# 5L−6　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

下村秀樹　酒井貴子　並木美太郎　中川正樹　高橋延匡
(東京農工大学　工学部)

## 1．はじめに

仮名漢字変換において，最近使用した語を次の変換時に優先する学習方式（以下，最近使用語優先学習方式）は，一般に広く採用されている．この方式は，変換精度の向上に非常に大きな影響を与えることが知られている[1]が，次のような点と変換精度の関係は，まだ十分に明らかにされていない．

（1）「最近」をどれくらいの範囲にすべきか

（2）変換結果の優先度を決める学習以外の情報（文節数最小，2文節最長一致など）とどちらを優先させるか

我々は，最近使用語優先学習方式をモデル化し，上記の問題の評価を試みている．本稿では，そのモデル化と「最近の範囲」と変換精度の関係についての実験を報告する．

## 2．最近使用語優先学習

### 2.1　最近使用語優先学習のモデル

最近使用語優先学習とは，「最近使用した語を含む変換結果を優先する」という学習方式である．これを簡単に次のようにモデル化する（図1）．まず，ある一連の仮名漢字変換で変換結果として選択された単語を，その順番に，

$W_1, W_2, W_3, \cdots, W_k$

とする．（注：このときの添字 1〜k は，単語を使用した論理的な時刻を示していると見ることができる．）このとき，$W_k$ からさかのぼって t 単語の範囲に含まれる語を「最近使用したので優先すべき語」と定義する．この範囲を「学習区間」，t を「学習区間長」，学習区間に含まれる語を「学習語」と呼ぶ．この「学習語」を含む変換結果を優先するのが最近使用語優先学習方式である．

### 2.2　学習語生存確率

以上のモデルで，最近使用語優先学習の効果が現れる必要条件は，変換結果として得たい語が学習区間内に含まれていることであり，その確率は最近使用語優先学習方式を評価する一つの尺度となる．以下，この確率を「学習語生存確率」と呼ぶことにする．

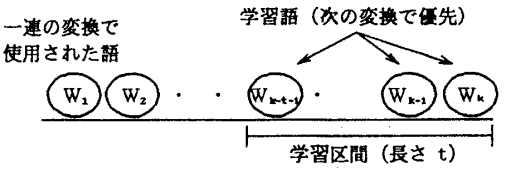


図1　最近使用語優先学習方式のモデル

この確率の推定方法は，計算機システムの仮想記憶管理のモデルから類推できる[2]．ある時刻 k における学習語生存確率 P(t,k) を k に関して平均した値 P(t)は，単語の使用間隔の確率分布から求めることができる．すなわち，ある単語を使用してから次にその単語を使用するまでの間に別の単語を X 個使用する（単語使用に関する論理的な時刻での時間間隔が X である）ときの確率分布を F(X) とすると，次のようになる．

$$P(t) = \sum_{i=1}^{t} F(i) \quad ---(1)$$

もし学習区間長を大きく採れば，学習語生存確率 P(t) は1に近付く．しかし，それに従って学習情報の格納に必要な記憶領域が大きくなること，語の優先順位の多くが語の出現順番で決まってしまうという問題点など，学習区間長をむやみに大きくすることのデメリットも大きい．したがって，学習区間長の設定は重要な問題である．

### 2.3　最近使用語優先学習の実現方式

最近使用語優先学習を行うためには，学習語の情報を格納する必要がある．その方式としては，大きく分けて次の二つが考えられる．

（1）学習語を学習用のバッファ（以下，「学習辞書」と呼ぶ）に格納する

これは，仮想記憶と同じ方式であり，学習辞書が主記憶に，基本辞書が2次記憶に相当する．一般に学習辞書には登録語数の制限があり，辞書が一杯になった場合には，LRU (Least Recently Used) 方式等で古い語を追い出す．登録可能語数の制限は，近似的な学習区間の限定と見ることができる．

（2）変換に使用したので優先すべき語であることを基本辞書に直接書き込む

この方法を実現する最も簡単な方法は，基本辞書の各単語に使用時刻を記録する領域を用意することである．学習区間の定義は，「現在の時刻からある一定時間以前に使用された語は学習語として扱わない」ことによって行える．

どの方式を採用するかは，実現上の問題が大きい．ただし学習辞書を使う方式では，語の重複使用によって学習辞書の登録可能語数よりも見かけ上の学習区間長が長くなる．したがって，学習辞書の登録可能語数を学習区間として (1)式から学習語生存確率を求めると，実際の値の下限値になることに注意が必要である．

---

An Analysis and Evaluation of The Recently Used Word Priority Method in Kana-to-Kanji Translation
Hideki Shimomura, Takako Sakai, Mitarou Namiki, Masaki Nakagawa and Nobumasa Takahashi
Department of Computer Science, Faculty of Technology,
Tokyo University of Agriculture and Technology.

### 3．単語使用間隔分布と学習語生存確率に関する実験

第２章で述べたモデルを検証するために，以下に述べる実験を行った．

#### 3.1 実験環境

実験環境として，学習辞書による最近使用語優先学習方式を持ち，またほかの情報（２文節最長一致，文節数最小など）との優先関係の変更を評価関数の変更だけで行える仮名漢字変換システムを実現した[3,4]．

#### 3.2 実験方法

（１）実験用に論文３編（情報処理学会論文２編，工学系卒業論文１編：自立語数 約 12000，2000 種）を用意する．

（２）学習辞書の登録可能語数を極端に大きく（LRU による追出しが起こらないように）し，実際に変換を行いながら，単語使用間隔の分布を測定する．またその結果から，２章の議論に従って，学習語生存確率の理論値（学習辞書を使う方式での下限値）を計算する．

（３）学習辞書の登録可能語数を実際にいくつかの値に制限して変換を行い，学習語生存確率を実測し，理論値（下限値）と比較する．

なお，学習は自立語に関してだけ行った．

#### 3.3 実験結果

実験の結果得られた，学習語生存確率の理論値（下限値）と実測値を図２に示す．まず，学習語生存確率の実測値は，登録可能語数 500 程度で 97% となった．この値は文章の単語使用状況（例えば分野の違いなど）によって変わることが予想される．しかし，この程度の比較的少ない登録可能語数でも，十分な学習効果を期待できることがわかった．

次に，学習辞書の登録可能語数から計算した下限値と実測値を比べると，誤差 0.1 前後であった．提示したモデルに基づけば，これ以外の文章に関しても，学習辞書の容量から学習語生存確率をある程度予測できると思われる．

### 4．学習語生存確率と変換性能の関係に関する実験

次に，学習辞書の登録可能語数と変換性能という観点から，文献[5]に報告した我々の実験で比較的変換率のよい組合せであった「文節数最小＋最近使用語優先」，「２文節最長一致＋最近使用優先」（どちらも前にある情報を優先する）について，学習語生存確率と変換性能の関係を調べた．実験に用いた文章と変換率の定義は，文献[5]の実験と同じように，句読点単位入力で欲しい結果が１回の変換で完全に得られた場合を成功とした．なお，基本辞書の同音語の優先順位情報は，変換結果の尤度の評価に加えずに実験を行った．

結果を図３に示す．図では，学習語生存確率と変換率がほぼ正の相関関係になっている．二つの手法間の変換性能に若干の差があるが，これは手法の能力の問題であろう．一般的には，文節数最小法の方が広い範囲からの情報を利用しているので，変換率が高くなる．

以上の結果から，基本的な手法と学習を比較的うまく組み合わせた場合，学習語生存確率は変換性能にほぼ線形に影響することがわかった．変換性能と学習区間長（学習辞書の登録可能語数）を直接結び付けるのではなく，文章の統計的性質（単語の使用間隔分布）から得られる学習語生存確率の概念を間に挟むことによって，変換性能をより解析的に議論することができると思われる．

### 5．おわりに

本稿では，仮名漢字変換における最近使用語優先学習方式をモデル化し，学習語生存確率の概念を提示するとともに，実際の文章についてその検証を行った．その結果，学習語生存確率は，単語使用間隔の分布から求めることができ，かつ変換性能を解析的に議論するうえで重要な要因であることがわかった．今後は，さらに複雑な学習方式のモデル化，また大規模なデータを用いての精密な検討を行うことが課題である．

#### 参考文献

[1]酒井他：日本語ワードプロセッサの仮名漢字変換における変換処理と精度についての考察，情処 HI 研，35-10（1991）
[2]高橋他：オペレーティングシステムの機能と構成，岩波書店（1983）
[3]下村他：OS/omicron 仮名漢字変換システム第２版の設計思想，情処 42 全大，5Q-1（1991）
[4]酒井他：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，情処 44 全大，4P-12（1992）
[5]酒井他：仮名漢字変換における変換手法と変換精度についての比較実験，情処 46 全大，5L-8（1993）

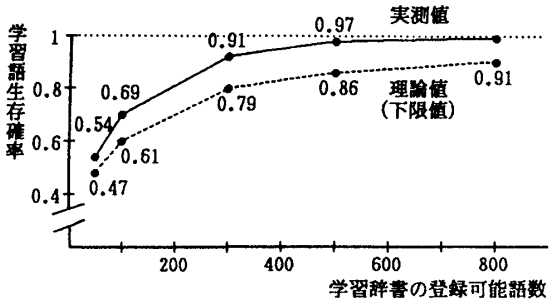


図２　学習語生存確率の理論値と実測値

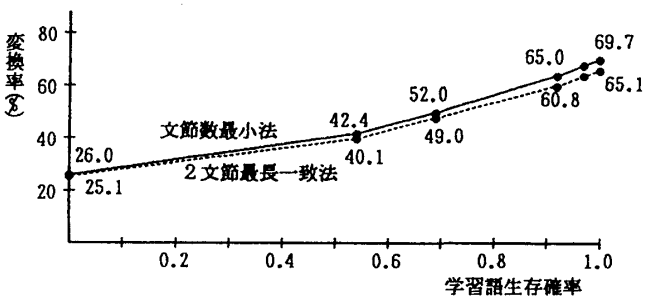


図３　学習語生存確率と変換精度（変換率）の関係