自然言語処理 129−10
(1999. 1. 21)

予測に基づく入力支援機能を備えたかな漢字変換システムの開発

市村 由美　齋藤 佳美　木村 和広　平川 秀樹

(株) 東芝 研究開発センター
〒210-8582　川崎市幸区小向東芝町1
yumi.ichimura@toshiba.co.jp

あらまし　本論文では、「入力予測」と呼ぶ支援機能を備えたかな漢字変換システムの開発について述べる。本システムは、入力された読み情報からの予測とかな漢字変換とを組み合わせたものである。特定の指示キーなしに自動的に予測が起動され、提示された予測候補が望むものでなかったとしても取消操作を必要としないので、ユーザは通常のかな漢字変換システムと同じ感覚で利用することができる。本システムの予測方式は、(1) 入力文字列の任意の位置から予測を開始する、(2) システム辞書とユーザ辞書を用い、システム辞書はコーパス中の出現頻度を元に計算した「確信度」の情報を持つ、(3)「確信度」と「有用度」に基づき候補を評価し、確からしく有益と思われるものだけを候補として提示する、といった特徴を有する。3種類のテキスト (ビジネス文、手紙、新聞) を用いた評価により、入力操作量を従来の65.2〜75.2%に削減できることを確認した。

キーワード　かな漢字変換、入力支援、予測、確信度、有用度

Kana-Kanji Conversion System with Input Support Method Based on Prediction

Yumi ICHIMURA,　Yoshimi SAITO,　Kazuhiro KIMURA,　Hideki HIRAKAWA

Toshiba R&D Center
1, Komukai-Toshiba, Saiwai-ku, Kawasaki-shi
Kanagawa, 210-8582 Japan
yumi.ichimura@toshiba.co.jp

Abstract　In this paper, we propose kana-kanji conversion system with input support method based on prediction. This system is composed of prediction from kana information and kana-kanji conversion. It automatically shows candidates of character strings which users intend to input, and cancel operation is not needed if candidates do not contain what users want. So, users can use this system in the same manner with ordinary kana-kanji conversion system. The prediction method features: (i)Prediction starts at arbitrary positions of input strings. (ii)System dictionary and user dictionary are used, and each entry of the system dictionary has "certainty factor" calculated from frequencies of corpora. (iii)"Certainty factor" and "usefulness factor" are used to estimate candidates, and only certain and useful candidates are shown. The proposed system could reduce users' key input operations to 65.2-75.2% from original ones in our experiments.

key words　kana-kanji conversion, input support, prediction, certainty factor, usefulness factor

1 はじめに

当社が日本語ワードプロセッサを商品化してから20年が経つ。我々はこれまで日本語入力の基本機能として、かな漢字変換の高精度化に注力してきた。変換精度を向上させることが、ユーザの入力負担の軽減につながると考えたからである。しかし、このような取り組みもキーを探しながら入力する初心者ユーザにとってはまだ不十分であり、キー入力量そのものを軽減する仕組みが必要であった。

キー入力量を削減してユーザの負担を軽減するツールとしては、UNIXシェルのヒストリやGNU Emacsのコンプリーションが広く利用されている。また、Darraghらは、以前に入力されたテキストを用いて、自由文の予測を行うシステムReactive Keyboardを提案している[1]。このシステムは、過去に入力されたテキストの中から入力中の文字列を含む部分を予測ウインドウに表示し、ユーザはそれを選択することで入力を効率化できる。このシステムで文字列予測に用いているn-gram方式は、字種が多い日本語に対しては単純には適用できない。

一方、日本語に関する研究としては、福島らによる予測ペン入力インタフェース[2]や、増井によるペン計算機向けの文章入力システムPOBox[3][8]が提案されている。予測ペン入力インタフェースでは、ペンで文字を書くと、書き込みに一定以上の時間があいた時点で、次に入力されるであろう文字列が予測表示される。その次の文字を書き込む位置により、表示された予測文字列を選択するかどうかが判定される。このシステムの予測方式では、手書き入力された文字を末尾とした同一字種文字列の先頭を予測の開始位置としている。キーボード入力では、ペン入力よりも一般に速い速度で入力されるので、いつ予測表示を行うのか問題になる。また、読み情報しか入力されないので、どこから予測を開始するのか問題になる。

POBoxでは、ソフトキーボードを使用して読みを入力すると、その読みで始まる候補単語の集合が提示され、その中から単語を選択することで文を入力していく。POBoxの予測方式では、2単語の連接からなる文例辞書と単語辞書の2つを利用する。どちらの辞書も、選択した文例や単語が辞書の先頭に追加されるようになっており、辞書の先頭から順番に読みとマッチングすることで、候補集合を提示する。ソフトキーボードによる入力速度は一般にそれほど速くないので、選択速度が入力速度を上回ると思われ、この方式は効果的である。一方、通常のキーボードでは、入力速度が選択速度を上回る場合が多いと思われ、必ずしも効果的とは言えない。また、ソフトキーボードでは、読みの入力と候補の選択が同一画面内での同様な操作で行えるのに対し、通常のキーボードでは、1文字入力するごとに予測候補が提示されるとキートップと画面の両方を交互に注視しなくてはならなくなり、かえって煩わしさが増す恐れがある。

これらに対して、我々は「入力予測」と呼ぶ支援機能を備えたかな漢字変換システムを開発した。本システムは、入力された読み情報からの予測とかな漢字変換とを組み合わせたものである。特定の指示キーなしに自動的に予測が起動され、提示された予測候補が望むものでなかったとしても取消操作を必要としないので、ユーザは通常のかな漢字変換システムと同じ感覚で利用することができる。これを利用すれば、キーボード初心者も日本語入力を効率的に行えるようになる。

このようなかな漢字変換システムを開発する上で、大きく3つの課題があった。

(1) 自動的に予測を行う場合、どの部分から予測を開始するのか。

(2) ユーザに適応し、かつ、精度の良い日本語の予測文字列をどう生成するのか。

(3) 予測候補を選択するだけで入力を行うことが前提ではないため、予測可能なものがあるときに常に候補を表示するのでは、使いやすいシステムにはならない。どのタイミングで候補を提示するのか。

我々はこれらの課題に対して、次のようなアプローチを取った。

(1) 現在のかな漢字変換では連文節変換方式が広く利用されており、ユーザは句や文節ごとに読みから漢字かな混じり文字列に変換しているとは限らない。よって、予測開始位置を入力の先頭に限定すると、入力方式によっては予測が起動される場面が非常に限定されてし

まうため、入力中の読みの任意の位置から予測を開始する。

(2) システム辞書とユーザ辞書を用いる。システム辞書の各見出しにはコーパス中の出現頻度を元に計算した**確信度**の情報を与える。

(3) **確信度**と**有用度**に基づき候補を評価し、確からしく有益と思われるものだけを候補として提示する。

以下、2. でシステムの使用例、3. で機能の実現方法について述べ、4. でキー入力操作量の削減効果を示す。

2 システムの使用例

システムの概要を、例を通じて示す。図 1 に本システムを用いた文章入力例を示す。「かきのかいぎをかいさいしますのでごさんしゅうねがいます」という文を入力例とする。

「か」「き」と順に読みを入力していき、「の」まで入力されると、自動的に入力予測候補ウインドウが開き、「下記の住所にささやかながら」「下記の住所にささやかな」という 2 つの候補が提示される (図 1(a))。1 番目の候補が反転表示されている。望む候補がある場合にはカーソルで選択し、ない場合は読みを続けて入力する。ここでは、望む候補がないので続けて「か」と入力すると、ウインドウが消える (図 1(b))。さらに、「い」「ぎ」と入力すると、再び入力予測候補ウインドウが開き、「会議を開催しますのでご参集願」「会議を開催します」「会議を行います」「会議を執り行」の 4 つの候補が提示される (図 1(c))。1 番目の候補が望む候補であるので選択キーを押すと、ウインドウが消え、入力領域に候補が挿入される。さらに、「かきの」の部分が「下記の」と自動的にかな漢字変換される (図 1(d))。最後に、「い」「ま」「す」と語尾を入力する (図 1(e))。

通常のかな漢字変換システムでは、27 文字の読みを入力する必要があるが、本システムでは「をかいさいしますのでごさんしゅうねが」の部分の 18 文字の入力を省略でき、9 文字の入力で済む。

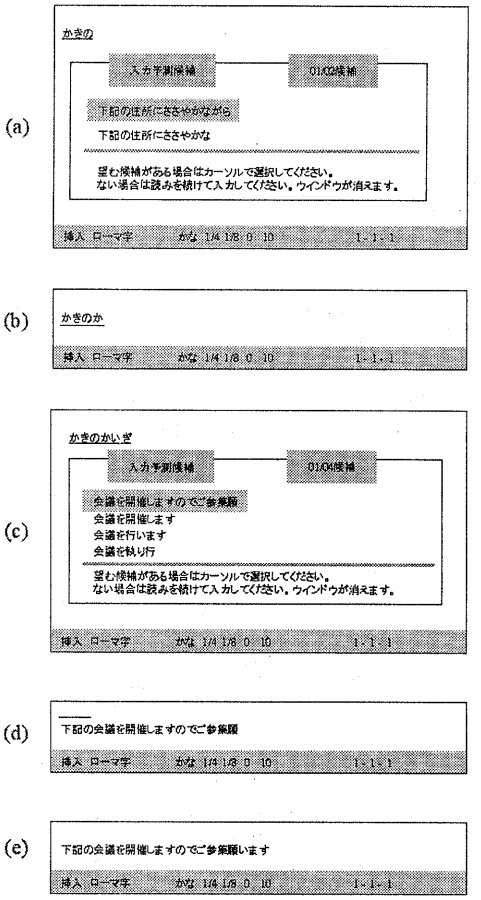

図 1: 文章入力例

3 機能の実現方法

3.1 システムの概要

図 2 にシステムの概略図を示す。ユーザの入力した読みを以下のように処理し、かな漢字変換・予測候補を生成してユーザに提示する。

(step1) 検索文字列の生成 : 入力読み文字列から予測辞書を検索するための検索文字列群を生成する。長さ l の入力読み文字列に対して、その文字列の最後尾の文字を含む長さ 2 以上の $l-1$ 個の部分文字列を生成し、それらを検索文字列とする。

(step2) 予測辞書の検索 ：各検索文字列で辞書を前方一致検索する。

(step3) 予測候補の評価 ：各検索結果に評価値を付与し、その検索結果を候補として提示するか否かを決定し、提示すべき候補の順位付けを行う。

(step4) かな漢字変換の実行 ：予測開始位置の前側に入力読み文字列が残っている場合には、かな漢字変換を行う。

(step5) 候補提示 ：かな漢字変換および予測候補を提示する。

(step6) 学習 ：提示された候補をユーザが選択し確定すると、ユーザ辞書に自動登録する。

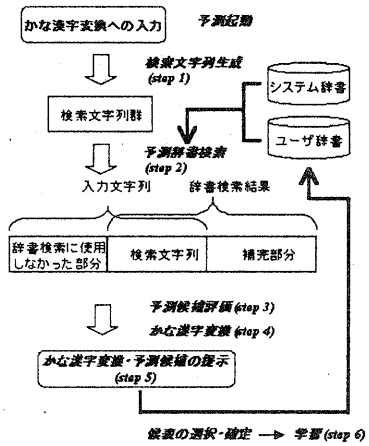

図 2: システムの概略図

3.2 予測辞書の構成

予測の情報源としては、システム辞書とユーザ辞書の 2 種類を用意し、各辞書には、単語、共起、慣用句、文などを区別なく格納する。

システム辞書 ：一般に使用頻度の高い表現を格納する。

ユーザ辞書 ：ユーザが過去に作成した文書や入力履歴から学習した表現を格納する。

辞書の形式を表 1に示す。cfは、3.3 で述べる確信度を示す情報である。

表 1: 辞書の形式

システム辞書	読み; 表記; 品詞;cf;
ユーザ辞書	読み; 表記; 品詞;

3.3 予測候補の評価方法

予測辞書から検索された候補を評価するために、**確信度** と**有用度**の 2 つの評価尺度を用いる。候補の順位づけは確信度と有用度を総合して行い、確信度と有用度が一定の値以上のものを候補とする。

確信度 ：候補が入力時点でどれだけ確からしいかを示す値。確信度はシステム辞書とユーザ辞書に分けて計算する。具体的な計算方法は 3.4 に示す。

有用度 ：候補が提示されたときの役に立つ度合を示す値。有用度は候補中の補完される読み文字列の長さに基づく。この有用度は検索文字列長に応じて変化するので、有用度が一定の値以上のものを候補とすることで、候補提示のタイミングを調整できる。

また、入力の途中からも予測を開始する方式を採用しているため、予測開始位置を考慮しなければ確からしい候補であっても、開始位置によっては文法的に不自然な場合がある。これを防ぐために候補の文法的妥当性を評価し、文法的に誤っているものは、確信度と有用度の値にかかわらず候補としない。

3.4 確信度の計算方法

3.4.1 システム辞書

文字列の結合強度は、単語内でもっとも強く、単語間、文節間の順に弱くなると考えられる点に注目して、コーパス中の文字連鎖確率を用いてシステム辞書の各見出しに対する確信度 (cf) を、以下の式で計算する。

$$cf = \frac{A}{B}$$

A = ある漢字表記コーパスにおけるその表現の出現頻度

B = 対応する読み表記コーパスにおけるその表現を検索した検索文字列の出現頻度

「かな漢字変換」という語を例に、確信度の計算例を示す。筆者が過去に作成した文書 (読み 22 万文字) 中の文字の出現頻度を表 2に示す。これから、「かな漢字変換」の確信度は、「かな」と入力した時点では 70/191=0.366、「かなか」と入力した時点では 70/114=0.614 と計算される。

辞書中には検索文字列ごとの確信度をあらかじめ計算して記述しておき、辞書検索の段階で検索文字列長に対応する確信度を辞書中から読み出す。

表 2: 確信度の計算例

文字列	出現頻度	
	読み表記ファイル中	漢字表記ファイル中
か	6720	
かな	191	
かなか	114	
かなかん	94	
かなかんじ	87	
かなかんじへ	78	
かなかんじへん	77	
かなかんじへんか	76	
かなかんじへんかん	76	
かな漢字変換		70
仮名漢字変換		5
カナ漢字変換		1

3.4.2 ユーザ辞書

ユーザ辞書では、登録される表現がダイナミックに変わり、また表現の出現頻度もあまり多くならない。そのため、システム辞書と同様な方法で確信度を計算するには問題がある。

ユーザ辞書の確信度を計算するために、筆者が過去に作成した文書 (読み 78,000 文字) を 6 等分し、13,000 文字ずつ入力していった場合を想定してユーザ辞書を作成し、その読みと出現頻度を調べた。

確信度を、長さ n の検索文字列により検索される平均語彙数の逆数で表した。図 3のグラフは、X 軸が検索文字列長 n、Y軸が入力テキスト量 m(単位:万文字) における確信度を示している。m により幅はあるが、同様な傾向の増加曲線を描いている。このグラフから m に応じて確信度を決定する。

システム辞書、ユーザ辞書ともに、この確信度は検索文字列長に応じて変化するので、確信度が一定の値以上のものを候補とすることで、候補提示のタイミングを調整できる。

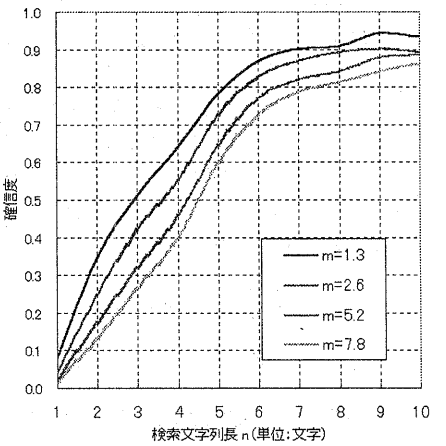

図 3: 検索文字列長と確信度の関係

3.5 学習の方法

ユーザが選択・確定した単語と単語間の共起をユーザ辞書に自動的に格納する。単語は、読みの長さが一定の値以上の場合、また、共起は、あらかじめ設定されたパタンを満たす場合に、格納する。

たとえば、「かいぎにしゅっせきする」を入力例とする。「か」「い」「ぎ」と順に読みを入力すると、図 1(c) に示すような 4 つの候補が提示される。しかし、ここには望む候補がないので、「にしゅっせきする」と入力を続け、かな漢字変換キーを押すと、「会議に出席する」と変換される。これをユーザが確定すると、「会議」「出席する」という単語と、「会議に/出席する」という共起が学習され、ユーザ辞書に自動登録される。すると、次に「かいぎ」と入力したときには、「会議に出席する」が候補として提示されるようになる。

4 評価実験

予測を利用した場合に、ユーザのキー入力操作がどれくらい削減されるかを評価した。本システムの方式では、予測を起動させるためのキー操作は不要であり、また提示された予測候補が望むものでなかったとしても取消操作は必要ない。一方、提示された予測候補を選択するためには、選択キーの操作が必要である。しかし、その部分を通常のかな漢字変換により入力した場合にもかな漢字変換候補に対する選択キーの操作が必要になるため、選択キー操作の増減はほぼ相殺されると考え、カウントしない。したがって、予測補完される文字数がそのまま入力操作の削減量になると考えられる。

そこで、もとの操作量に対してどれだけの割合の操作量で入力が可能になるかを示す値として、以下に定義する操作比率を用いる。

$$\text{操作比率} = \frac{A-B}{A} \times 100 \quad (\%)$$

A = 本来入力が必要な読み文字数

B = 予測により補完される読み文字数

4.1 入力のシミュレーション

予測による効果を求めるためには、予測補完される読み文字数を数えればよい。そこで、今回の評価は、人間が対話的にシステムを操作するのではなく、読みテキストファイルと、正解文書としてそれに対応する漢字テキストファイルとを与えて、予測補完される文字数を自動的に計算するプログラムを作成して行った。

4.2 評価条件

評価には、表3に示すビジネス文、手紙、新聞の3種類のテキストを用いた。これはシステム辞書の確信度を計算したコーパスには含まれないデータである。システム辞書は登録数37,926件の辞書を用いた。確信度が0.1、有用度が2以上のものを候補として提示し、候補ウインドウに表示する候補数は最大5個とした。

予測による入力操作量の削減効果を測定するとともに、辞書と学習の各々の寄与についても分析する。そのため、各テキストに対し、以下の条件で測定を行う。

表3: 評価に用いたテキスト

テキスト	読み文字数
ビジネス文	2369
手紙	950
新聞	2064

(A) 学習せずに、システム辞書のみで予測した場合。

(B) システム辞書を用いずに、学習のみで予測した場合。

(C) システム辞書と学習の両方を用いて予測した場合。

ただし、ここでの学習とは、評価テキストからあらかじめユーザ辞書を作成してシミュレーションを行ったものである。よって、一度入力したテキストを再度入力した場合と見なせる。表4に各テキストに対して学習されたユーザ辞書の登録内容を示す。登録件数は、ビジネス文では228件、手紙では69件、新聞では191件であった。テキストの大きさにばらつきがあるため、100文字あたりの登録件数を計算したところ、ビジネス文では9.6件、手紙では7.3件、新聞では9.3件であった。また、登録された見出し1件あたりの読みの長さを計算したところ、ビジネス文では8.7文字、手紙では7.7文字、新聞では8.4文字であった。

表4: ユーザ辞書の登録内容

テキスト	登録件数	100文字あたりの登録件数	1件あたりの読みの長さ
ビジネス文	228	9.6	8.7
手紙	69	7.3	7.7
新聞	191	9.3	8.4

4.3 結果と考察

(A)(B)(C) の3通りのケースで測定した操作比率を表5に示す。

入力操作量の削減効果：システム辞書と学習の両方を用いるケース (C) では、ビジネス文で

表 5: 予測を利用した場合の操作比率 (%)

テキスト	(A) 辞書のみ	(B) 学習のみ	(C) 辞書+学習
ビジネス文	85.9	70.3	65.2
手紙	82.0	84.9	73.8
新聞	95.0	76.1	75.2

65.2%、手紙で 73.8%、新聞で 75.2%の操作比率を得られた。本システムは日本語入力効率の向上に効果があることを確認できた。

システム辞書の効果 : 学習を用いずにシステム辞書だけで予測を行うケース (A) では、手紙 < ビジネス文 < 新聞 の順に操作比率が小さくなっている。手紙にはコーパスでの出現頻度の高い定型文がよく使われるため、最も効果が大きいと考えられる。一方、新聞では多彩な表現が利用されるため、システム辞書だけでは対応しきれず、効果が少ないと考えられる。システム辞書は定型文書に対して有効であると言える。

学習の効果 : システム辞書を用いずに学習だけで予測を行うケース (B) では、ビジネス文 < 新聞 < 手紙の順に操作比率が小さくなっている。表 4に示したユーザ辞書の登録内容がその要因の一つと考えられる。テキスト 100 文字あたりの登録件数は、ビジネス文 > 新聞 > 手紙の順に多くなっており、さらに、1 件あたりの読みの長さもビジネス文 > 新聞 > 手紙の順に長くなっている。これは削減効果の大きさの順と同じ傾向である。つまり、より長い見出しがより多く学習されているテキストほど、学習の効果が高くなっていると言える。

システム辞書と学習の寄与の比較 : システム辞書を用いずに学習だけで予測を行うケース (B) と、システム辞書と学習の両方を用いるケース (C) では、操作比率の差は 0.9～11.2%であった。一方、学習を用いずにシステム辞書だけで予測を行うケース (A) と、システム辞書と学習の両方を用いるケース (C) では、操作比率の差は 8.2～20.7%であった。したがって、この実験ではシステム辞書に比べて学習の寄与の方が大きくなっている。

5 まとめ

「入力予測」と呼ぶ支援機能を備えたかな漢字変換システムを開発した。本システムは以下のような特徴を有する。

(1) 特定の指示キーなしに自動的に予測が起動され、また提示された予測候補が望むものでなかったとしても取消操作を必要としないので、通常のかな漢字変換と同じ感覚で利用できる。

(2) 入力中の読みの任意の位置から予測を開始するので、連文節変換方式の場合にも利用できる。

(3) 検索文字列長に応じて変化する確信度と有用度に基づき候補を評価し、確信度と有用度が一定の値以上のものだけを候補とすることにより、候補提示のタイミングを調整し、確からしく有益と思われるものだけを候補とする。

(4) 候補の文法的妥当性を判定することにより、文法的に誤っているものは確信度と有用度の値にかかわらず候補としない。

(5) ユーザが過去に作成した文書や入力履歴から単語と単語間の共起を抽出して、ユーザ辞書に自動登録する。

3 種類のテキストを用いた評価により、入力操作量を従来の 65.2～75.2%に削減できることを示した。今後は、入力された読み文字列に加えて、分野や文脈の情報を利用することで、予測の精度向上を図っていく予定である。

参考文献

[1] John J. Darragh, Ian H. Witten, and Mark L. James: The Reactive Keyboard: A Predictive Typing Aid, *IEEE Computer*, Vol.23, No.11, pp.41-49, November 1990.

[2] 福島俊一, 山田洋志: 予測ペン入力インタフェースとその手書き操作削減効果, 情報処理学会論文誌 Vol.37, No.1, pp.23-30, 1996.

[3] 増井俊之: ペンを用いた高速文章入力手法, インタラクティブシステムとソフトウェア IV: 日本ソフトウェア科学会 WISS'96, pp.51-60, 近代科学社, December 1996.

[4] 杉本正勝: 片手操作キーカード (SHK) による日本語入力, 情報処理学会 モーバイルコンピューティング研究会 1-1, 1997.

[5] Alice Carlberger, Johan Carlberger, Tina Magnuson, Sira E. Palazuelos-Cagigas, M. Sharon Hunnicutt and Santiago Aguilera Navarro: Profet, A New Generation of Word Prediction: An Evaluation Study, *Proceedings of the ACL Workshop on Natural Language Processing for Communication Aids*, pp.23-28, July, 1997.

[6] Nestor Garay-Vitoria and Julio G. Abascal: Word Prediction for Inflected Languages. Application to Basque Language, *Proceedings of the ACL Workshop on Natural Language Processing for Communication Aids*, pp.29-35, July, 1997.

[7] 増井俊之: 動的パタンマッチングを用いた高速文章入力手法, インタラクティブシステムとソフトウェア V: 日本ソフトウェア科学会 WISS'97, pp.81-86, 近代科学社, December 1997.

[8] Toshiyuki Masui: An Efficient Text Input Method for Pen-based Computers, *Proceedings of the ACM Conference on Human Factors in Computing Systems (SHI'98)*, Addison-Wesley, pp.328-335, April 1998.