頻度情報を用いた漢字辞書の評価法
—知識ベースの漢字入力に向けて—

○堀　幸雄†
池村　匡哉†

Evaluation method for the Japanese Kanji dictionary using frequency information

—For knowledge base kanji input system—

○ Yukio Hori††
Masaya Ikemura††

Abstract

This paper reports an evaluation method for Japanese Kanji input system based on frequency information and user's input history information. Japanese Kanji Frequency information follows Zipf's law, user's input history information is follow LRU algorithm's hit rate, reports the result conducted evaluation experiment of this. And considere about a future Japanese Kanji input system.

1　はじめに

我々は人間の言語および知識獲得能力の解明とその実現を目的として研究を行なっている．このような研究はこれまでにも存在するが，現実に有効な日本語入力システムを完成するまでには至っていないのが現状である．

現在利用されている日本語入力システムでは辞書に頻度情報を考慮したものを使用し，また利用者の選択した漢字の履歴情報を考慮することが一般に使われている．利用者の履歴を学習機能と呼ぶには無理があるが，変換率の向上には良いとされている．しかしそもそもこの頻度情報の特徴について分析し，その結果を用いて「どの程度の変換率が期待できるか」等，現在は明確な基準なしに行なわれている．

そこで本論文ではまず日本語，特に単語の使用頻度が **Zipf** の法則 [1] [2] に従うという経験則をベースにその特徴について分析し，選択された単語の履歴についてはキャッシュヒット率の改善等の手段として広く研究された **LRU(Least Recently Used)**[3] を考慮してその性能を推定する．日本語入力において頻度情報，履歴情報がどの程度評価できるかについて定量的に検討し，評価実験の結果について報告する．そして今後の日本語入力方式の在るべき姿について考察する．

2　頻度情報の利用

日本語入力で使用される漢字辞書において，頻度情報を持ちはじめてのは **1978** 年に発表された日本初の日本語ワードプロセッサ **JW-10** [4] である．
漢字における頻度情報の利用とは，

1. 同音異義語の単語の中でそれぞれの単語が使用する分野や個人によって偏りがあること．
2. あるドキュメントを作成しているときに出現した同音異義語は再びそのドキュメント中で出現したときは同じ単語を使用する可能性が高いこと．

という 2 つの経験則に基づく変換率の向上を計っている．前者を長期的な頻度情報，後者を短期的な頻度情報という．短期的な頻度情報とはまさに利用者の履歴情報である．この方式を現在の日本語入力システムの大部分が何らかの形で採用している．図 1 はかな漢字変換システム **SKK** [5] の辞書の例である．

†神奈川大学院理学研究科情報科学専攻
††Department of Information Science, Grad. School of Science, Kanagawa University

長期的頻度情報 (入力システムが持つ)	短期的頻度 (履歴) 情報 (利用者が持つ)
いっせい /一斉/一世/一生/一成/一誠/	いっせい /一斉/
いこう /以降/移行/意向/移項/偉功/威光/遺稿/	いこう /移行/移項/
ふごう /符号/負号/符合/富豪/	ふごう /富豪/符号/

図 1: 頻度情報を持った辞書 (SKK 辞書)

3 頻度情報を用いた漢字辞書の評価法

3.1 実際の入力方式

これまで述べたように漢字入力において，頻度，履歴を大部分の漢字入力システムがなんらかの形で利用しているにも関わらず，その評価には頻度，履歴のどちらも考慮されていない．[6] [7]

まず簡単に漢字を入力する流れについて，SKK の場合を例に説明する．下の図 2 では実際に上の図 1 の状態の辞書を持っていたときに「移行」という漢字を決定するまでの過程である．履歴を考慮した場合 1 回目の視察で漢字が決定し，頻度情報のみの場合は 2 回の視察で目的の漢字が決定される．

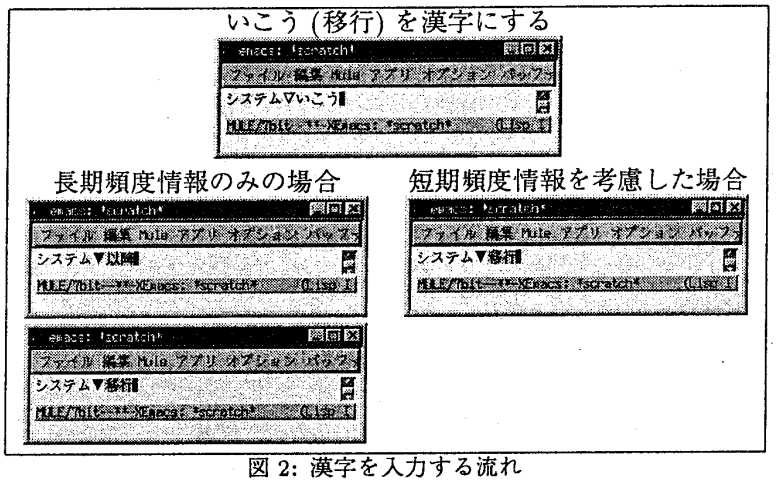

図 2: 漢字を入力する流れ

3.2 頻度，履歴を考慮した評価法

漢字の頻度情報より，どの程度の視察数が期待できるのかに Zipf の法則を用いる．この法則は f を単語の使用度数 r を使用度数の大きい法から振った順位とし C, k を頻度パターンを決定する定数としたときに $fr^k = C$ が成り立つという経験則である．これは i 番目の頻度を持つ単語の使用度数は $(1/i)^k$ に比例していく．この Zipf の法則により簡単に単語の平均順位 $\bar{r}$ を求めることができる．

$$\bar{r} = \frac{n}{logn}$$

また履歴となる短期的な頻度情報は LRU アルゴリズムを用いたキャッシュ方式のヒット率 $\mathbf{H_{LRU}(S)}$ で表わされ，使用順位 r 位の単語が履歴として記憶されている確率 $P(r)$ から推定できる．具体的には，保存する履歴の回数 D だけ，過去にその単語が表われなかった確率 $(1-P(r))^D$ を考える．

$$\begin{aligned} H_{LRU} &= \int_0^n P(r) \times (1-(1-P(r))^D) dr \\ &= \int_0^n P(r)dr - \int_0^n P(r) \times (1-P(r))^D dr \\ &= 1 - \int_0^n P(r) \times (1-P(r))^D dr \\ &= 1 - \frac{((1-k)C^{1-\frac{1}{k}})^{\frac{1}{k}}}{k} \times \int_{1-kC^{-\frac{1}{k}}}^{\infty} t^{-\frac{1}{k}}(1-t)^{\frac{S}{1-k}} dt \end{aligned}$$

ただし，$D = \frac{S}{1-k}, t = P(r), n =$ 同音異義語数

以上の考え方で，頻度，履歴に関する変換率が推定できる．頻度情報では同音異義語数が増えるに従って何回目の視察で目的の決定できるか，履歴情報ではどの程度のヒット率なのかについて実験する．実際には，世の中すべてのドキュメントに対する C, k の計測は困難であるがある程度特定のドキュメントから求めるものとする．

4 評価実験

入力データとして **NACSIS** テストコレクション (**1999** 年度版) [8] の **NTCIR1** を使用し，その中の各学会から集められた論文の日本語要旨 (約 **33** 万件，単語数 **5346** 万語) を簡単な形態素解析処理により抽出された単語を用いて，頻度情報と履歴情報による視察数，ヒット率の測定を行なった．

履歴情報を評価する **LRU** アルゴリズムのヒット率で計測を行なった結果，同音異義語数に対して 10 % ほどの履歴情報の中に目的となる漢字が含まれる確率が約 50 % 程となっている．

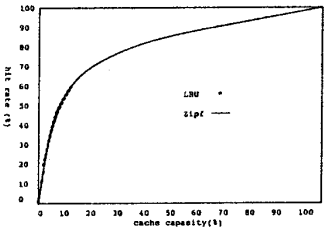

図 3: 履歴によるヒット率

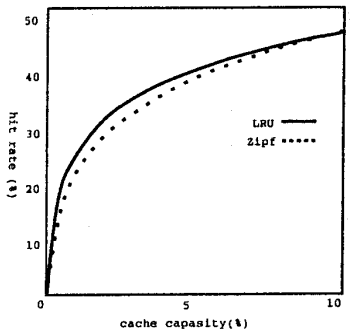

図 4: キャッシュ容量 (履歴保持数) 10% までの拡大図

頻度情報による視察字数の評価では **Zipf** の法則では過小評価してしまう結果となった．同音異義語数が 100 個存在しても平均で 10 回の視察で目的の漢字を決定している．(図 3)

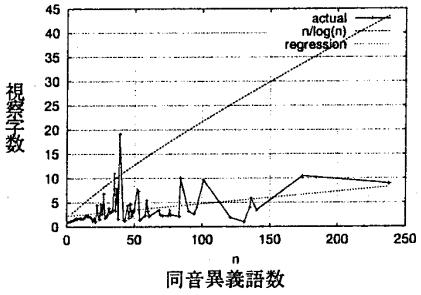

図 5: 頻度情報による視察数

5 頻度情報の辞書の評価のまとめ

今回の評価実験では日本語入力において文法知識，意味解析を一切行なわない **SKK** の評価となるものである．頻度情報は単語の局所参照性より有益な情報ではあるがその限界は示されている．また漢字が整列されるような情報 (画数，面数，品数) を持てば 2 分探索 ($\log_2 n$) で目的の漢字を探すことが可能となるが，その利用方法については検討を要するものである．今後詳細なデータの収集と評価モデルの改良が残された問題である．

6 知識ベースの辞書

現在の日本語入力システムの主流である「かな漢字変換システム」は，履歴学習型である．これは，あらかじめ選定された辞書の並び順をもとに，利用者が確定した漢字の優先順位を上げることで，システムを使っている利用者がよく利用する漢字の出てくる順番を早めていく．それにより，利用者にとって使いやすい辞書を「育て上げていく」ことができる．

計算機の処理能力の向上により，辞書の語彙の増加，変化アルゴリズムの改良が重ねられ，平均 90 % 以上の確率 (ATOK 使用) で希望する漢字を変換できるようになっている．さらに近年，かな漢字変換システム ATOK においては「ATOK 監修委員会」がおかれ，国文学の立場から辞書の選定が行われるようになった．それをふまえ，より使いやすい，文章作成支援としての日本語入力システムに必要なものとして，知識ベースの辞書を提案する．

現在までのかな漢字変換システムは，共起処理などの変換アルゴリズムの進歩によってその変換効率を高めてきた．同時に辞書に含まれる語彙数も飛躍的に増加した．辞書の選定にも国文学などの専門家があたるようになり，急速に発展してきた「書き言葉/打ち言葉」における日本語の規範としてふさわしいものになってきた．

だがまだ，人間が鉛筆で紙に文章を書くのと同じ感覚で，キーボードから文章を打ち込むようにはなっていない．その差異として考えられるのが，語彙のデータ，すなわち辞書が構造化されているか否か，ということである．人間は漢字かな混じり文を書くとき，その漢字自体の意味を知っており，使い分けることができる．これは，漢字自体の意味という知識を持っているためで，ある．それは，部首・総画数・つくりといった文字自体の知識であったり，漢字の間の微妙な意味の差異であったりする．

例えば,「兄弟」「姉妹」はそれぞれ「きょうだい」と読むが，後者は「しまい」とも読む．これは，単純に「きょうだい」という言葉だけでなく，その言葉の持つ意味あいを知らないと正確な文字を書くことはできない．ではなぜこれらの使い分けができるかというと，それは人間がそれぞれの意味を知っているからである．このように，かな漢字変換システム用の辞書にも構造化された知識のデータを加え，これを変換に使うことでより手書きに近い感覚での入力が実現する．知識の構造化には，C-TRAN 法，SS-KWEIC 法，SS-SANS 法の格手法を用いる予定である．

7 今後の展望

日本語入力システムとして,「かな漢字変換システム」は「漢字変換プログラム」に偏るきらいがある．よくいわれる,「パソコンで文章を作成すると漢字が多くなる」ことのゆえんである．もちろん，それは多くの語彙を持っていることの証明でもあり，漢字が多くなるのは，利用者が安易に変換された漢字をそのまま使ってしまうことが原因の一つである．文章を書くときに，大きな辞書を見ながら，あらゆるかなを漢字に変換している行為に近い．ユーザが使いやすいように，支援するのが目的であるから，それは

加えて，通常の定型文書やメールなどの日常の文章を書くエンドユーザと，文筆家やライターなどのエンドユーザでは，使う辞書も異なって当然である，ということがある．説明に必要な漢字と，描写に必要な漢字の違いなど，ユーザによって必要な漢字，日本語は異なってくる.「はこ」という漢字は，通常なら「箱」であるが小説家なら「匣」かもしれない．後者は通常「はこ」の読みでは変換することはできず，JIS コードから探す等の手段を持って見つけ，登録して使うことになる．もちろん，辞書は各自の癖をも学習していくものであるから，こうして「育てて」いくことが必要であるというのは一つの意見である．しかし，不完全であるものを，ユーザが手間と時間を掛けてようやく使えるものになるというのは製品としておかしいのではないだろうか．よりユーザに近いレベルで，その目的にあった分野での辞書選択，また辞書の構造化が必要である．

また，日本語が漢字かな混じり文であることを考えると，日本語入力システムもかな文字を漢字に変換するだけでなく,「漢字かな混じり文」を製作するための道具であることが望ましい．そこで，文章，単語に関する知識をもった辞書とそれを生かすことのできる変換アルゴリズムの搭載で，思考を中断するのではなくそれを助けることができるようなシステムの構築を目指す．その為の知識ベース辞書である．

参考文献

[1] 水谷静夫：数理言語学 培風館 1982

[2] 影浦峡：計量情報学 —図書館/言語研究への応用— 丸善 2000

[3] 西川記史 細川貴史 森靖英 吉田健一 辻洋: WWW トラフィック特性に基づくキャッシュ方式の提案 情報処理学会論文誌 Vol. 41 No. 9 pp.2625-2637 2000

[4] 森健一 八木橋利昭： 日本語ワープロの誕生 丸善 1989

[5] 佐藤雅彦：かな漢字変換システム SKK bit Vol.23,No.5 共立出版

[6] 中山剛 黒須正明：日本文入力方式評価法の研究情報処理学会論文誌 Vol.26 No.11 1985

[7] 森田正典：各種日本文入力方式の性能の定量的比較電子情報通信学会論文誌 D Vol. J70-D No.11 pp.2182-2190 1987

[8] NACSIS テストコレクション: http://research.nii.ac.jp/ntcir/index-j.html

[9] 電脳辞書の国語学: 箭内敏夫 おうふう

[10] 日本語学のみかた: 朝日新聞社 1997

堀 幸雄 (1,2,3,4,5 節)　神奈川大学院理学研究科情報科学専攻
Yukio Hori　(*horiyuki@goto.info.kanagawa-u.ac.jp*)
池村 匡哉 (6,7 節)　神奈川大学院理学研究科情報科学専攻
Ikemura Masaya　(*ike@goto.info.kanagawa-u.ac.jp*)