ヒューマンインタフェース 57－6
（1994. 11. 10）

仮名漢字変換システムにおける自動単語登録手法の設計と実現

丸山芳男[†]，酒井貴子[††]，下村秀樹[†††]，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
[††]日立製作所システム開発研究所
[†††]NEC 情報メディア研究所

本稿では，仮名漢字変換システムにおいて，ユーザが単語を登録する際の品詞を自動的に判定し，ユーザの手間を軽減する手法の設計と実現について述べる．従来のシステムでは，表記だけでなく品詞まで指定しなければ単語を登録することができず，ユーザに品詞に関する知識と手間を要求していた．今回提案する自動単語登録手法では，単語登録の際には仮の品詞で登録しておき，パタンセットとのマッチングによって後から品詞を判定する．本手法では，ユーザは品詞を選択する手間をかけずに，また，品詞に関する知識がなくとも単語登録が可能になるという特徴がある．今回，本手法を実現するために実験によって定めなければならないパラメータを明らかにし，実験環境の整備を行なった．

Design and Implementation of an Automatic Word Registration Technique for Kana-to-Kanji Translation Systems

Yoshio MARUYAMA[†], Takako SAKAI[††], Hideki SHIMOMURA[†††],
Eiichi HAYAKAWA[†], Mitarou NAMIKI[†], and Nobumasa TAKAHASHI[†]

[†] Dep. of Computer Science, Faculty of Technology,
Tokyo University of Agriculture and Technology
[††] Systems Development Laboratory, Hitachi Ltd.
[†††] Information Technology Research Laboratories, NEC Corporation

This paper describes the design and implementation of an automatic word registration technique for a Kana-to-Kanji Translation System. This technique decreases the user's workload by carrying out registration in place of the user. In previous systems the user must understand parts of speech as well as registering the words themself because these techniques of registration don't demand just the word but also its part of speech. The technique proposed by this paper determines the part of speech by pattern-matching after first the registering the words with a default part of speech. This technique has the merit that it makes it possible to register a word without work or knowledge of parts of speech. Through this research it was made clear how parameters must be determined for implementation of this technique, and an environment for experiments was produced.

1．はじめに

近年、日本語ワードプロセッサ（以下、ワープロと記す）は高性能化が進み、広く普及している。基本機能である文字入力には、キーボードを用いた仮名漢字変換システムが主流であり、その性能は変換アルゴリズムと変換用辞書の語彙数の両側面から向上が計られてきた。

このうち、辞書の語彙数は重要な問題である。基本的には、辞書に格納されていない単語は変換することはできない。次候補群の中にも希望の候補が含まれない、といった変換不能の事態を回避するためには大量の単語が登録された変換用辞書が必要とされる。

仮名漢字変換に用いられる変換用辞書には、十数万語の単語が登録されており、一般的によく用いられる語句が変換できないということは少なくなっている。しかし、特定のユーザ集団ごとに存在する固有名詞や特殊な表記ルールの語句など、ローカルな語彙に対しては十分であるとはいえない。すべてのユーザ集団に存在するローカルな語彙を辞書に格納しておくのは、辞書の容量の点から非現実的な解決策であり、また当然、どのような語彙がローカルに存在するかをあらかじめ知ることも困難である。すなわち、ユーザが使っていくうちに登録してやらなければ、ローカルな語彙を充実させることは不可能であるといえる。

そこで、ほとんどのワープロではユーザに単語登録の機能を公開し、インタフェースを提供している。しかし、変換用辞書への単語登録は本来、ワープロの開発あるいは保守を担当する人間が行なうべき処理であり、エンドユーザにとっては難解な作業となっている。

筆者は、ユーザによる単語登録の際のインタフェースを改善する必要性を感じ、単語登録を自動化する手法を設計した。本稿では、読み及び変換後表記の取得法、品詞の判定法について述べる。

2．単語登録時インタフェースの問題点

ユーザが単語を登録する場面は、次の二通りに分けられる。

a）辞書の整備を目的に、明らかに辞書に登録されていないと思われる単語をあらかじめ登録する。

b）文章入力中、変換できない単語が現れたので、次に現れるときに備えて登録しておく。

利用者の側からすると、b）の場面で登録を行なうのが自然であり、a）の場面と比べても多いはずである。

現在多くのワープロが採用している、変換作業中の単語登録のインタフェースでは、ユーザは次のような情報を入力しなければならない。

①読みの表記の語幹部分
②変換後の表記の語幹部分
③品詞（活用の種類）

固有名詞しか登録しないのであれば、語幹や品詞を指定する必要はないが、1．で述べたように、本研究の目的はローカルな表記を対象とした単語登録の自動化である。ユーザ集団ごとに存在するローカルな表記ルールには、送り仮名の規定や教育水準に合わせた使用漢字の限定などが考えられるが、これらは特定の語彙に対してのみあてはまるわけではなく、そのユーザ集団が作成する文書中に現れるすべての語彙に一貫して適用されるルールである。すなわち、わずかでもローカルなルールがあれば、ユーザが品詞や語幹を間違えやすい、活用する語（用言）も多数、登録の対象になるはずである。

用言の単語登録が困難であるため、本来存在していたローカルな表記ルールを適用せずに、ワープロが採用している表記ルールに従ってしまうユーザも多い。これはワープロの功罪といえるだろう。

3．設計方針

2．で挙げた、登録時に必要な①～③の三つの情報のうち、ユーザの負担が最も大きく、最も改

善が望まれるのは品詞情報の入力である。また、語幹を指定するという作業も、品詞や活用の種類を指定するのと同様に、多くのユーザにとっては困難であり、これらを同時に解決できなければ、単語登録のインタフェースを改善したことにはならない、と考える。

本方式では、次の三点を設計方針とする。

1）品詞の知識を要求しない。
2）登録する表記の範囲は、語幹の指定を要求しない。
3）登録する用表記の作成によって、学習に悪影響を与えない。

また、本方式は、2．で示した二つの登録方法のうち、特にユーザが利用する場面が多いであろうと思われるb）の方法を対象として、ユーザインタフェースの向上を計るものである。

4．初版の設計

4.1 本方式の特徴

3．で述べたように、本方式はユーザに品詞の指定を要求しないと同時に、用言の語幹指定も要求しないことを方針としている。よって、（語幹のみとは限らない）不確実な範囲の表記を蓄積していき、蓄積された複数の表記をもとに品詞の判定を行なう。品詞を判定するための直接の材料となるのは、活用語尾または付属語と思われる部分である。システムが要求する語幹指定に従わず、ユーザが自然だと思う区切りで指定した表記範囲の、本来余分であるはずの部分を利用して品詞の判定及び語幹の切出しを行なう点に、本方式の特徴がある。

4.1 登録までの流れ

本方式では、登録単語の表記を得た時点では、仮登録語として専用辞書に格納される。用言の場合には活用形＋αの表記で格納され、蓄積した仮登録語は、同じ単語の活用形と思われるもの同士でまとめられて1グループになる。品詞はグループごとに判定し、語幹部分の表記で正規の辞書に本登録する。

読み及び変換後の表記は、仮登録機構と呼ぶモジュールによって取得される。この時点では品詞の判定は行なわず、変換に利用できるよう、仮の品詞情報を付加して専用辞書に仮登録する。

専用辞書に仮登録語が蓄積されると、品詞学習機構と呼ぶモジュールが起動する。品詞学習機構は、品詞の判断材料となる仮登録語を同一単語ごとにまとめるグルーピング処理部と、各グループの品詞を実際に判断するマッチング処理部からなり、グルーピング処理が成功すればマッチング処理に、しなければ品詞学習は持ち越しとなる。マッチング処理まで成立すれば品詞が判明するが、仮登録語の不足でマッチングが成立しなければ品詞学習は持ち越しとなる。

仮登録から品詞が判断されるまでの流れを、図1に示す。

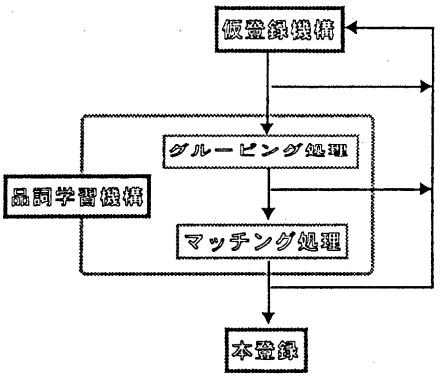

図1 仮登録から品詞学習までの流れ

4.2 学習の対象とする品詞

品詞学習機構で学習する品詞は、いくつかの種類に限定している。副詞、接続詞のような品詞は、仮名漢字変換の処理において重要な意味を持つことが多く、あらかじめシステムの辞書に格納しておくべきだからである。また、例えばカ変動詞のような例外的なものも、未知語として現れるとは

考えにくい。

そこで、本方式で用意する活用パタン、すなわち学習の対象とする品詞は、次の６種類に設定した。活用語尾、助詞は国文法[1]に従ったが、単純な文字列マッチングに使用するものなので、仮定形と命令形など、表記が重複するものはひとつにまとめてある。実際には、動詞はさらにア行からワ行までのそれぞれに対してパタンが用意される。

五段動詞 …… ９パタン
下一段動詞 …… １２パタン
上一段動詞 …… １０パタン
形容詞 …… １パタン
形容動詞 …… １パタン
普通名詞 …… １パタン

活用パタンは、活用語尾を検索する正規表現をもとにした文字列の集合である。一例として、カ行下一段活用のパタンを示す。

{ *け*, *ける*, *けれ*, *けろ*, *けよ* }

4.3 品詞学習方式

品詞の判断材料である仮登録語がある程度蓄積されたら、まず、グルーピング処理によって本来同じ単語の異なる活用形であると思われるもの同士をまとめる。理想値として充分な数の仮登録語があれば、グループ内にはすべての活用形が含まれていることになる。これを、マッチング処理で活用パタンセットと文字列比較し、該当する品詞を判定する。

4.3.1 グルーピング処理

仮登録語の切出し範囲は不定であり、活用語尾や付属語を含んでいる可能性が高い不確実な情報である。しかし、登録対象の前に余計な文字列が付属しているとは考えない。すなわち、仮登録語は次のいずれかの形をしていることになる。

体言 ［＋ 付属語］

用言［＋ 活用語尾］［＋ 付属語］

よって、仮登録語の先頭から数文字が等しいもの同士を同じ単語とみなしてグルーピングする。一つの仮登録語が複数のグループに重複して属することも有り得る。

このとき、グループ内の仮登録語に共通な先頭数文字が語幹または体言部分として認識される。

また、この時点でグループ内のメンバが不足しているときには、品詞学習には時期早尚と判断して処理を中断する。

仮登録された単語と、それをグルーピングしたものの例を図２に示す。

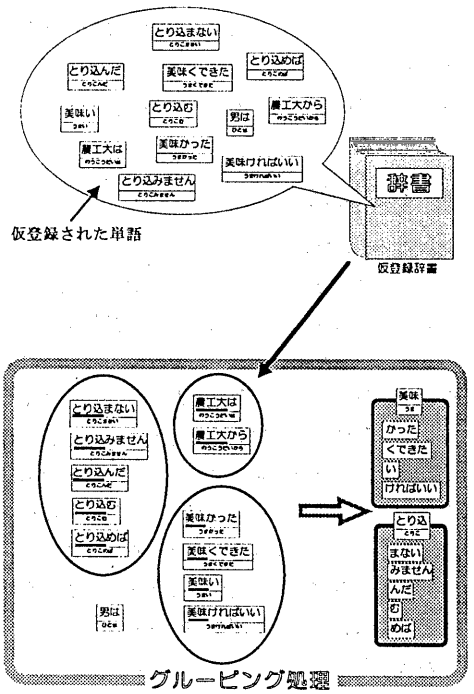

図２ 仮登録語のグルーピング例

4.3.2 マッチング処理

活用する語（用言）の品詞を判定するために、グループ内の各単語表記とあらかじめ用意した活用パタンセットとを比較する処理である。グループ内の仮登録語は 4.3.1 で示したように、語幹と

思われる先頭数文字に付属語や活用語尾などの文字列が続いたものであるが、この語幹の後の数文字を比較対象にする。比較の結果、活用パタンのすべての要素にマッチすれば、マッチした活用パタンで示される品詞であると判定して、語幹部分を辞書に本登録する。

また、活用しない語（体言）は一括して普通名詞として扱い、活用語尾の代わりとして、格助詞、副助詞、終助詞をもつパタンセットをマッチングに用いる。

活用パタンセットとのマッチングの様子を図３に示す。

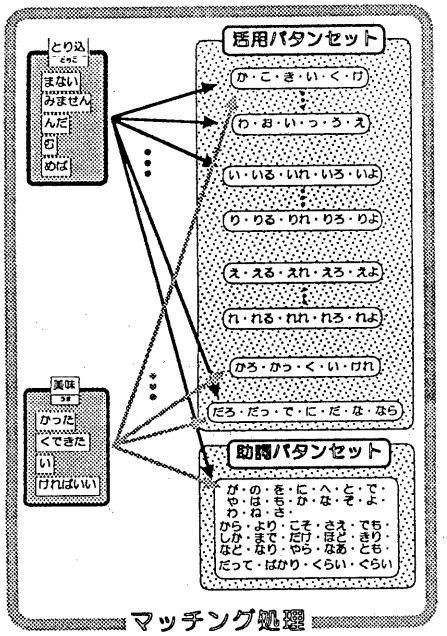

図３ 活用パタンセットとのマッチング

4.3.3 仮登録語の品詞予想

マッチングが成功するまでは、仮登録語には仮の品詞として、普通名詞、サ変動詞の属性を付加して変換に利用する。しかし、この段階では品詞だけでなく、表記も不確実なままである。品詞を判定し、語幹を切り出すまでには到らないが、仮登録段階でも、少しでも正確な情報を随時追加していくべきである。そこで、マッチング処理に移行できる程度にメンバが揃ったグループについて、次の予想を立てる。

①ある活用パタンのほとんどの要素にマッチした
↓
おそらく用言として学習される。

②助詞パタンのほとんどの要素にマッチした
↓
おそらく体言として学習される。

この予想がたった時点で、そのグループの語幹部分を新たに仮登録語として追加する。その際、①のグループなら「仮用言」という拡張品詞の属性をつける。仮名漢字変換エンジンでは、この品詞をもつ仮登録語の後ろになんらかの活用語尾が続いた場合、無条件に接続可能であるとして処理する。

②の場合には、最初の仮登録語同様、普通名詞およびサ変動詞の属性とする。

4.4 仮登録機構

読み及び変換後の表記を得て、専用辞書に仮登録するモジュールである仮登録機構は、直接ユーザとの入出力を処理するフロントエンド部分に組み込まれ、入力時の情報を利用して表記を得る。

4.4.1 読み表記の取得

登録単語の読みの表記はすべて入力中の文字列に含まれているので、その中から該当する部分を抜き出して利用する。

通常、変換候補選択時には先頭から文節を確定していくが、文節単位で希望の候補が現れないときには次のような操作が行なわれる。

変換する
↓
文節長を指定する
↓
次候補を順次呼び出す
↓
候補が尽きる

すなわち、分かち書き[2]を要求しないことを前提とした連文節変換でも、希望の候補が現れないときにはユーザは変換したい単語の範囲を指定していることになる。この伸張操作時の情報を利用すれば、登録すべき単語が含まれている文節の読みに該当する仮名文字列を切り出すことができる。

ただし、活用する語（用言）の場合、辞書のレコードサイズを抑えるため、あるいは一つの単語に複数の品詞情報を格納するために、通常、ユーザ辞書には語幹部分だけを格納し、活用語尾はあらかじめ変換エンジンの内部データとして持っていたり別の専用辞書から読み込んだりするようになっている。

文節単位で切り出された文字列には、活用語尾や付属語といった、本来登録対象に含めない文字列が含まれている可能性が高い。しかし本方式では、品詞学習機構のグルーピング処理により、語幹部分を切り出すことができる。

4.4.2 変換後表記の取得

希望の候補が得られないとき、ユーザは未確定の文をキャンセルまたはいったん確定した後、別の（変換できそうな）読みで変換して表記を得、編集するしかない。本方式でもこの基本的な手順を自動化するには到っていないが、読みの表記が入力中に得られるようになったことで、ユーザが指定しなくとも自動的に登録モードに移るようになり、次の点が改善された。

①確定後の編集が省略できる

未確定の文を一旦キャンセル、あるいは確定してから編集するのでは、入力した文字列が無駄になる。本方式では入力中の文字列を元に自動的に登録モードに移るので、未確定の文をキャンセルすることもなく、また表記を得るための作業が専用のウィンドウ内で行なわれた後、作成された表記は該当する未確定文字列に差し替える形で挿入されるので、編集作業を最低限に抑えることができる。

②学習辞書への登録を自動的に中断できる

変換できない単語の漢字表記を得るために別な読みで変換する場合は、入力中の文書に関係のない単語であることも多い。確定した候補を専用辞書に登録していく「学習機能」の目的は、候補選択の基準に「最近（現在の文書入力中に）使われた単語が再び使われている可能性が高いので、優先順位を上げる」ことである[3]。したがって、この場合には学習をするべきではない。熟練したユーザなどは、細かく学習機能を手動で中断させたりもするが、本方式では自動的に登録モードに移り、表記の作成は専用ウィンドウを介して行なわれるので、モードの切替え時に自動的に学習機能を制御することができる。

5．実現における問題点と解決策

本方式を仮名漢字変換システムに実装するにあたり、設計どうりの性能を発揮して実用化するためには、次に述べるような点が問題点となる。

5.1 学習機構の起動タイミング

グルーピングおよびマッチングからなる品詞学習機構は、ある程度の仮登録語が蓄積された時点でないと新たな語が学習されることは少ないと予想される。さらに、仮登録が起こるたびに品詞学習を試みると、速度的に実用性を欠くことになる。しかし、明らかに、普通名詞とサ変動詞の品詞しか持たない仮登録語のままでは正変換率に最良の効果を発揮するとは思えないので、品詞判定可能な情報が揃ったなら、即座に判定し、本登録語にするのが理想である。よって、実験によって品詞学習機構を起動する適度なタイミングを検討する必要がある。

実験により、サンプルテキスト中の未知語の分布、活用形が揃うまでに変換エンジンに渡される文字数、グルーピング及びマッチングの処理速度等を測定し、タイミングを決定する。

5.2 マッチング成否境界の設定

設計では、マッチング処理によって品詞が判定

されるのはグループ内の仮登録語が、ある活用パタンのすべての要素にマッチした場合、すなわちすべての活用形が揃った場合である、としていた。しかし、実使用において、どの程度の文章を変換すればすべての活用形が揃うのかを予測しておかないと、ほぼ永遠に仮登録のままになっていては本方式の意味がない。すべて揃うのに非実用的な入力数が必要であると仮定すると、妥協して何割かのマッチングに成功した時点で品詞を学習してやらなければならない。しかし、マッチング成功率が低いうちは複数の活用パタンにマッチしてしまう可能性があるし、グルーピング処理での語幹部分切出しが誤っている可能性も高くなる。そこで、マッチする要素の数に対する、誤って品詞を判定してしまう確率を実験によって測定し、品詞判定のスレッシュホールドを定めてやる必要がある。

また、普通名詞を判定するために活用パタンセットの代わりに導入した助詞パタンセットは、活用パタンセットに比べて要素数が多いので、独自のスレッシュホールドを定めるべきであろう。

5.3 品詞予想の発生タイミング

品詞の学習を予想して新たな仮登録語を生成するのは、4.3.3 で述べたように、グルーピングはできたがマッチングに失敗した場合である。しかし、マッチングに失敗する、すなわち充分な種類の活用形（用言の場合）が揃っていない段階では、グルーピングに失敗している可能性がある。誤ったグルーピングの結果から新規の仮登録語を生成しても、変換性能が落ちるだけである。また、仮用言はあらゆる品詞の活用語尾と接続するが、様々な仮用言が登録されていると誤変換の原因になるかもしれない。

品詞予想を立てるときのマッチング成功率は、5.2 で述べた品詞学習のスレッシュホールドよりさらに低く設定しなければならない。上記の問題を考慮し、品詞学習のスレッシュホールド同様、実験によって用言と体言（普通名詞）それぞれに値を定める必要がある。

5.4 実験環境

5.1 から 5.3 で示した問題点を解決するための実験環境を、次に述べる。

5.4.1 仮名漢字変換システム

自動単語登録機能は、筆者の所属する研究室で開発された OS/ｏ仮名漢字変換システムＶ２[4][5]に組み込む形で実現される。この変換システムでは、すべての単語を木構造に展開し、候補の優先度を、木をたどる際のコストの合計で表わす。変換木の各ノードにはコストを決める評価関数によって評価値が付けられ、文節の長さ、前後の単語との接続、検索元辞書の種類、最終使用日時（学習語）など、評価関数でのパラメータの扱い方を替えるだけで様々な変換手法に対応できるようになっている。

仮登録語は、表記の範囲、品詞ともに不確実な単語であるため、正式な登録語と比べて優先順位を下げるべきであろうが、特定の品詞、あるいは特定の辞書から検索された単語の評価値を下げることも容易に実現できるようになっている。

5.3.2 ツールセット

実験を行なうためには、変換エンジンに渡すための仮名テキストや、ツール群が必要である。ツール群には、過去に変換精度測定のために作成されたツールを再利用するほか、新たにいくつかのツールを作成する。最終的に自動単語登録機能としてまとめる仮登録・品詞学習のルーチンも、個別にリンクする形で、まずはツールとして実現する。

実験の流れと使用するツールを図４に示す。

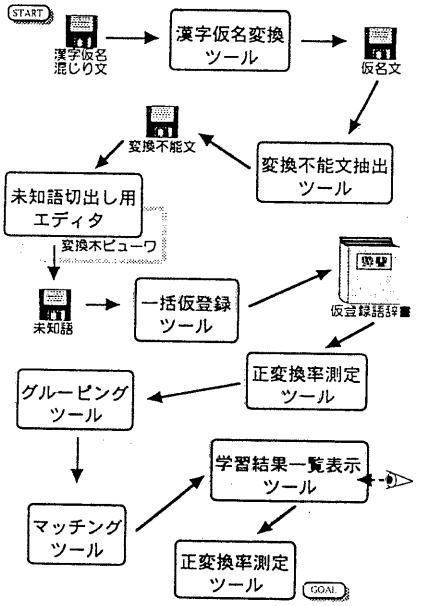

図4 ツールを使った実験の流れ

6．おわりに

本報告では、仮名漢字変換システムに自動的に単語を登録する手法を設計し、実現に向けて解決しなければならないパラメータを明らかにした。

本研究の成果を次に示す。

①読み及び変換後の表記を得るためのユーザインタフェースを設計し、実現した。

②品詞を判定するための機構を設計した。

③品詞学習機構実現のために定めなければならないパラメータを明らかにした。

④実装する仮名漢字変換システムを拡張し、辞書ハンドラの追加や扱う品詞の追加など、仮登録語を扱うための枠組みを入れた。

⑤実験用ツールセットを一部実現した。

7．今後の予定

仮名漢字変換システム側の実装準備は整ったが、5．で述べた問題点が解決するまでは実装できないので、早急に実験用ツールセットを整備して、大量のサンプルテキストを用いた実験を行なう。

自動単語登録機能は、仮登録処理、グルーピング処理、マッチング処理、本登録処理の四つのモジュールの集合であり、OS/ｏ仮名漢字変換システムＶ２に組み込まれて実現されるものである。各モジュールは実験ツール用のルーチンとして実現されており、すべての未解決パラメータが決定され次第、実装する予定である。

8．参考文献

[1] 村上本二郎：初歩の国文法（口語・文語），昇龍堂出版，1990

[2] 牧野寛，木澤誠：べた書き文の分かち書きと仮名漢字変換－二文節最長一致法による分かち書き，情報処理学会論文誌，Vol.20，No.4，pp.337-345，1979

[3] 下村秀樹，酒井貴子他：仮名漢字変換における最近使用語優先学習方式のモデル化，情報処理学会論文誌，Vol.35，No.3，pp.426-435，1994

[4] 酒井貴子，下村秀樹他：仮名漢字変換における最尤候補選択アルゴリズムの実験，情報処理学会第44回全国大会論文集4P-12，pp.191-192，1991

[5] 酒井貴子，下村秀樹他：仮名漢字変換の手法と学習に関する一評価，情報処理学会論文誌，Vol.34，No.12，pp.2489-2498，1994