5L−4

仮名漢字変換システムにおける単語自動登録の一方式

丸山芳男、下村秀樹、酒井貴子、並木美太郎、高橋延匡
(東京農工大学　工学部　電子情報工学科)

1. はじめに

現在、計算機で日本語を扱うために仮名漢字変換システムが主流となっており、仮名漢字変換の正変換率が計算機の使いやすさを大きく左右するといえる。

正変換率を向上させるためにさまざまな研究が行われているが、基本的な変換手法[1]がある程度確立されていることから、筆者は変換用の辞書を整備することが重要な課題であると考えた。システム側であらかじめ用意しておく辞書(システム辞書)の内容はもちろん、ユーザの手によって単語を追加登録していく辞書(ユーザ辞書)も理想的な内容になっていることが望ましい。そのために、多くのシステムでは単語登録の機能を提供することで辞書の整備をユーザに任せている。ユーザ辞書に単語を登録することはシステム辞書の不備を補うため、あるいはユーザに固有な単語に対応するために有効である。しかし現状の単語登録は、登録する表記の範囲指定、それに対応する読みの入力、さらに複雑な品詞情報の特定を要求しており、ユーザへの負担は大きい。本稿ではそのような負担をかけないために、単語登録を自動化するための一方式について述べる。

2. 単語自動登録機能の位置付け

本単語自動登録機能は、本研究室で開発された OS/omicron 仮名漢字変換システム第2版[2]上に実現されるものである。仮名漢字変換システム全体の中での単語自動登録機能の位置付けを図1に示す。

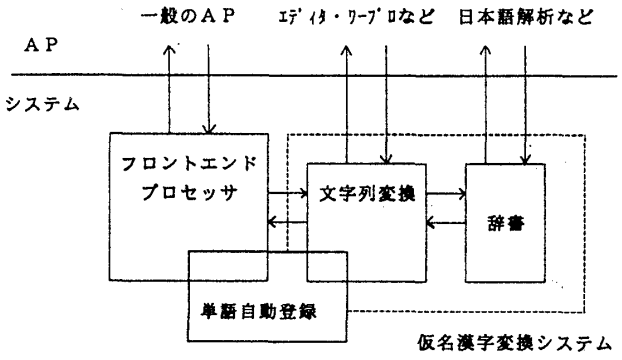

図1 単語自動登録機能の位置付け

3. 単語自動登録機能のモデル

単語自動登録を実現するために解決しなければならない問題は、次の二点である。

(1) 登録対象となる表記の特定
(2) 品詞の判定

特に実現が困難であり、かつ単語自動登録機能の有効性を左右するのが品詞の判定方法である。従来の登録方式ではユーザが品詞を指定するので、登録の際に品詞情報も一緒に辞書に格納される。しかし品詞を自動的に判定するとなると、登録すべき表記からだけでは即座に品詞を決定することは困難である。そこで本方式では、辞書への単語登録と品詞の判定を同時には行わず、名詞として仮登録し、後から品詞を学習していくモデルを採用する。

4. 登録対象となる表記の特定

登録対象となる表記は、未知語と複合語の二つに分類される。複合語は、その構成単語それぞれが辞書に登録されてあれば複数の単語として希望通りの変換結果が得られることもあるが、最近使用語学習[3]や変換アルゴリズム上の理由で次回入力時には同様な変換結果が得られない可能性がある。このような不安定な変換を改善するために、複合語は複数の文節で示されたまま確定されても、一つの単語として自動登録する。また、次候補を選択し続けた後、候補が尽きたら常に未知語として表記の作成をユーザに促す。ここでいう未知語とは、文節を区切り直しても(複数の単語に分割しても)希望する表記が得られないものである。未知語の表記を作成した場合は常に自動登録の対象にする。未知語作成中は最近使用語の学習は中断され、ユーザが希望した文節長に対応する入力文字列を変換中文字列とは別環境の枠内で変換・編集する。また、片仮名キーや無変換キーなど、変換作業に入る前に確定した表記についても登録対象とする。

次に、ユーザが一つの単語のつもりで入力した文字列が複合語として変換されたときに、区切りはともかく希望どおりの表記が得られた場合、次の二通りの対応が考えられる。

(1) そのまま確定する
(2) 希望どおりの読みに文節長を合わせる

A Method for The Automatic Dictionary Entry in Kana-Kanji Translation System
Yoshio MARUYAMA, Hideki SHIMOMURA, Takako SAKAI, Mitarou NAMIKI, and Nobumasa TAKAHASHI
Department of Computer Science, Faculty of Technology, Tokyo University of Agriculture and Technology

(2)の場合の「対応する読み」は、変換キー打鍵などによる候補取り出し操作時にフロントエンドプロセッサによって特定され、前者の場合は仮名漢字変換処理の内部情報を見ることによって特定される。

5. 品詞の判定方法

活用語尾・付属語まで含めた表記で仮登録された用言の品詞を判定するための材料は、その表記（字面）と読みだけである。市販ソフトの中には、同じ読みの単語が既に存在する場合にはその単語の同音異義語として品詞情報を継承していると推測できるものもあるが、同音異義語に限定しないように、また、品詞の異なる同音語も登録できるように本方式では仮登録された表記をもとに品詞が学習される。品詞の学習は、仮登録語を本来一つの単語であると思われるもの同士にグルーピングし、あらかじめ用意した活用パタンセット群とマッチングさせることによって実現される。また、正式に登録し直す際には語幹部分を切り出して登録する。

5.1 単語のグルーピング

仮登録単語群を同一単語と思われるグループにまとめる処理の様子を図２に示す。グルーピングの処理は、簡単な正規表現（ワイルドカード）にマッチするものをピックアップすることによって実現される。このとき、複数のグループに属する単語は、属する全てのグループに存在するものとして重複して処理する。また、構成メンバが一定数に満たないグループの単語については、品詞判定が次回学習時まで持ち越しとなる。構成メンバが揃わないうちはいつまでも仮登録のままであるが、仮登録語のままでも変換に利用できる場合がある。例えば、「泳（およ）：ガ行五段」という単語が辞書になかった場合、「泳いだ（およいだ）：普通名詞」という語が仮登録されていれば、「僕は泳いだ（ぼくはおよいだ）」という文を形態素解析した結果は「僕は人間」と同じであり、変換できるということになる。

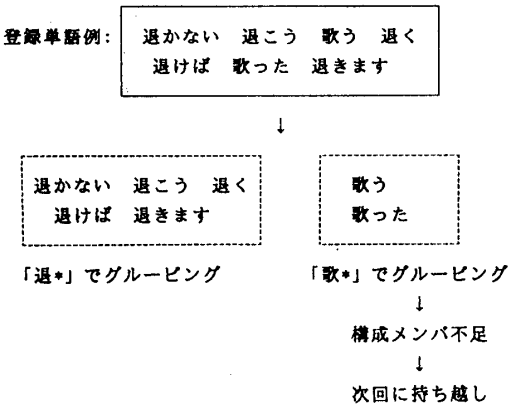

図２ 仮登録単語群のグルーピング

5.2 活用パタンセット群とのマッチング

活用語尾を含む文字列の正規表現を考え、その集合を活用パタンセットとする。一つのセットには音便も含めて５～１０通りのパタンを登録しておき、グルーピングした単語群の要素と活用パタンセット内の要素数個とがマッチすればその活用パタンセットと同じ品詞であると判断する。全ての活用パタンセットとマッチングを行っても品詞が判定できなかったものについては次回学習時まで持ち越しとなる。また、マッチングを行う際に語幹の切り出しも同時に行い、マッチしたグループのメンバを辞書から削除した後、語幹部分を改めて登録する。

マッチングの様子を図３に示す。

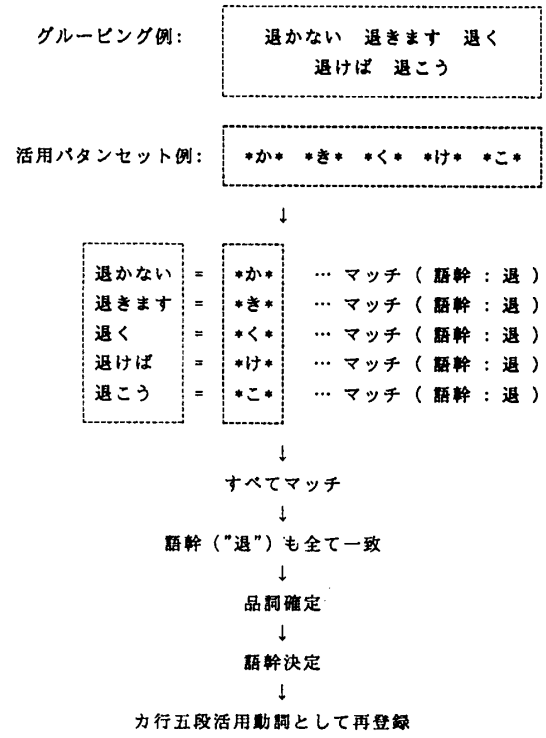

図３ 活用パタンセットとのマッチング

6. おわりに

本稿では、仮名漢字変換システムにおける単語自動登録機能の一方式について述べた。現在は基盤となるフロントエンドプロセッサ、および未知語作成ウインドウの部分が実現済みであり、今後は複合語の認識、品詞学習の機能を実現・評価する予定である。

参考文献

[1] 吉村賢治他：最長一致法と文節数最小法について、情報処理学会人工知能と対会報告 24-1, 1982

[2] 本宮志江他：OS/omicron 仮名漢字変換システム第２版の設計と実現、情報処理学会第 42 回全国大会論文集 5Q-2, pp.287-288, 1990

[3] 下村秀樹他：仮名漢字変換における最近使用語優先学習方式の解析と評価、情報処理学会第 46 回全国大会論文集 5L-6, 1993（予定）