

インターネットを用いた日本語入力システム

奥野陽[†] 萩原将文^{††}
[†]慶應義塾大学理工学研究科

本論文では、インターネットを用いた日本語入力システムを提案する。従来の日本語入力システムはインターネットが普及する以前の状況を前提に開発されてきた。一方で、近年の Web アプリケーションの台頭に見られるように、クライアントサイドの機能をサーバーサイドに移す動きが顕著である。提案システムはインターネットのメリットを最大限に活用するため、インターネットを通してサーバー側で変換を行う。インターネットを用いることで、次のような利点がある。(1) Web 上の文章から抽出された大規模な統計量を用いることができる。(2) サーバーサイドの豊富なハードウェアリソースを利用できる。(3) ユーザーが登録した単語を共有することで、専門用語や流行語などの単語を変換できるようになる。提案システムの評価を行うため、初心者ユーザーを想定して文章を入力する評価実験を行った。実験の結果では、提案システムは Microsoft Office IME 2007 と比べて入力時間が平均 21%、キータイプ数が平均 26%削減された。

Japanese Input Method based on the Internet

YOH OKUNO[†] MASAFUMI HAGIWARA^{††}
[†]Keio University

In this paper, we propose a Japanese input system based on the Internet. The advantages of usage of the Internet are the following three merits; (1) The large-scale statistic extracted from the Web can be used; (2) the rich hardware resource of the server-side can be used; (3) the words such as a technical term or the vogue word can be converted by sharing the words that users registered. From the result of the experiments, as for the proposed system, an average of 21% in input time and 26% in the number of the key types were reduced in comparison with the Microsoft Office IME 2007.

1. はじめに

近年メールやブログの普及により、初心者も PC で日本語の文章を入力する機会が増えている[1]。しかし PC 初心者にとってかな漢字変換は大きな負担となる[2]。また初心者だけでなく、高齢者や体の不自由な方にとっても、かな漢字変換は大きな障害である。

初心者が日本語を入力するための補助をする方法として、予測変換が注目されている[3]。Window においては、Microsoft Office IME 2007 やジャストシステム社の ATOK 2008 等、広く用いられているソフトウェアが予測変換機能を提供している。これらの日本語入力システムにおける予測変換機能は、ユーザーが過去に入力したことのある単語を提示する。また、あらかじめ設定された定型文を予測候補とするものもある[4]。しかしながら、このような方式の予測変換を有効に利用できるケースは少ないと考えられる。このような現状において、PC 上で動作する予測変換システムを構築する事は、大きな意義があると考えられる。

一方、近年のブロードバンドの一般化により Web アプリケーションの形態をとるソフトウェアが台頭している[5]。Web アプリケーションには、サーバー側にデータや処理系を持つという特徴がある。そのため、複数のユーザー同士でデータを共有する事ができる、どの PC でも同じデータを利用可能であるといったメリットがある。このように、近年ではクライアントサイドの機能をサーバーサイドに移す動きが顕著である。

従来の日本語入力システムは、インターネットが普及する以前の状況を前提として開発されていた。そのため、変換に用いる辞書や単語の使用頻度などのデータは基本的にインストールした時点のものがそのまま使われる。しかし、このような方式の日本語入力システムには次のような3つの問題点がある。

1 番目の問題は、単語の使用頻度のデータが少ないという点である。2 番目の問題は、使用できるリソースが少ないという点である。3 番目の問題は、デフォルトの辞書に入っていない専門用語を変換できないという点である。

以降、第2章でインターネットを用いた日本語入力システムを提案し、第3章で提案システムの変換エンジンで用いる確率モデルについて説明する。第4章で評価実験により提案システムの有効性を確かめ、第5章を結論とする。

2. インターネットを用いた日本語入力システム

以上のような背景から、本論文ではインターネットを用いた日本語入力システムを提案する。インターネットを用いることで、次のような利点がある。

(I) Web 上の文章から抽出された大規模な統計量を用いることができる

Web から抽出された大規模な統計量を用いることにより、かな漢字変換や予測変換の曖昧性を解決する事ができる。この理由について、以下で説明する。

現在, Web 上の文章についての日本語の N グラム統計量が Google によって提供されている[6]. これを Web 日本語 N グラムと呼ぶ. N グラムとは, 文章中で N 個の単語が連続して使われた頻度についての統計量である. Web 日本語 N グラムは, 抽出元の文章は約 200 億文を含み, 抽出されたデータは 100GB にも及ぶ大規模な統計データである.

一方, 日本語入力においては, 同音異義語が複数ある場合や, 単語の境界が曖昧な場合に, 変換結果を一意に決めることができないという問題がある. 予測変換では, 読みが入力されていない部分を予測するため, 特に曖昧性が高い. Web 日本語 N グラムのように大規模な統計量を用いると, 使用頻度の高い言語表現を DB (データベース) の中から検索することができる. これにより, 変換候補の中から適切な候補を選択し, 日本語入力における曖昧性を解決することができる. 特に予測変換では前述のように曖昧性が高いため, 大規模な統計量を有効に活用できると考えられる.

(2) サーバーサイドの豊富なハードウェアリソースを使用することができる

前述のような大規模なデータをクライアント PC に直接インストールし, 常にメモリー上にロードして使用することは現実的ではない. それに対し, 専用のサーバーでデータを扱うことで, ハードディスクやメモリーの容量などのハードウェアリソースを豊富に使うことができる. また, サーバーを複数台用意することで, より多くのデータを扱うことができる.

(3) ユーザーが登録した単語を共有することで, 専門用語や流行語などの単語が変換できる

多くのユーザーが利用するサーバー型の利点を活かして, 登録された単語をユーザー間で共有することができる. これにより, 専門用語や流行語などの, 普通では変換できない単語を変換することができるようになる.

2.1 提案システムの概要

図 1 に提案システムの概要を示す. 提案システムは PC で動作するクライアント側のソフトウェアと, サーバー側のソフトウェアから構成されている. クライアントは入力された読みをサーバーに送信し, サーバーで変換を行って, 変換結果をクライアントで表示する. クライアントは Windows 上の日本語入力システムとして違和感なく使えるよう実装されている. 本論文では, 主にサーバー側の変換アルゴリズムについて説明する.

ここでいう変換とは, 予測変換とかな漢字変換のことである. 予測変換のみでは, 長文が入力された場合

のシステムの対応が困難である. そのため, 予測変換だけでなくかな漢字変換を行い, それぞれの候補が統合されている. かな漢字変換と予測変換の両方において, Web 日本語 N グラムが用いられる.

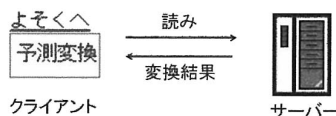


図 1 提案システム の 概念

Fig. 1 The proposed system

2.2 サーバー構成

図 2 にサーバーの構成を示す. サーバーは, Web アプリケーションでは一般的な LAMP (Linux, Apache, MySQL, PHP) と呼ばれる構成を採用している. クライアントとは HTTP 上の Web API で通信を行う. 変換は PHP による変換アルゴリズムと MySQL による DB の検索によって行われる. DB は, 主に語彙 DB と N グラムの DB からなる. N グラムの DB は単語 ID の形式で N グラムとその頻度が格納されている. このように単語 ID で表すことで, DB のサイズを減らすことができる. また, 予測変換では検索に時間がかかるため, キャッシュが行われる.

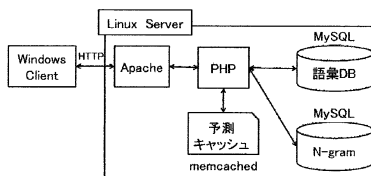


図 2 サーバー構成

Fig. 2 The Design of the Server

2.3 変換アルゴリズム

図 3 に提案システムにおける変換アルゴリズムの概要を示す. まず, かな漢字変換と予測変換の候補を DB から検索する. 次に, 変換候補の統合を行い, 候補の組合せを表現するデータ構造を生成する. 最後に, 生成された変換候補を確率モデルに従ってソートする.

変換候補の検索と統合によって, 入力された読みの変換結果として可能な解が多数生成される. しかし, それらの中には変換候補として適切でないものも多い. そのため, 変換候補をどのような順番でユーザーに提示するかが重要となる. 近年ではかな漢字変換の順位付けのために, 確率論に基づく統計的な手法を用いることが一般的となっている[7]. 提案システムにおいても, 生成された変換候補を順位付けするために, 確率

モデルを用いる。確率モデルと大規模な統計量を組み合わせることで、シンプルかつ強力な変換を行うことができる。提案システムで使用される確率モデルについては、第3章で詳しく説明する。

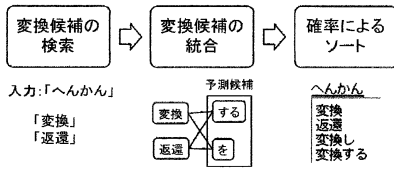


図 3 変換アルゴリズム

Fig. 3 Conversion Algorithm

2.4 変換候補の検索

図 4 に変換候補の検索方法を示す。まずかな漢字変換のために、部分文字列検索を行う。部分文字列検索では、与えられた文字列に含まれる全ての部分文字列を検索する。これにより、かな漢字変換の候補を漏れが無いように列挙することができる。次に、予測変換のために前方一致検索を行う。前方一致検索では、文末を含む部分文字列で検索する。これにより、予測変換の結果として考えられる候補を列挙することができる。これらは N グラム DB から検索された後、語彙 DB を検索して単語の詳細な情報が付与される。

前方一致検索においては、少ない文字数で検索を行うと大きな時間がかかる。これは、検索結果が多くなってしまい、そのソートに時間がかかるためである。そこで提案システムでは、2 文字以下の予測変換のみ、あらかじめソート済みの結果の上位 50 件を予測キャッシュとして保持している。

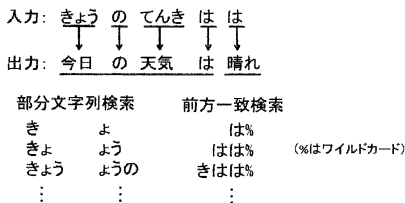


図 4 変換候補の検索

Fig. 4 The retrieval of the candidates

2.5 変換候補の統合

検索された変換候補は、かな漢字変換と予測変換の候補を統合して組み合わせられる。図 5 に変換候補の統合により生成されるデータ構造を示す。かな漢字変換の候補は隣り合う単語同士で結合される。また、予測候補は文末に統合される。このデータ構造により、変換候補の組み合わせをコンパクトに表現することができる。これらの変換候補の組み合わせの中から、確率

の高い経路を探索することによって、かな漢字変換を行う。

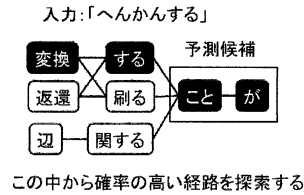


図 5 変換候補の統合

Fig. 5 The uniform the candidates

2.6 変換候補の探索

変換候補の探索では、前述のデータ構造において、最も確率の高い経路を探索する。単純な方法として全探索があるが、これは入力文字列の長さに対し指数オーダーの計算量がかかる方法であり、効率が悪い。動的計画法の一種であるビタビアルゴリズム[8]を用いることにより、余分な経路を省略して線形オーダーで探索を行うことができる。

図 6 にビタビアルゴリズムの概要を示す。ビタビアルゴリズムでは、先頭から探索を行い、続く単語を順に辿っていく。図では左から右に探索することに相当する。このとき、ある単語にたどり着くまでの経路は、確率が最大の経路を1つだけ残すようにする。図では、ある単語の左側に接続される線は最大1つに制限することに対応する。これにより他の経路を省略することができるので、入力の長さに対して線形オーダーにより探索を行うことができる。また、探索された経路は確率が最大となることが保証されている。

実際に提案システムでは、全探索では7文字程度の変換であっても数分かかるのに対し、ビタビアルゴリズムでは0.1秒以内にかな漢字変換を行うことができる。

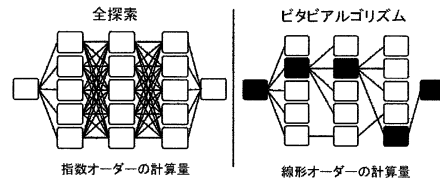


図 6 変換候補の探索

Fig. 6 The search of the candidates

3. 変換エンジンの確率モデル

本章では、変換候補の順位付けに用いられる確率モデルについて説明する。

3.1 かな漢字変換の確率モデル

かな漢字変換の候補の組み合わせの中から正しい候補を選ぶため、確率モデルによる順位付けが行われる。入力された読みを y ，出力する変換候補を x とすると、変換確率は条件付き確率 $P(x|y)$ により表される。つまり、 $P(x|y)$ の値が最大になるように前述のビタビアルゴリズムによって探索を行えば、最も適切と考えられるかな漢字変換の候補が得られる。

このとき、ベイズの定理によって次式が成り立つ。 $P(x|y) \propto P(x)P(y|x)$ (1)

ここで $P(x)$ は確率的言語モデルと呼ばれ、文 x が書かれる確率を表す[9]。 $P(y|x)$ は確率的かな漢字モデルと呼ばれ、文 x が読み y のように読まれる確率を表す。読み y は与えられており全ての候補で同一なので、上式が等式の場合の右辺の分母 $P(y)$ は考慮しなくてよい。すなわち、 $P(x)P(y|x)$ が最大となるよう探索すれば、適切な候補が得られることになる。

提案システムでは、確率的言語モデルは N グラムによってモデル化し、Web 日本語 N グラムを用いて推定される。また確率的かな漢字モデルは単語ごとにモデル化し、McCab[10]によって推定される。

3.2 確率的言語モデル

確率的言語モデルとして、単語 N グラムモデルを用いてモデル化する。 N を大きく取ると、より前後の文脈を考慮することができる。しかし、計算時間の問題から、かな漢字変換においては $N=2$ とした。かな漢字変換では読みが与えられているということが強い制約となるため、 N が小さいことが原因となる誤変換は少ない。しかし、後述するように予測変換では最大 7 グラムまで用いる。

$N=2$ の場合の N グラムは、次式で表される。

$$P(x) = P(w_1) \prod_{i=2}^n P(w_i | w_{i-1}) \quad (2)$$

ここで w_i は文 x を構成する i 番目の単語、 n は文 x に含まれる単語数である。このモデルは、文 x を単語列 $\{w_i\}$ のマルコフ連鎖とみなす。なお、日本語入力では、文単位で変換されるとは限らないため、文頭と文末は区別しない。

連接確率 $P(w_i | w_{i-1})$ および生起確率 $P(w_i)$ は次のように推定される。

$$P(w_i | w_{i-1}) = \frac{C(w_i, w_{i-1})}{C(w_{i-1})} \quad (3)$$

$$P(w_i) = \frac{C(w_i)}{C} \quad (4)$$

ここで $C(\cdot)$ はコーパス中の頻度を表す。具体的には、 $C(w_i, w_{i-1})$ は単語 w_{i-1} と w_i が連続して現れる頻度であり、 $C(w_i)$ は単語 w_i が現れる頻度である。また、 C はコーパス中の全単語数である。このように、 N グラム確率はコーパス中の頻度を用いて推定される。提案システムは頻度として Web 日本語 N グラムのデータを用いる。

3.3 予測変換の確率モデル

これまでに、かな漢字変換の確率モデルが 2 グラムとかな漢字モデルで求まることを説明してきた。ここからは、予測変換の確率モデルについて説明する。

予測変換は、Web 日本語 N グラムの 1~7 グラム全体を候補とし、文末に付与される。かな漢字変換の候補 x が与えられたときの、予測候補の確率は次式で表される。

$$P(w_{n+1}^{n+N-1} | x) = P(w_{n+1}^{n+N-1} | w_n) \quad (5)$$

ここで、 w_{n+1}^{n+N-1} は w_{n+1} から w_{n+N-1} までの $(N-1)$ 個の候補を表す。 N は N グラムの次数である。上式のモデル化は、 N グラムの最初の 1 語をかな漢字変換との接続に用いる文脈とし、残りの $(N-1)$ 語を予測に用いることに相当する。

なお、かな漢字変換の候補 x は、入力された読み全体の変換結果である必要はない。例えば x が空文字の場合は、入力を前方一致検索した結果がそのまま変換候補となる。これは、短い入力に対してはかな漢字変換が行われず、予測変換だけで変換候補を示せることを意味している。また、部分文字列の変換結果に、予測候補を付与することもできる。これにより、入力途中で単語として不完全な読みを入力されても、予測候補を表示する事ができる。

3.4 予測変換のトレードオフ

これまでに説明したモデルにより、かな漢字変換と予測変換の候補を統合し、確率の高い候補を選択することができる。しかし、予測変換においては、確率の高い候補を選択すると問題が生じる。それが、予測変換におけるトレードオフの問題である。

図 7 に予測変換のトレードオフの例を示す。予測変換においては、予測の長さが短いほど当たる確率は大きく、予測候補が長いほど当たる確率は低い。しかし、長い候補の予測ほど、入力せずに済む文字数が大きい。そのため、当たったときのメリットは大きくなる。Web 日本語 N グラムのように最大 7 グラムまで使用できると、かなり先まで予測することができる。しかしながら、どの程度先まで予測するべきか、あるいは、どこで予

測を止めるべきか、について明確な指針はこれまでに存在しないという問題があった。

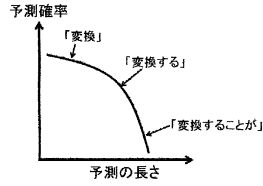


図 7 予測変換のトレードオフの例

Fig. 7 An example of the tradeoff of prediction

3.5 入力時間のモデル化と最小化

前節で説明したトレードオフの問題を解消するために、入力時間のモデル化を行う。そして、入力時間の削減量の期待値で予測候補をソートするという方法を用いた。

入力時間のモデルは予測成功時と予測失敗時に分けられ、次式で表される。

$$T = \begin{cases} \alpha(L - L_0) & \text{if succeeded} \\ -\beta & \text{if failed} \end{cases} \quad (6)$$

ここで T は予測変換による入力時間の削減量であり、 L は変換候補の長さ、 L_0 は入力済みの読みの長さ、 α 、 β は正の定数である。このモデルでは、予測が成功したときの削減量は、予測部分の文字数 $(L - L_0)$ に比例する。 α は 1 文字の入力にかかる時間である。一方、予測が失敗した時は β だけ時間のロスがあるものとしている。

このとき T の期待値 $E[T]$ は、変換確率 P を用いて次のように求まる。

$$\begin{aligned} E[T] &= \alpha(L - L_0)P - \beta(1 - P) \\ &= \alpha \left(L - L_0 + \frac{\beta}{\alpha} \right) P - \beta \end{aligned} \quad (7)$$

ここで、変換確率 P はこれまで説明したかな漢字変換の確率 $P(x|y)$ と予測変換の確率 $P(w_{n+1}^{n+N-1}|x)$ の積である。

$$P = P(x|y)P(w_{n+1}^{n+N-1}|x) \quad (8)$$

よって、以下のスコア値 s で予測候補を降順にソートをすれば、入力時間の期待値を削減できる順に並ぶことになる。

$$s = (L - L_0 + \beta/\alpha)P \quad (9)$$

上式の s は、変換確率 P に候補の予測の長さ $(L - L_0)$ を考慮した係数をかけたものと解釈することができる。定数 α と β の比は設計者が決めるパラメーターであ

る。 β/α が無限大の時は、予測候補の長さを考慮せず、変換確率 P の高いものほど選ばれやすい。すなわち、同じ読みで始まる候補は短い候補が優先される。

4. 評価実験

提案システムと Microsoft Office IME 2007(MS-IME)の入力効率を比較するため、評価実験を行った。また、提案システムが実時間で利用可能であることを確認するため、変換にかかる時間を調べた。また、実際に予測変換を行った場合の候補の確認を行った。

4.1 実験方法

日本語入力の入力効率を調べるため、実際に PC 初心者者に文章の入力を行ってもらい、その所要時間を測定した。また、上級者には擬似的に初心者の代わりとして、片手の人差し指で入力してもらった。評価基準は、入力時間とキータイプ数である。被験者数は 15 人である。

入力文章としては、2 種類の文章を想定した。1 つはメールを想定した文章であり、もう 1 つはブログを想定した文章である。これらの文章は、MS-IME で誤変換が生じないような一般的な文章とした。そのため、実験結果には提案システムの予測変換機能の効率性が反映されたと考えられる。パラメーター設定は、予備実験により $\beta/\alpha = 5$ とした。

4.2 実験結果

図 8 に入力時間の実験結果を示す。MS-IME と比べ、提案システムは文章 1 と文章 2 において入力時間とキータイプ数を削減することができた。入力時間は平均 21%削減され、キータイプ数は平均 26%削減された。入力時間の削減率がキータイプ数の削減率よりも小さいのは、予測候補を目で見て確認する時間が加わったためと考えられる。本実験では初心者ユーザーを想定して実験を行ったが、上級者は見るよりも打つのが早いいため、予測変換の効果が薄いことが報告されている[3]。

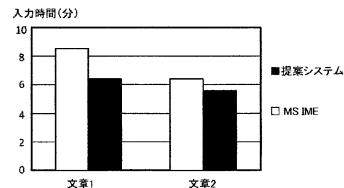


図 8 入力時間

Fig. 8 The input time

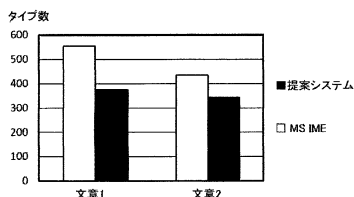


図 9 キータイプ数

Fig. 9 The number of key types

4.3 予測変換の例

図 10 に変換の例を示す。入力中はリアルタイムに予測候補を表示し、いつでも選択することができるインターフェースを用いた。また、長い文章でも、前半部分をかな漢字変換によって変換することができることが確認できる。

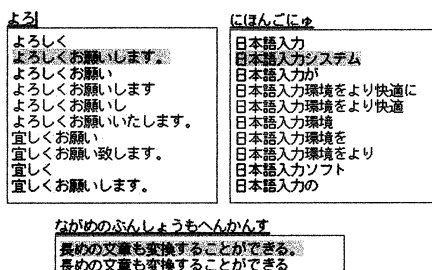


図 10 予測変換の例

Fig. 10 An example of the prediction

5. 結論

インターネットを用いた日本語入力システムを提案した。提案システムはインターネットのメリットを最大限に活用するため、インターネットを通してサーバー側で変換を行う。インターネットを用いることで、次のような利点がある。(1) Web 上の文章から抽出された大規模な統計量を用いることができる。(2) サーバースайдの豊富なハードウェアリソースを利用できる。(3) ユーザーが登録した単語を共有することで、専門用語や流行語などの単語を変換できるようになる。提案システムで用いる変換アルゴリズムには、次のような特長がある。(1) 予測変換とかな漢字変換が統合されている。(2) 入力時間のモデル化を行うことで、候補の長さを考慮した予測変換が可能である。提案システムの評価を行うため、初心者ユーザーを想定して文章を入力する評価実験を行った。実験の結果では、提案システムは Microsoft Office IME 2007 と比べて入力時間が平均 21%、キータイプ数が平均 26%削減された。このようなシステムを実装することで、インターネッ

トを用いた日本語入力システムの可能性が示された。今後の課題として、学習による個人適応への対応が挙げられる。

謝辞 本研究は 2007 年度に情報処理推進機構の未踏ソフトウェア創造事業の支援を受けて行われた。

参考文献

- 1) インターネット白書 2008, インプレス (2008)。
- 2) できる日本語入力—Windows 版, インプレス (2000)。
- 3) 市村由美, 斉藤佳美, 木村和広ほか: 入力予測機能を組み込んだ仮名漢字変換システム, 電子情報通信学会論文誌 D-II, Vol.J85-D-II, No.12, pp.1853-1863 (2002)。
- 4) ズバリ簡単入力, ソースネクスト (2005)。
- 5) 田中辰雄, 矢崎敬人, 村上礼子: ブロードバンド市場の経済分析, 慶應義塾大学出版会 (2008)。
- 6) 工藤拓, 賀沢秀人: Web 日本語 N グラム第 1 版, 言語資源協会発行 (2007)。
- 7) 森伸介, 土屋雅稔, 山地治, 長尾真: 確率的モデルによる仮名漢字変換, 情報処理学会論文誌, Vol.40, No.7, pp.2946-2953 (1999)。
- 8) Fomey, G.D., Jr.: The viterbi algorithm, Proc. IEEE, Vol. 61, No. 3, pp.268-278 (1973)。
- 9) 北研二, 辻井潤一: 確率的言語モデル, 東京大学出版会 (1999)。
- 10) 工藤拓: Mecab, 入手先
<<http://mecab.sourceforge.net/>> (参照 2009-02-03)。