仮名漢字変換システムにおける単語自動登録機能の実現

3N－7

丸山芳男，早川栄一，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

現在広く普及している日本語ワードプロセッサ（以下、「ワープロ」と記す）の入力方式として、仮名漢字変換方式が定着してきている。仮名漢字変換システムが用いる辞書には、あらかじめ十数万語もの単語が登録されるようになったが、基本的な語が充実するにつれ、予測ができないローカルな語彙が変換できない場合が、相対的に顕著になってきた。

ローカルな語彙は固有名詞だけではなく、送り仮名の規定、教育水準などに合わせた使用漢字の限定などの表記ルールが存在するので、膨大な数になる。これらに対応するためには、ユーザが自分で登録するしかない。

しかし、現在普及しているワープロが提供する単語登録インタフェースでは、登録単語の語幹部分と品詞情報を指定しなければならず、初心者にとっては困難な作業である。

筆者は、ユーザによる登録時の負担を軽減するために単語自動登録の研究を行なっている。本稿では、初版の設計[1]に基づいて実現した結果、明らかになった問題点に対処するため、新たに見直した設計内容について述べる。

２．単語自動登録機構の設計

本章では、ユーザによる単語登録時の負担を軽減するために筆者が実現した、単語自動登録機構初版の設計のうち、特に性能を左右する品詞判定の部分について述べる。

2.1 全体設計

本方式では、ユーザが単語を登録する際に、語幹および品詞情報の入力を省略できるように、活用語尾や付属語まで含めた表記を仮登録語として登録する。その際、品詞はすべて名詞としておき、仮登録語が蓄積した時点で改めて品詞の判定を行なう。

仮登録された単語の品詞判定は、グルーピング処理、マッチング処理から成る品詞学習機構によって行なわれる。まずグルーピング処理によって、蓄積された仮登録語の先頭数文字が等しいものを一つのグループとしてまとめる。グルーピングに複数の解釈が成り立つ場合、重複して処理される。

マッチング処理では、体言用の助詞１種類、用言用の活用語尾 33 種類から成るパタンセットとの文字列マッチングによって品詞を判定する。すべての要素がマッチしたパタンの持つ品詞情報をそのグループの品詞と判定し、グルーピングの際に抽出した先頭数文字を語幹とみなして正式に登録する。

３．実現

仮名漢字変換システムを拡張する形で、単語自動登録機構の初版を約 5000 行のＣ言語で実現した。実験などを考慮して、仮名漢字変換本体から切り離して独立したツールとして利用することも可能にした。よって、長期間の使用を再現した大量の仮登録語をバッチ処理で品詞学習させることができるほか、マッチング条件を変えて実験する際には、一度

Impremention of an Automatic Word Registration Technique for Kana-to-Kanji Translation Systems
Yoshio MARUYAMA, Eiichi HAYAKAWA, Mitarou NAMIKI and Nobumasa TAKAHASHI
Department of Computer Science, Faculty of Technology, Tokyo University of Agriculture and Technology

グルーピングすればその結果をファイルに出力し、マッチングだけやり直すことができる。

4 結果

評価のために、対象とする工学系の文書からベンチマークテキストを作成した。11,632文字からなるテキスト中、本方式で対応する品詞の単語 737 個から 2288 個の仮登録語を登録し、品詞学習機構にかけてみた結果、仮登録語の活用形について表1に示すような分布が得られた。

表1 テキスト中に現れる活用形の集計

品詞	種類	単語数
名詞	2 3 4 5以上	93 39 24 9個以下
カ行五段	2	7
ガ行五段	2	1
サ行五段	2 3	28 5
タ行五段	2	3
ナ行五段	2 3	58 1
ラ行五段	2 3	5 1
ワ行五段	2 4	3 1
タ行下一段	2	1
ナ行上一段	2	3
形容詞	2	4
形容動詞	2 3	25 2

5．考察

今回の結果は、全ての要素にマッチしたときに品詞を判定するとしたものなので、表1に示したベンチマークテキストではマッチするものはなかった。

より大量な文章を入力すれば、品詞判定に必要な活用形が揃う可能性もあるが、今回用いたベンチマーク程度の文書で効果を発揮するためには、品詞を判定する際のスレッシュホールドを3程度にして、3種類以上の活用形が揃った時点で正式に登録し直してやる必要がある。

また、用言のパタンセットである活用語尾は、各品詞ごとに 5～6 通りであり、あまり低く設定すると誤判定が起こるので、スレッシュホールドを二段階にし、最初の段階では単に「品詞情報を付加した仮登録語」として辞書に格納する。最も多くマッチしたパタンの品詞情報を付加した語幹部分を仮登録語として新たに登録することで、品詞判定までの期間の正変換率を改善する。

6．おわりに

仮名漢字変換システムに単語登録する際に必要な情報、特に品詞を自動判定することで、ユーザによる単語登録時の負担を軽減する方式について述べた。今後の課題として、誤判定が起こらず実用的な時間で品詞を判定するためのスレッシュホールドを決定することが重要である。

5．参考文献

1）丸山芳男他：仮名漢字変換システムにおける単語自動登録の一方式，情報処理学会第46回全国大会5L-4，1993