

帰納的学習による入力文予測を用いた中国語ピンイン入力手法

烏 日那[†] 荒木 健治[†] 柄内 香次[‡]

[†] 北海道大学大学院工学研究科 〒060-8628 札幌市北区北13条西8丁目

[‡] 北海学園大学大学院経営学研究科 〒062-8605 札幌市豊平区旭町 4-1-40

E-mail: [†] {wrn, araki}@media.eng.hokudai.ac.jp, [‡] tochinai@econ.hokkai-s-u.ac.jp

あらまし 中国語のピンイン入力システムにおいては、打鍵数が多い及び語候補から選択することは入力スピード向上の制限になっている。本研究では、以上の現象に着目し、帰納的学習による入力文予測を用いた中国語ピンイン入力手法を提案する。本システムは、帰納的学習を用いて既に入力された文章の字面情報により、予測用のルール（文字列）を獲得し、ルール辞書を自動生成する。また、隣接文字列の情報を利用してルール再生成を行うことにより、正予測率の向上を図っている。ルールを再生成する際、文脈に依存した文字列を獲得し、ユーザに動的に適應することができる。正予測率が40%の場合、打鍵数が39%程度減少していることから、本手法は入力スピードを向上させるために有効であると考えられる。ルール再生成を行う場合と行わない場合の正予測率を比較する実験を行った結果、後者の正予測率が平均4%程度増加していることが確認された。

キーワード 帰納的学習 入力文予測 予測候補 ルール再生成

A Chinese PinYin-Input Method for Association of Input Word Using Inductive Learning

Rina WU[†] Kenji ARAKI[†] and Koji TOCHINAI[‡]

[†] Graduate School of Engineering, Hokkaido University Kita 13 Nishi 8, Kita-ku, Sapporo-shi, 060-8628 Japan

[‡] Graduate School of Business Administration, Hokkai-Gakuen University
Asahimachi 4-1-40, Toyohira-ku, Sapporo, 062-8605 Japan

E-mail: [†] {wrn, araki}@media.eng.hokudai.ac.jp, [‡] tochinai@econ.hokkai-s-u.ac.jp

Abstract We introduce an efficient Chinese PinYin input method of intelligent association using Inductive Learning. As for the Chinese PinYin input system, a large number of keystrokes and the selection from candidate words limit the speed increasing. Association of input words is one of the most important methods for Chinese character input to decrease keystrokes and to improve the input speed. Our proposed method enables us to promptly input the Chinese words into a computer, that is, not necessary to spell complete character of PinYin only by choosing the words from candidates. Moreover, the Inductive Learning approach gives system two characteristics. One is the dynamically adaptation to current situation, and the other is the capability of acquisition rules even if no initial rules. Furthermore, we make use of the relationship between Chinese words and characters in context to generate a rule. When the correct association rate is 40%, the decreasing rate of keystrokes is 39% in the experiment for the performance evaluation. It shows that this method is effective for Chinese input.

Keyword Candidate rule, Intelligent association, Inductive Learning, Rule generation

1. まえがき

計算機の普及に伴い、計算機入力システムの研究及び開発が盛んに行われている[1][2][3][4]。中国語には、ピンインと漢字があり、ピンインはローマ字を用いて漢字の読み方を表す手段で、それぞれ1から6文字のローマ字によって構成される。漢字は中国語のテキストを構成する唯一の文字で、現在約4万字存在している。英語と同じように中国語の漢字をキーボードを用いて計算機へ直接入力することはできないため、何らかの入力手法が必要である。現存の主要な中国語入力手法としてはピンイン入力手法と字形入力手法の2

種類の方法が存在する。その中でピンイン入力手法が約93%を占め、一番普及している入力方法になっている。通常の中国語ピンイン入力手法のプロセスを以下に示す。

- (1) ピンインを構成しているローマ字を入力する。
- (2) ピンイン漢字変換を行う。
- (3) 変換結果の語候補から正しい語を手で選択する。

中国語の漢字には同音異型字が多いため、ピンイン

で入力する際候補から選択する手間が掛かり、連続入力ができないという欠点がある。また、平均で一文字漢字のピンインは3文字のローマ字から構成されるので、打鍵数が多くなり、ピンイン入力システムのスピードの向上も困難であると考えられる。入力語予測手法は、上述の問題を解決し、入力スピードを向上させる有効な方法として、それに対する研究[1]も盛んに行われている。

入力語予測手法として、基本的な方法は予めシステムに予測語辞書を与え、漢字を入力すると辞書を参照し次の入力漢字を予測する。このようなシステムにおいては、予測を精度よく行うために大規模の辞書を作成する必要がある。しかしながら、予めシステムに与えた辞書は、作成者の知識に大きく依存し、全ての言語現象を辞書に与えること及び全てのユーザの言葉遣いに対応することは困難である。このような問題を解決するために、統計的手法に基づく研究[1]が行われている。この手法においては、予測を精度よく行うために大量のコーパスが必要である。しかしながら、これはコーパスを作成する労力が必要である。また、科学技術の発展に伴い、各分野に新用語や専門用語が日々新出語として出現する。固定されたコーパスにおいて、統計的手法により獲得された知識によって、全ての分野、全てのユーザに適應できることは困難である。

これまで、中国で公開されている入力語予測機能を持つピンイン入力システムの一つとしては“智能ABC入力手法[注1]”が広く使用されている。このシステムは、入力文のピンインを全て入力した後、そのピンインを用いて予めシステムに与えた辞書を参照することにより適当な語を予測候補として出力する。ユーザが候補から正しい語を順次選択することにより連続入力を実現する。しかしながら、全てのピンインを入力する必要があるだけでなく、予測候補から正しい入力語を選択する手間も掛かる面では、入力スピードの向上は有限であると考えられる。そこで、我々は中国語ピンイン入力において入力の打鍵数を減らし、変換を精度よく行うための入力手法として、帰納的学習[5]による入力語予測を用いた中国語ピンイン入力手法を提案する。

本手法においては、システムが帰納的学習を用いて、既に入力された文の字面情報から共通部分と差異部分を抽出することによりルールを獲得する。更に、直前に入力した語と獲得されたルールを用いて次の入力を予測する。帰納的学習手法の自然言語処理においての有効性はすでに確認されている[5][6]。予測用の文字

列を帰納的学習により獲得することの利点としては、システムはユーザがすでに入力した文の字面情報により、学習を行って知識を獲得し、更に、この知識を利用して次の入力で役立てるため、大量のコーパスを作成する必要がない。また、動的にユーザに適應することができるということが挙げられる。

2. 概要

本研究で提案するシステムを図1に示す。システムは変換処理、校正処理、予測処理、学習処理、フィードバック処理によって構成される。

本手法においては、学習により得られたルールをルール辞書に登録し、ユーザの直前に入力した文字列を用いてルール辞書を検索することにより次の入力を予測する。また、隣接文字列情報により、ルール再生成を行い、文脈に依存した予測用の文字列を獲得する。ルール再生成については以下の3.2.3.節で説明する。これによって、正予測率及び動的にユーザに適應する能力の向上を実現している。

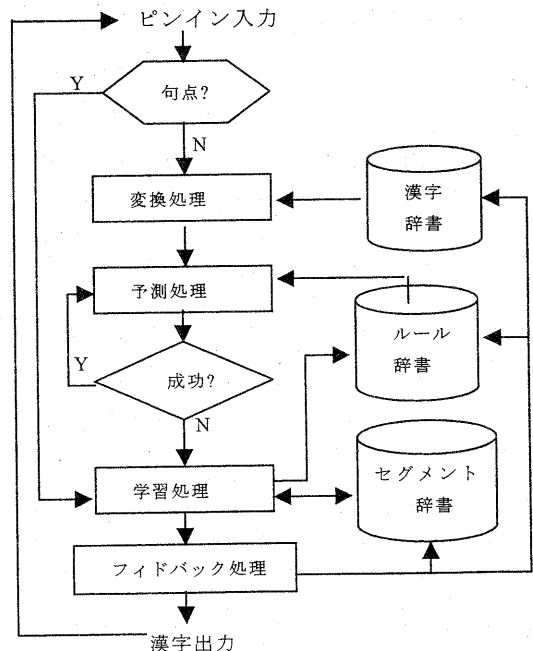


図1 システムのプロセス

3. 処理方法

3.1. 変換処理

変換処理では、入力されたピンインを用いて、漢字辞書を参照することによりピンイン漢字変換を行う。

[注1]本手法の情報は以下のホームページより獲得した。

<http://www.znabc.com>.

漢字辞書は、予めシステムに与えられている。これは中国語基本漢字に基づいて作成されており、漢字の最近使用された情報によって決められた尤度情報を持っている。変換は一文字を単位として行う。ピンインを入力すると漢字辞書を参照しながらそれに当てはまる全ての同音異型字を求め、システムは尤度が一番高い字を入力漢字として自動的決定するとともに、残りの同音異型字を候補として出力する。システムが自動決定を誤った場合、ユーザが校正処理を行う。この処理では手動で候補から入力漢字を選択すると同時に、選択された漢字の尤度を最も高くし、次の入力に役立てる。

3.2. 学習処理

学習処理では、既に入力された文の字面情報により、帰納的学習を用いてルールを獲得する。この際、文の入力ごとにこの文を用いて既に入力された文に対して検索を行い、2文字以上一致した場合、それらを語として抽出する。これをW(Word)と呼び、Wの前後の文字列をそれぞれFW(Front Word),BW(Behind Word)と呼ぶ。更に、システムはW,FW,BWを用いてルールを獲得し、式(1)に示す尤度評価式により獲得されたルールを尤度付でルール辞書に登録する。差異共通部分を抽出する例を表1に示す。表1の文2は既に入力された文の中の一文である。文1を入力するとこれを用いて文2の検索を行い、一致した部分(κ η θ ι)をWとして抽出し、差異部分をそれぞれ文1に対してはFW1(γ δ ε),BW1(κ λ μ)と記述し、文2に対してはFW2(μ ν ξ ο),BW2(π ρ σ)と記述する。

表1: 差異共通部分の抽出例

文1	γ δ ε ξ η θ ι κ λ μ
文2	μ ν ξ ο ζ η θ ι π ρ σ
W	ξ η θ ι
FW1	γ δ ε
FW2	μ ν ξ ο
BW1	κ λ μ
BW2	π ρ σ

3.2.1. 新セグメントの獲得

セグメント辞書に登録されているセグメント(以下はSと呼ぶ)は、共通部分として抽出されたWであり、単語になる確率の高い文字列である。これは次に述べる処理でルール再生成の際使用される。Wの場合は、Wが5文字以内(5を含む)の場合Sとしてセグメント辞書に登録する。Wが5文字以上の場合、セグメント辞書に登録されているSを用いてマッチングすることにより、Wを分割し、マッチしたSの尤度数を1増加

させ、残りの部分をNS(New Segment)としてセグメント辞書に登録する。セグメントを獲得する例を表2に示す。表2でWが5文字以上(ν ξ ο π ρ σ)の場合、セグメント辞書のS(ο π)を用いて分割し、新たにν ξ, ρ σというNSを獲得する。Wが5文字以内(ν φ χ φ)の場合、新たに(ν φ χ φ)をNSとしてセグメント辞書に登録する。登録されるNSの尤度数は初期値として1と設定しており、マッチングを行うごとにその尤度数が1増加する。

表2: セグメントの獲得例

W	ν ξ ο π ρ σ	ν φ χ φ
Wの文字数	>5	<5
S	ο π	
NS	ν ξ, ρ σ	ν φ χ φ

3.2.2. ルールの獲得

FWとBWの文字数によりそれぞれFW+W+BW,FW+W,W+BW,W等4種類のルールを獲得する。獲得されるルールと対応する条件を表3に示す。本稿ではFWとBWの文字数をそれぞれNFW(words Number of FW)とNBW(words Number of BW)と呼ぶ。文字数の条件を5と限定している原因は、まず、中国語では4文字の慣用語が1万個以上であり、これらの慣用語はほとんど1文字の助詞と一緒に使用されている。次は、中国語の単語はほとんど2文字と3文字の漢字で構成されており、また、多数の複合語はこれらの単語により構成されている。以下の例は中国語の一文を単語に分割した結果である。この例では、下線の部分は2文字あるいは3文字の単語を示し、括弧の中は複合語を示す。

例:

漢語 是 世界上 最 (古老的 語言) 之一, 但是 (現代 漢語的) (語法 理論), 大多 取材 于 (西方 語言的) (語言學 理論), 並不 (十分 適合) (漢語 分析的) 需要。

以上の例では、2文字の単語は15個(62%)、3文字の単語は7個(29%)であり、複合語はほとんど4文字あるいは5文字で構成されている。

表3に示す通り、システムがNFWとNBWによってルールを獲得する手順は:

- (4) NFWとNBWが全て5文字以内の場合、FW+W+BWというルールを獲得する。
- (5) NFWが5文字以上、NBWが5文字以内の場合、W+BWというルールを獲得する。
- (6) NBWが5文字以上、NFWが5文字以内の場合、FW+Wというルールを獲得する。
- (7) NFWとNBWが全て5文字以上の場合、Wというルールを獲得する。

ルール辞書はルールの形によって二つに階層化されており、それぞれのルールの形は表3に示す通り第一階層ではW+BWとWの2種類であり、第二階層ではFW+W+BWとFW+Wの2種類である。

表3：獲得するルールとその条件

	NBW	0<NFW<=5	5<NFW
NFW			
0<NBW<=5		FW+W+BW	W+BW
5<NBW		FW+W	W
階層		第二階層	第一階層

ルール辞書に登録されているルールの尤度値Vは以下の式(1)により定義されている。

$$V = \alpha \times A - \beta \times B + \gamma \times F + L \quad \dots\dots(1)$$

A：正予測頻度

B：誤予測頻度

F：出現頻度

L：ルールの文字数

α, β, γ ：係数

3.2.3. ルールの再生成

ルールの再生成は、学習処理で獲得されたFWあるいはBWの文字数が5文字以上の場合、セグメント辞書に登録されているSを用いて分割することにより、尤度が高いFW+W+BW形のルールを新たに生成することである。ルール再生成の例を表4に示す。表4で、“ηθικ”はセグメント辞書に登録されている単語になる確率が高い文字列である。これを用いて文字列(γδεζηθικμνξοπ)を分割し(γδεζ+ηθικ+μνξοπ)というルールを獲得する。

複数のセグメントが同時にマッチした場合、システムはマッチしたSの文字数と頻度数及びマッチした位置の情報によって新ルールを決定する。マッチング位置Pは次の式(2)によって決定される。

$$P = \frac{|Lfs - Lbs|}{Lfs + Lbs} \quad \dots\dots(2)$$

Lfs：マッチした位置の前に存在する文字列の長さ

Lbs：マッチした位置の後に存在する文字列の長さ

システムは以下の手順で新ルールを判断する。

- (1) マッチされたセグメントの頻度数が最大
- (2) Pが最小
- (3) マッチされたセグメントが最長

表4：ルール再生成の例

FW(BW)	γδεζηθικμνξοπ
S	ηθικ
New rule	γδεζ+ηθικ+μνξοπ

3.3. 予測処理

予測処理では、既に入力された漢字列の最後の5文字を用いてルール辞書の検索を行い、尤度が高い順にルールを検出し、予測語候補として出力すると共に、尤度値が最も高いルールをシステムが自動的に予測語として出力する。システムの自動選択が失敗した場合、ユーザが予測語候補から正しい語を選択することにより入力語の予測を実現している。予測の際、まず、入力された文字列の最後の1文字及び2文字を用いてルール辞書の第一階層を参照し、マッチするルールを得る。次に、その後ろからそれぞれ2, 3, 4, 5文字を用いてルール辞書の第二階層を参照しマッチするルールを得る。最後に、得られたルールを尤度の高い順に予測候補として出力するとともに、システムは尤度値が最高のルールを自動的に次の入力語として決定する。システムの判断が正しい場合、そのルールの正予測頻度を1増加させる。誤った場合、ユーザの選択により正しい語を決定し、ユーザに選択されたルールの正予測頻度を1増加させるとともに、システムに選択されたルールの誤予測頻度を1増加させる。

3.4. フィードバック処理

フィードバック処理では、人手により校正済みの結果に含まれる情報を辞書に登録し、漢字辞書とルール辞書それぞれの尤度を更新する。まず、変換処理でユーザにより選択された漢字の同音異型字列中の順位を一番高くし、次回の変換に役立てる。次に、予測処理で正予測されたルールの正予測頻度を1増加させ、誤予測されたルールの誤予測頻度を1増加させることにより、回次の予測に役立てる。このように、フィードバック処理により、辞書の活性化を行い、システムをユーザに動的に適応させる。

4. 評価実験

本手法の有効性を確認するため、本手法に基づく処理のプロセスを実現する実験システムを作成し、評価実験を行った。

4.1. 予備実験

尤度評価式(1)の係数 α, β, γ の値を決定するために予備実験を行った。実験に用いたデータは、中国語で書かれた論文[注2]の8,000文字である。まず、 $\beta = 1, \gamma = 1$ として、 α の値を変化させて実験を行った。 α の値を変化させるごとに、ルール辞書を空にして実験を繰り返した。また、同一の条件・方法で、同一の実験データを用いて実験を行い、 β と γ の値を決

[注2] 実験データは次のホームページより獲得した
<http://www.zizw.net>

定した。実験結果を表5に示す。表5より、平均正予測率が最大となった係数の値を適用した。 α, β, γ はそれぞれ $\alpha=5, \beta=7, \gamma=2$ の値を用いた。

表5 予備実験結果

α	1	2	5	10	15	20	25
平均正予測率(%)	18.9	18.7	19.7	19.5	19.0	18.8	18.8
β	1	2	3	5	7	9	10
平均正予測率(%)	18.9	19.4	19.0	20.5	21.3	21.0	21.0
γ	1	2	3	5	7	9	10
平均正予測率(%)	18.9	20.0	19.7	19.7	19.3	18.0	18.6

4.2. 実験方法

実験は、中国語の6,870個の基本漢字を含む漢字辞書を作成し、学習の有効性を確認するために、ルール辞書とセグメント辞書は空の状態から行った。実験では、中国語で書かれた三分野の文章約6万字[注3]のデータを用いて評価実験を行った。また、本実験では、予測候補を20個と限定している。これは、獲得されるルール数の増加に従い、予測候補数が無限に増大することを防止するためである。

4.2.1. 予測性能の評価

以下の式(3)に示す正予測率を用いてシステムの予測能力の評価を行った。

正予測率 = 正予測文字数 / 入力総文字数 …… (3)

正予測文字数：予測された候補の中に正解が含まれている場合の正解の総文字数

入力総文字数：予測された文字数を含む入力総文字数

4.2.2. 打鍵数の評価

以下の式(4)に示す打鍵数減少率を用いてシステムの入力スピードにおける有効性の評価を行った。

$$DKR = CAR(1 - \frac{1}{\omega \times AL}) \quad \dots\dots (4)$$

DKR：打鍵数減少率

CAR：正予測率

ω ：漢字のピンインを構成するローマ字数（本実験では、平均値 $\omega=3.1$ を用いた）

AL：予測された文字列の平均長

[注3] 実験データは次のホームページより獲得した。

<http://www.zjzw.net>, <http://www.hnclnp.com>,
and <http://www.ahetc.gov.cn/>

4.2.3. ルール再生成の有効性の評価

ルール再生成の有効性を確認するために、ルール再生成を行う場合と行わない場合の正予測率の評価を行った。4.2.の実験データを用いて2万文字を入力してからルール再生成なしで実験を行って、ルール再生成を行う場合の実験結果と比較した。

4.3. 実験結果

実験結果をそれぞれ図2、図3、図4に示す。図2は正予測率と入力文字数の関係を示し、図3は打鍵数の減少率と入力文字数の関係を表す。また、図4にルール再生成を行う場合と行わない場合の正予測率の比較結果を示す。図4は、入力文字数が2万文字から6万文字までの4万文字を用いて比較実験を行った結果である。

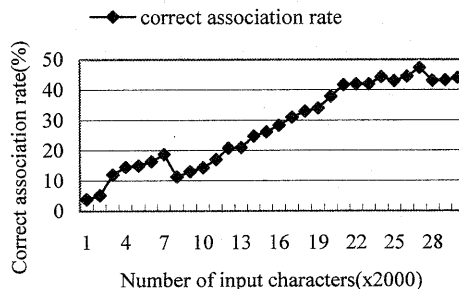


図2: 正予測率と入力文字数

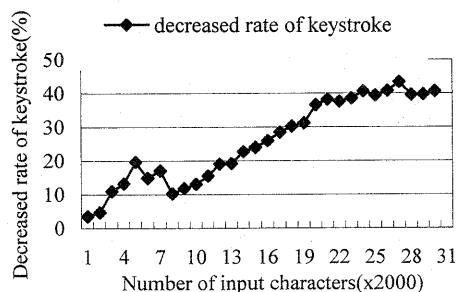


図3: 打鍵数の減少率と入力文字数

5. 考察

図2からわかるように、入力文字数の増加に従い、ルール辞書に登録されたルール数が増加し、予測に使用されるルールが増大するとともに、正予測率が向上している。この結果からシステムの学習の有効性が確

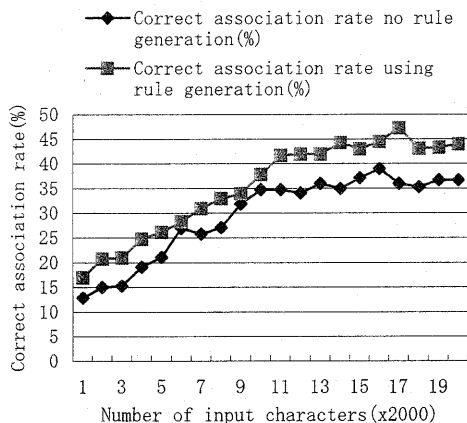


図 4: ルール再生成前後の正予測率

認された。また、図 3 で示したように、打鍵数減少率が入力文字数の増加に従って増加していることから、打鍵数の減少により入力スピードの向上を図る本研究の目的が実現されたと考えられる。同時に、図 4 からルール再生成を行うことにより、正予測率が約 4% 増加していることが確認された。

また、図 2 から分かるように、4 万文字までは、ルール数の増加に従い、正予測率が向上している。しかしながら、4 万文字以後は、ルールが増加しても正予測率の向上が低く、5 万文字から 6 万文字の間は 43% 前後で安定している。これは、4 万文字入力までは、学習により獲得されたルール数が増加し、適用ルール数も増加している。従って、予測が成功する確率が高くなり、正予測率が高くなる。しかしながら、本実験では予測候補数を 20 個と限定しているため、ルール数がある程度増加すると、予測候補が多くなり、予測を行っても成功する確率が低くなり、正予測率の増加が緩慢になっていると考えられる。

図 2 では、入力文字数が 5.4 万の点で正予測率が 47.3% になっている。これは、実験データの中で、5.3 万から 5.4 万の 1 千文字の間に文字列の重複がかなり多いため予測の成功率が高くなっていることである。

6. むすび

本稿では、入力語予測の中国語ピンイン入力システムにおける利点に着目し、帰納的学習による入力文字予測を用いたユーザに動的に適應できる中国語ピンイン入力手法を提案し、その実現方法について述べた。システムの同音異型字候補からシステムの自動選択の精度を向上させるために、漢字辞書に尤度情報を与え、フィードバック処理により、同音異型漢字の尤度を変更し、

次の自動選択に役立てる。また、フィードバック処理では、ルールの尤度を変更し、正予測率を向上させることにより、入力打鍵数の減少を実現した。また、本手法では、隣接語の情報を利用してルール再生成を行うことにより、適応性の高いルールを獲得している。以上の基本的な考え方に基づくシステムを作成し、ルール辞書が空の状態では評価実験を行った結果、入力文字数の増加に従い、正予測率が向上し、6 万文字入力すると約 40% の正予測率が得られ、本手法の有効性が確認された。ルール再生成を行う場合と行わない場合の正予測率を比較する実験を行った結果、後者の正予測率が平均 4% ぐらい増加していることが確認された。それとともに、正予測率の向上に従い、打鍵数が大幅に減少していることも確認された。

今後は、ルールを再生成する手段を再検討し、正予測率を向上させ、より実用的なシステムに向けて研究を進めて行く予定である。

文 献

- [1] Liu Changsong, Wu Zhenjun, Qiao Chunlei, Li Yuanxiang, "Intelligent Association for Chinese Input Using Statistical Method," Journal of Chinese Information Processing Vol.14 No.1, pp32-38, (2000).
- [2] Masui, T. Integrating Pen Operations for Composition by Example. In proceedings of the ACM symposium on User Interface Software and Technology (UIST'98) (November 1998), ACM Press, pp.211-212.
- [3] Zhao lei, Lu Qiang, Yang Jiwen, Zhu Qiaoming, "The Design of the Chinese Character Input Method Class," Journal of Chinese Information Processing, Vol.10 No.4, pp40-49, (1996).
- [4] Jun Wu, Zuoying Wang, Jin Guo, Zhengxian Wang, "Smart Spelling Inputting Method---A Chinese Character Inputting Method Based on Language Understanding," Journal of Chinese Information Processing, Vol.10 No.2, pp56-61, (1996).
- [5] 荒木健治, 高橋祐治, 桒内佳雄, 桒内香次, "帰納的学習を用いたべた書き文のかな漢字変換," 信学論, Vol.J79-D-II, No.3, pp.391-402, March 1996.
- [6] K. Araki and K. Tochinal, Effectiveness of Natural Language Processing Method Using Inductive Learning, Proceedings of the IASTED International Conference ARTIFICIAL INTELLIGENCE AND SOFT COMPUTING, pp.295-300, May, 2001, Cancun, Mexico.
- [7] 荒木健治, 桒内香次, 永田邦一, "多段階分割によるべた書き日本語文のかな漢字変換," 情処学論, Vol.28, No.4, pp.412-421, Apr 1987.
- [8] 王忠建, 荒木健治, 桒内香次, "帰納的学習を用いた単語分割手法の中国語への適用とその性能評価," 電子情報通信学会論文誌 D-II, Vol. J85-D-II, No.1, pp.56-65 (2002).