2J-5

日本語入力補助手法としての曖昧変換を用いた自動登録システム*

蓮井洋志†

室蘭工業大学情報工学科‡

1 はじめに

仮名漢字変換の精度の向上にともなって、一般の人にもコンピュータで日本語文書を作成することが容易にできるようになった。小山ら [1] が提案したシステムは、第一番目の変換候補の正解率が 88.1%で、既存のシステムよりも 8.9%向上した。しかし、いくら精度の高い変換であったとしても、頻出する長い語を繰り返し入力するのは面倒である。科学技術論文や雑誌、新聞の社説の中にはキーワードとなる複合語や片仮名表記の外来語が頻繁に出現する。何度も入力している間に表記に揺れが生じる場合もある。

我々は、こういった問題を解決するために入力補助システムを提案してきた。本論文では、曖昧変換システムを提案する。このシステムは、前の文脈に現れる語を自動的に辞書に登録し、平仮名文字列をそれと類似した読みを持つ辞書内の語に変換する。ユーザの入力した表記が短縮形であっても、少々誤った表記であっても正しい表記に変換することができる。短縮形で入力すれば、繰り返し現われる長い語を打鍵数を少なく入力することができる。この変換および辞書は、既存の変換や辞書とは別に作る。

本論文においてまず、2 章では短縮形の曖昧変換について述べる。3 章では曖昧変換システムの構成および実現方法を述べる。

2 短縮形の曖昧変換

2.1 曖昧変換

曖昧変換とは、平仮名文字列を辞書内のそれと類似した読みを持つ語に変換することである。表記にずれがあっても正しい変換ができる。また、読みを短くした形で変換すれば仮名入力の手間を減らすこともできる。平仮名文字列と辞書内の表記の間の類似度を定義し、閾値以上の類似度を持つ語を変換結果とする。

表 1: 短縮形を基準とした類似度の統計

語	平均値	標準偏差	最小値	最大値
自然言語処理	0.64	0.65	0.43	0.88
視覚研究	0.48	0.49	0.33	0.57
仮名漢字変換フロントエンドプロセッサ	0.46	0.50	0.23	0.76
機械翻訳システム	0.56	0.61	0.27	0.90
情報工学ゼミナール	0.58	0.59	0.40	0.78
句構造表現	0.70	0.71	0.50	0.89
統語解析	0.56	0.59	0.36	0.92
入力支援	0.56	0.59	0.43	0.80
文書データベース検索支援	0.58	0.60	0.35	0.79
要約文生成法	0.60	0.63	0.53	0.80

2.2 表記の類似度

曖昧変換を行なう上で、類似した表記の度合を決定するために、表記の類似度 $Similarity(A,B)$ を以下の式で定める。システムが自動的に登録した表記をシステム表記と呼び、ユーザが入力する表記はユーザ表記と呼ぶ。ユーザ表記を A、システム表記を B とし、その A と B の間の類似度 $Similarity(A,B)$ を以下の式で定める。

$$Similarity(A,B) = \frac{SameLetterNumber \times 2}{AllLetterNumber} - \frac{DifferentLetterNumber(A)}{AllLetterNumber}$$

$SameLetterNumber$ は A と B とで共通に存在する文字数を表す。
$DifferentLetterNumber(A)$ は A にしか存在しない文字数のことである。
$AllLetterNumber$ は A と B のすべての文字数の和を表す。

2.3 閾値の設定

短縮形は人の主観によって異なる。大学生 24 人に対して 10 個の複合語のアンケートをとった。

短縮形と実際の読みの間の類似度を計算し、各々の語に対して平均値、標準偏差、最大値、最小値を求め

*Automatic Registration System with Approximate Translation for Japanese Typing-Aid Method

†Hiroshi Hasui

‡Department of Computer Science and Systems Engineering in Muroran Institute of Technology

表 2: 人を基準とした類似度の統計

人	平均値	標準偏差	人	平均値	標準偏差
1	0.73	0.74	13	0.54	0.57
2	0.49	0.59	14	0.59	0.60
3	0.63	0.65	15	0.51	0.53
4	0.70	0.71	16	0.56	0.57
5	0.68	0.69	17	0.51	0.51
6	0.78	0.79	18	0.59	0.61
7	0.73	0.74	19	0.50	0.54
8	0.52	0.54	20	0.57	0.57
9	0.59	0.60	21	0.57	0.64
10	0.55	0.58	22	0.49	0.49
11	0.52	0.53	23	0.51	0.52
12	0.55	0.55	24	0.50	0.51

た。それを表 1 に示す。最小値は、類似度が 0.0 のものを除外した。これらの類似度 0.0 の短縮形は 1 文字目がユーザ表記とシステム表記で異なったものであった。この結果からいえることは、語によって平均値が大分異なる。つまり短縮形の類似度は語によって大きな差がある。また、標準偏差は平均値と比較して大きく、語によって人の作る短縮形の類似度には大きな開きがあることが推測できる。つまり、閾値は語によって異なることが望ましい。

表 2 はアンケート結果を人の面から統計をとったものである。各人の類似度の平均と標準偏差を表している。平均値を見ると類似度は個人差があることがわかる。また、平均値と比較して標準偏差が大きく、個人の中でも語によって類似度は異なることがわかる。閾値は人によって異なることが望ましい。

本研究の曖昧変換システムは、ユーザごとに辞書を別とし、語ごとに閾値を設ける。閾値は最初の 2 回の変換で入力時の類似度の小さい方の値とする。こうすることで、語にあった変換、人にあった変換を行なう。

3 曖昧変換システム

3.1 曖昧変換システムの構成

曖昧変換システム **uum.5a** のシステム構成を図 1 に示す。このシステムは、**Linux** 環境上で動作するフリーウェアの仮名漢字変換システム **FreeWnn-1.1** のクライアントである **uum** を改良して実現した。3 つの部品と 2 つのデータベースを追加した。サーバは **uum** と同じものを活用する。

部品:
(a) 曖昧変換部
(b) 自動登録部
(c) 自動登録辞書

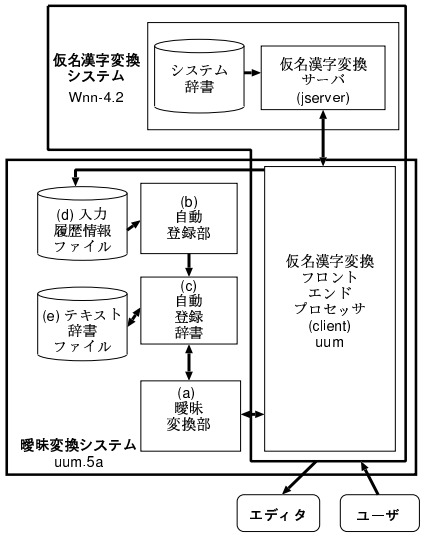

図 1: 曖昧変換システム **uum.5a** の構成

データベース:
(d) 入力履歴情報ファイル
(e) テキスト辞書ファイル

従来の変換はクライアント **uum** を通してサーバ **jserver** と接続し、変換結果を得る。

曖昧変換部 (a) は自動登録辞書 (c) の中の、語を曖昧検索する。その結果が変換候補である。この辞書 (c) はメモリ上にある。起動時に自動登録部 (b) が (c) にテキスト辞書ファイル (e) 内の語を登録する。また、ユーザが従来の変換で入力する度に、その単語列が入力履歴情報ファイル (d) に入り、その中の語を (b) が自動的に辞書 (c) に登録する。

(c) は単語の読みと変換結果と使用回数、閾値を 1 エントリとして線形リストで実現した。(c) の最大登録語数は 3000 語で LRU で管理する。終了時に (c) に登録されているすべての語をテキスト辞書ファイル (e) に保存する。入力履歴情報ファイル (d) とテキスト辞書ファイル (e) はディレクトリごとに用意する。一つのディレクトリで一つの文書を書く場合が多い。疑似的に文書ごとに (d)、(e) を用意する。文書によって頻繁に使う語が異なるためである。

参考文献

[1] 小山泰男, 安武満佐子, 吉村賢治, 首藤公昭. 連語データを利用した仮名漢字変換. 情報処理学会論文誌, Vol. 39, No. 11, pp. 2978–2987, 1998.