曖昧変換を利用した自動短縮登録システム

4M－4

蓮井洋志†　魚住 超†　小野功一†
†室蘭工業大学情報工学科

1　はじめに

日本語の入力は、普通仮名漢字変換システムを用いる。このシステムは、性質上複合語同音異義語誤りや表記が揺れた外来語が入力されてしまうという欠点を持つ。そのため、正しい入力を促進するために入力を支援するシステムが必要である。こういったシステムを入力補助システムと呼ぶ。

我々はこれまでに入力補助システムとして予測システム [1]、自動短縮登録システム [2] などを提案してきた。この中で、自動短縮登録システムは正しい入力を促進する効果があるだけでなく打鍵数を減らす効果もある。

自動短縮登録システムは以前に入力した単語列を短縮形で辞書に登録しておき、ユーザがそれを逆変換することで複合語を入力する。正確な表記で入力しておけば、正しい表記に変換できる。しかし、短縮形は覚えにくいし、とっさの入力時には思い浮かばない場合も多い。

本研究では、この逆変換に対して曖昧変換を応用した方法を提案する。曖昧変換とはユーザが入力した短縮形と類似した語を結果とする変換のことである。入力したい語の短縮形が少々誤っていても正確な逆変換ができる。

本稿の 2節では自動短縮登録システムについて説明し、3節では自動短縮登録システムに対する曖昧変換の応用方法について説明する。4節でそのシステムの構成を述べ、5節で曖昧変換を利用した入力例を分析する。最後に 6節でそれらを考察する。

2　自動短縮登録システム

自動短縮登録とは、過去に入力した語を辞書に短縮形で自動的に登録することである。ユーザは登録した短縮形を元の語に逆変換することで入力ができる。短縮形は実際の読みよりも短いために仮名入力に要するキーの打鍵数が少なくてすむ。この逆変換のことを短縮変換と呼ぶ。

このシステムでは、文書ごとに自動登録する辞書を変更する。文書ごとによく入力する単語は決まっている。以前に入力した語だけを自動登録することで、同じ辞書に 2つ以上の同音語が入る可能性を減らす。それゆえ、短縮変換の結果が 1つしかない語が多く、変換結果の選択の必要性が少なくなる。

3　自動短縮登録システムにおける曖昧変換

3.1　曖昧変換

曖昧変換は表記が類似した語を変換結果とする。入力中の語の読みの誤りや揺れを吸収する働きがある。また、長い語の変換では入力途中の読みでも変換できる。その場合には、残りの部分を予測する効果もある。その反面、普通の変換と異なって、多くの類似した変換結果ができてしまうという副作用がある。

3.2　表記の類似度

本研究では短縮変換に曖昧変換を応用する。ユーザの入力した短縮形と登録された短縮形の間の類似度がある閾値よりも大きいものを変換結果とする。表記の類似度 Sim は下式で計算する。

$$Sim = \begin{cases} \frac{SameLetterNumber \times 2}{AllLetterNumber} & (If\ First\ Letter\ is\ Same) \\ 0 & (Otherwise) \end{cases}$$

$SameLetterNumber$ は 2つの短縮形に共通した文字の数のことである。それに対して、$AllLetterNumber$ は 2つの表記の文字数の和である。また、類似度は最初の文字が同じでない場合は 0である。

閾値は 0.4とした。この値は経験的に定めている。

3.3　短縮形

自動短縮登録システムは、変換結果の数が少ないという利点があった。曖昧変換ではそれが増えてしまう。そのために、同音語の数を増やさないための工夫が必要である。

自動短縮登録システムでは名詞節も登録したが、曖昧変換システムは 1語からなる自立語と複合語しか登録をしない。

また、短縮形は語の読みを規則に従って短くしたものである。自動短縮登録システムでは短縮形の規則は 3つあるが、本研究ではその中の 1つの規則だけを用いる。この規則は頭の自立語の 2文字と 2番目以降の自立語の 1文字を並べたものを短縮形とするというも

Application of Approximate Translation to System for Automatic Registration of Contraction
Hiroshi Hasui, Takashi Uozumi, Koichi Ono at Department of Computer Science and Systems Engineering in Muroran Institute of Technology

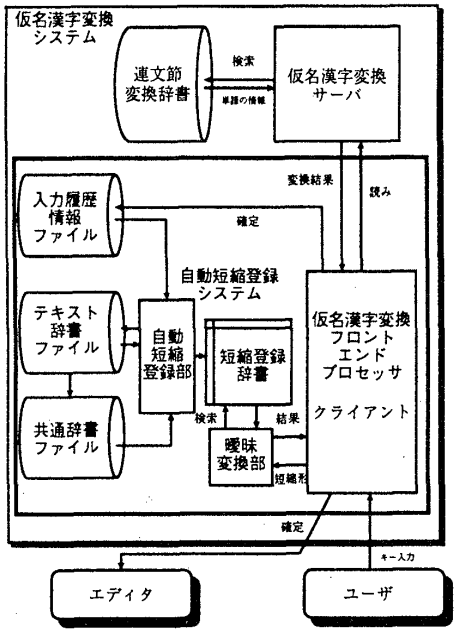

図 1:. 曖昧変換システムの構成

のである。例えば、「言語処理」という単語の短縮形は「げんし」となる。

3.4 候補の順位づけ

変換候補の順序で選択に使う手間の量が決まる。第1候補にユーザの入力したい語があれば、選択の手間がいらない。

候補の順位づけは下式で計算し、その値の大きいものから並べる。

$$Value = Sim * 50.0 + Useco$$

評価値は $Value$、類似度は Sim、候補語を短縮変換によって入力した回数は $Useco$ である。類似度が大きくて入力回数の多い候補の順番が小さい。

4 曖昧変換システムの構成

曖昧変換システムの構成は図1で表される。構成は自動短縮登録システムと同じである。既存の仮名漢字変換システムを土台として以下の2つの機能と3つのファイルを付け加えた。

機能は曖昧変換部と自動短縮登録部である。ユーザが入力を確定すると自動短縮登録部が自動的に短縮登録辞書に単語列を登録する。短縮変換を行なう時には短縮形が曖昧変換部に渡り変換される。

また、曖昧変換システムは入力履歴ファイル、テキスト辞書ファイル、共通辞書ファイルも持つ。これらのファイルは辞書データを保存する。

5 曖昧変換の動作例

曖昧変換の動作が良い方に働いた例と悪い方に働いた例を以下にあげる。矢印の左側が仮名入力した読みで右側が変換結果である。括弧の中はその語の短縮読みを表す。

良い例：

(1) 一般に使われている省略表現で変換できる

むろこうだい ⇒ 室蘭工業大学 (むろこだ)

かけんひ ⇒ 科学研究費 (かがけひ)

(2) 仮名入力途中で短縮変換を使いたくなった時に即座に変換に移れる

たんしゅくへ ⇒ 短縮変換 (たんへ)

(3) 普通の読みでも短縮変換できる

けいたいそかいせき ⇒ 形態素解析 (けいか)

悪い例：

(1) 短縮形が短い場合、変換結果の候補が多い

たん ⇒ 短縮、単語、短縮形など 15 個

6 考察

日本語入力で単語の読みを即座に短縮形に変換できる人は少ない。途中まで書いた後に短縮変換を使おうとする場合が大半である。曖昧変換はこういったユーザの心理を配慮した変換である。

日本語には省略表現が多い。日本人は省略表現をあらかじめ知っているために、短縮形を作る必要がない。省略表現の変換に本システムは向いている。

曖昧検索を利用した入力システムには動的曖昧検索システム ASearch[3] がある。このシステムは辞書内の類似した英単語を検索し入力する。本研究の曖昧変換はこれを日本語の仮名漢字変換に応用した。

閾値や変換結果の順位づけのパラメータの最適値を獲得することが今後の課題である。

参考文献

[1] 蓮井洋志, 西野順二, 小高知宏, 小倉久和：日本語入力における平仮名文字列の予測, 情報処理学会第 53 回全国大会講演論文集 (2), 125 - 126 (1996)

[2] 蓮井 洋志, 西野順二, 小高知宏, 小倉久和：入力補助システムとしての自動短縮登録システムの検討, 電子情報通信学会論文誌 (D-I), Vol. J81, No. 1 , 38 - 45 (1998)

[3] 増井俊之：動的曖昧検索, UNIX マガジン, Vol.13, No.1, 65 - 69 (1998)