤りとなるのは、不要な濁点（半濁点）の入力、または濁点（半濁点）の直前に誤った文字キーを押下したかのいずれかである。前者は1文字、後者の場合、文字として入力されるのは濁音文字（半濁音文字）のみである。

Nsを一定値31, Niを80としたとき、検索すべき見出し語の数を表1に示す。

表1　見出し語の数

文字長n	2	4	6	8
見出し語数	305	531	757	983

以上のように辞書検索の回数は膨大になる。そこで、探索時間を削減するために辞書をダブル配列[5]で構成した。ダブル配列は概念的にはtrie構造の状態遷移ネットワークであり、ノードを辿ることで探索がなされる。検索かな列の先頭位置から訂正位置までのノードの探索は1回ですみ、また、次の探索文字にマッチするノードがない場合には探索を打ち切ることができるので、高速な検索が可能となる。

3.4. 候補数の削減

本方式では候補数の増大はさけられない。そこで、辞書検索時に制約を設定する。

誤入力の場合、本来入力しようとした形態素列中、誤入力を含む形態素が分断されるので、この形態素に対応する誤入力文字列を検索すると、見出し語長の短い形態素が検索されると考えられる。もちろん偶発的に、見出し語中に誤入力を含む誤った形態素が検索されることもあるが、誤った形態素の見出し語長が、本来入力しようとした形態素よりも長くなる可能性は低いと考えられる。以上より[制約1]を設定する。

[制約1] 変形したかな列より得られた形態素の文字長は、変形なしに得られた形態素の最長の文字長以上とする。

この他に次の制約を設定する。

[制約2] 変形したかな列より得られた自立語（語幹）の文字長は2以上とする。

これは、1文字の自立語は信頼性が低いにもかかわらず、すべての文字位置に置いて多数の形態素を検索することになるためである。

[制約3] 検索語数に上限を設定する。

信頼性の高い形態素が削除されることのないように、検索語に優先順位を設定する。

優先順位付けには次の情報を使う。

- 文節長に基づく値：長い方を優先
- 誤りの種類：置換誤りを優先
- 置換誤りのうち隣接キーを押下する誤りを優先

4. 誤り訂正インターフェースとその実装

4.1. インターフェース

誤入力を許容する文節生成処理をかな漢字変換に組み込み、二つの誤り訂正機能を実装した。

選択型訂正は、ユーザの要求に応じて訂正文節を複数提示し、ユーザが正しい訂正文節を選択する機能である。また、自動訂正は、かな漢字変換時に自動的に誤入力を訂正する機能である。

選択型訂正と自動訂正の一例を図4に示す。

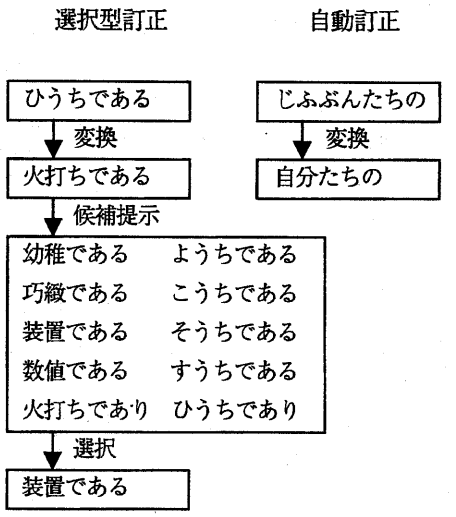

図4 選択型訂正と自動訂正の例

4.2. 選択型訂正の実装

誤りを許容した文節を生成し、優先度の高い複数の文節を出力するものである。複文節処理（文節区切り処理）と同音語処理で適用する各種ルールは使わず、生成された文節の情報のみに基づいて優先順位付けする。

優先順位付けには次の情報を使う。

- 文節長に基づく値：長い方を優先
- 誤りの種類：置換誤りを優先

4.3. 自動訂正の実装

誤りを許容した文節を生成し、複文節処理と同音語処理を実行するものである。複文節処理では、（二文節最長一致法に基づく）各種ルールを、訂正された文節と未訂正の文節の双方に適用し、最終的に最も優先度の高い文節列を得る。この時点で、かな文字列は唯一に決定され、訂正された文節があればこれを誤入力とみなす。

5. 実験

5.1. 選択型訂正の累積再現率

(1) 実験条件

以下の条件で上位20位までの累積再現率を調べた。

- 評価例文：分類2による単一誤りのうち、

記号類の誤入力を除いた文節
(誤入力文字が「かな」または「長音」の文節)

表2　例文の情報

分類	全体	置換	挿入	脱落
文節数	191	130	24	37
異なり数	149	117	24	8

脱落誤りでは「ワードプロセッサとは」を「ワープロセッサとは」とする誤入力が頻発したため、異なり数が少ない。

・ 辞書中の自立語の語彙数：約70000語
・ 再現率の定義

$$\text{再現率} = \frac{\text{正しく訂正された文節数}}{\text{単一誤りを含む文節数（例文数）}}$$

(2) 実験結果

実験結果を図5に示す。

・ 第1位だけで39%が正しく訂正でき、上位6位までで59%の正解が得られる。
・ 置換、挿入誤りに対する訂正精度は比較的よいが、脱落誤りに対する訂正精度が低い。これは、脱落誤りより優先度の高い置換誤りを想定した文節が多数生成されるためである。

5.2. 自動訂正の再現率

(1) 実験条件

選択型訂正の累積再現率の実験条件と同じ条件で、自動訂正の再現率を調べた。

(2) 実験結果

実験結果を表3に示す。

表3：自動訂正の再現率（%）

全体	置換	挿入	脱落
44	52	71	0

・ 置換誤りの52%を正しく訂正できた。
訂正に失敗した59文節のうち、22例は誤入力文字列が文節として成立し、20例は正しく訂正文節を生成しているが他の訂正文節が優先された。
・ 挿入誤りの71%を正しく訂正できた。
訂正に失敗した7文節のうち、4例は文節として成立する。
・ 脱落誤りは訂正できなかった。
訂正に失敗した37文節中、35例が文節として成立し、残り2例は未登録語である。

5.3. 自動訂正の誤り率

自動訂正では、ユーザが正しい文字列を入力したときに、誤って訂正することは避けねばならない。そこで誤りを含まない例文を入力して自動訂正を実行し、誤り率を調べた。

(1) 実験条件

・ 評価例文：誤まりを含まないかな文字列
・ 誤り率の定義

$$\text{誤り率} = \frac{\text{誤訂正された文節数}}{\text{総文節数（7356文節）}}$$

(2) 実験結果

誤り率は3%であった。
誤訂正の原因は主に未登録語による。

6. 考察

選択型訂正では、上位6位までに6割の正解が得られるので、訂正候補の提示・選択操作により入力効率の向上に効果があるものと思われる。実際に使用したときの評価が課題である。

我々が採取した誤入力データはその絶対数が少ないため十分な誤り傾向を抽出するに至っていない。このため、置換誤り、脱落誤りの訂正に多数の文字を割り当て、採取した誤入力へ特化されるのを抑制している。多数の誤入力データを詳細に分析し、分析結果を積極的に利用することにより選択型訂正の性能をさらに向上させることが可能と考えられる。

自動訂正では、再現率が多少低下しても誤り率を抑えることが重要である。実験では、再現率44%、誤り率は3%であるので、文節単位に変換する場合、入力効率の改善が期待できる。ただし、複文節単位の変換時には誤り率が大きくなる。直接的な原因は、複文節処理の各優先付けルールが訂正文節を考慮していないためである。優先付け

ルールの調整、または適切な訂正範囲を設定し、複文節単位の誤り率を削減するのが課題である。

なお、本方式は比較的高速に動作可能である。再現率評価用の例文を使用しワークステーション（SPARC 110MHz）で測定したところ、選択型訂正は平均 16msec／文節、自動訂正は平均 20msec／文節であった。

7. まとめ

入力操作全体を含めた効率を向上させるために、かな漢字変換上に誤入力を訂正する機能を試作した。

本かな漢字変換は、文節内に置換、挿入、脱落誤りのいずれかを一つ含むかな文字列を訂正する。ユーザの訂正指示操作に応じて誤り訂正候補を提示し選択する方法（選択型訂正）と、変換操作に応じて自動的に誤りを訂正する方法（自動訂正）の２つのインターフェースを想定し性能を評価した。実験の結果、選択型訂正の再現率は第 1 位が 39%、上位 6 位までで 59%であった。また自動訂正の再現率は 44%、誤り率は 3%であった。

本誤り訂正機能を使うことにより入力効率の改善が期待できる。

参考文献

[1] Damerau, “A technique for computer detection and correction of spelling errors”, Communications of the ACM, Vol.7,No.3, pp171-76(1964)

[2] 野田, “誤打鍵特性の調査と分析”, 情報処理学会第 47 回全国大会, 1W-3(1993)

[3] 荒木他,, “2 重マルコフ連鎖モデルによる日本語文の誤り検出ならびに訂正法”, 情報処理学会自然言語処理研究会, 97-5, pp29-35(1993)

[4] 村上他, “日本語文音節入力に対して 2 重マルコフ連鎖モデルを用いた漢字仮名交じり文節候補の抽出精度”, 電子情報通信学会論文誌, D-2,Vol.J75-D-2,No.1,pp.11-20(1992)

[5] 青江他, “パターンマッチングマシンの効率的記憶検索法”, 情報処理学会論文誌, Vol.24,No.4,pp.414-420(1983).

[6] Kukich, “Techniques for Automatically Correcting Words in Text”, ACM Computing Surveys, Vol.24, No.4,pp.377-439(1992)

図5 選択型訂正の累積再現率