Vol. 42 No. 9 情報処理学会論文誌 Sep. 2001

# 日本語形態素解析の分類問題への変換とその解法

新 納 浩 幸†

本論文では日本語形態素解析が分類問題として取り扱えることを示し，決定リストを利用してその問題を解くことを試みる．日本語形態素解析は単語分割と分割された単語への品詞付けの 2 つの処理から成り立っている．入力文中の単語を構成している各文字に対して，S（ 開始文字），M（ 中間文字），E（ 終了文字）そして I（ その文字自身が単語）のいずれかの記号を付与することで，単語分割が可能になる．また品詞ごとに上記 4 つの記号を用意すれば，同時に品詞付けも行える．つまり日本語形態素解析は入力文の各文字に，前述した記号を付与する単一の分類問題に変換できる．ここでは帰納学習手法の1 つである決定リストを利用して，訓練データから分類規則を学習し，その規則を利用して形態素解析を行った．1,000 文の解析結果を形態素解析システム「 茶筌」による解析結果と比較したところ，ほぼ同等の精度を得た．また「 茶筌」による解析結果を本手法により修正するという形をとれば，最終的に得られた結果は「 茶筌」よりも精度が良かった．

# Conversion of Japanese morphological analysis into classification problem and its solving

Hiroyuki Shinnou†,?

In this paper, we propose a new method for morphological analysis of Japanese texts. We convert Japanese morphological analysis into classification problem which is solved by the decision list method. Japanese morphological analysis consists of two processes: word segmentation and assignment of the part of speech to each segmented word. We can segment a sentence into words by assigning one of four symbols, which are S (start point of word), M (middle point of word), E (end point of word) and I (that character is the word), to every character in the sentence. By setting up these four symbols for every part of speech, we can also assign the part of speech to the word simultaneously. Therefore, Japanese morphological analysis can be converted into a classification problem. To solve the classification problem, we use the decision list method, which is inductive learning method of sorts. In experiment, we compared the results of our method for 1,000 test sentences with the results of the Chasen system. The result were comparable for the two systems. Our method was also applied to the correction of the Chasen system output, which lead to improvement.

## 1. はじめに

日本語情報処理において形態素解析技術は重要な要素技術であり，従来より様々な手法が試みられている．本論文でも1 つの新しい形態素解析手法を提案する．本手法は日本語形態素解析を入力文の各文字にクラスを付与するという分類問題としてとらえ，それを解くことで形態素解析を行う．

日本語形態素解析は，一般に，辞書を用いて可能な単語分割と品詞付けの組合わせを求め，その中から尤もらしいものを選択することにより行われる．そしてその選択の規則の優劣が精度を左右する．しかしその規則を人手で構築するには，コストや一貫性の点で困難であり，統計的言語モデルを用いた日本語形態素解析法の研究が提案された [1]．これによって日本語形態素解析はほぼ実用の域に達したと思われる．ただし日本語形態素解析には未知語の問題が残されており，未知語を扱える枠組みをもった形態素解析法が望まれている．

未知語を取り扱う最も単純な手法は，辞書を利用しない文字ベースの手法を採用することである．ただし文字ベースの手法は，形態素解析のタスクを単語分割に限定し，品詞付けに対しては別の処理を必要とすることが多い [2]．また 文字ベースの Hidden Markov Model (HMM) も提案されている [3] が，HMM の場合，状態遷移確率やシンボル出力確率を n-gram から得ている．n-gram では字種や括弧内の表現かどうか

† 茨城大学工学部システム工学科
Department of Systems Engineering, Faculty of Engineering, Ibaraki University

などといった n-gram 以外の情報が扱いにくい．

ここでは日本語形態素解析を文字にクラスを付与する分類問題として取り扱う．これは文字ベースの手法でもあるために，未知語に対して頑健である．また本手法は単語分割と品詞付けの 2 つのタスクを単一の分類問題により同時に解決する．この点で，単語分割と品詞付けの各々に対して分類器を作成する従来の k-近傍法 [4] や決定木 [5] を用いた手法とは異なっている．また分類問題として扱うことで，n-gram では扱いにくい情報も容易に扱える枠組みを持つ．また分類問題は機械学習や統計学の分野で活発に研究されているテーマであり，分類問題として取り扱うことができれば，そこでの研究成果を直接応用することができる．特に，近年，マージン最大化戦略に基づくサポートベクトルマシン [6] やアダブースト [7] の研究が盛んであり，分類問題に対してはこれらの手法を適用できる．

本手法の設定では，判別するクラスが結果的に 2,552 種類にもなる．このようにクラスの種類が多数である場合に，従来主流であった決定木 [8] や最大エントロピー法 [9] の利用は現実的に困難であり，どのような学習戦略を用いれば設定した分類問題を解けるかは明らかではない．ここでは決定リスト [10] を利用した．これによって，このような大規模な分類問題も決定リストによって現実的に解決できることも示した．

実験では 1 年分の新聞記事を訓練データとして決定リストを作成し，その決定リストを利用して，別の新聞記事 1,000 文を解析し，その解析結果を形態素解析システム「茶筌」[11] による結果と比較した．本システムで利用した品詞体系は品詞の細分類，活用形まで含めて「茶筌」の品詞体系と同一である．そのため公平に「茶筌」と精度の比較ができる．結果，わずかに精度は「茶筌」より低かったが，ほぼ同等といえる解析精度が得られた．本システムは「茶筌」と品詞体系が同じであるので,「茶筌」と同じレベルの情報が出力される．その上でほぼ同等の解析精度であったことは，本手法が実用的であることを示している．また「茶筌」の解析結果を本手法によって修正するという形をとれば，修正された解析結果の精度は向上した．

## 2. 形態素解析の分類問題への変換

### 2.1 基本的な考え方

日本語形態素解析のタスクは入力文を単語分割し，各々の単語に品詞を付与することである．簡単のために日本語の品詞を名詞，助詞，動詞の3 つとし,「ハナ子が本を読む」を形態素解析することを考える．形態素解析の結果として，図 1 のような単語分割と品詞付けが得られる．

ハナ子 /が /本 /を /読む
〈名詞〉 〈助詞〉〈名詞〉〈助詞〉〈動詞〉

**図 1** 単語分割と品詞付け
Fig. 1 Word segmentation and tagging of part of speeches

図 1 と同等の結果を得るためには，各単語の最初の文字に品詞別の記号をつければ良い．今，名詞の始まる位置に NS という記号，助詞の始まる位置に PS という記号，そして動詞の始まる位置に VS という記号を設けた場合，図 1 の結果を得るためには，各文字に図 2 のような記号を付与できればよい．ただし図 2 において No とは，NS , PS , VS を与えられない文字に与える記号である．

ハナ子 が 本 を 読む
NS No No PS NS PS VS No

**図 2** 形態素解析に対応するクラス付与 (1)
Fig. 2 Giving signs equivalent to morphological analysis (1)

上記の設定から，日本語形態素解析が入力文の各文字にある記号( クラス）を付与する分類問題に変換できることがわかる．

ここではさらにクラスを細分類して，各品詞 $H$ に対して以下の4 つのクラスを用意する．

$H_s$: その文字が品詞 $H$ の開始文字
$H_e$: その文字が品詞 $H$ の終了文字
$H_m$: その文字が品詞 $H$ の中間文字
$H_i$: その文字 1 文字が単語で品詞 $H$

つまり「ハナ子が本を読む」に対しては，各文字に対して 図 3 のようなクラスを割り当てる．

ハ ナ 子 が 本 を 読 む
Ns Nm Ne Pi Ni Pi Vs Ve

**図 3** 形態素解析に対応するクラス付与 (2)
Fig. 3 Giving signs equivalent to morphological analysis (2)

### 2.2 クラスの設定

上記の設定では，形態素解析で用意する品詞が $n$ 種類存在する場合，$4n$ 種類のクラスを用意することに

なる．そして日本語形態素解析は入力文の各文字にその $4n$ 種類の中のいずれかのクラスを付与する分類問題として取り扱うことができる．

基本的にこの設定で十分だが，用言の活用の問題に注意しておく．形態素解析のタスクは単語分割と品詞の付与であるが，用言の場合は活用形も判定する必要がある．本論文では活用形も含めて品詞を設定することにした．たとえば，5 段活用動詞の連用形と連体形では別の品詞と考える．こうすることによって完全に，日本語形態素解析を単一の分類問題へ変換できる．

本論文で設定した品詞（クラス）は形態素解析システム「茶筌」による品詞の細分類と活用形から得た．全部で 638 種類である．このため本論文で設定したクラスはこの数の 4 倍，つまり 2,552 種類のクラスである．

### 2.3 クラスの列の選択

入力文の各文字にクラスを付与するが，一意にクラスを付与した場合，現実には不可能な単語分割が生じることもある．たとえば，ある文字が名詞の始まりのクラスを与えられ，次の文字が動詞が終わるクラスを与えられたとしたら，そのクラス列に対する単語分割は不可能である．

このような事態を避けるために，入力文中の文字 $a$ に対して，$a$ がクラス $C$ に属する確率 $P(C|a)$ を求めることにする．

入力文が $a_1a_2\cdots a_n$（各 $a_i$は文字）の場合，分類器により，$P(C|a_j)$ が求まる．文字 $a_j$ に与えるクラスが $C_j$ とすると，形態素解析は，クラス $C_j$ と $C_{j+1}$ が接続可能という条件のもとで，以下の値が最も大きくなるような，$C_j$ の列を求めることに対応する．

$$\sum_{j=1}^{n} P(C_j|a_j)$$

これは一般に Viterbi アルゴリズムによって求めることができる．

本論文ではクラス間の接続可能性は，先の例に出したような，理論上有り得ないものだけを排除することで設定している．

## 3. 分類規則の学習

### 3.1 決定リストの利用

形態素解析は入力文の各文字 $a$ に対して，$a$ がクラス $C$ に属する確率 $P(C|a)$ を求めることによって実現できる．本論文では $P(C|a)$ を求めるために決定リストを利用する．

決定リストは帰納学習手法の一種であり，正解付きの訓練データから，分類規則を学習する．決定リストの場合，分類規則は証拠とクラスの組の順序付きの表となる．ここで証拠とは属性とその属性の値の組である．実際の分類はリストの上位のものから順に，その証拠があるかどうかを調べ，その証拠があれば，それに対応するクラスを出力する．

決定リストの作成は概ね以下の手順による．

**step 1** 属性を設定する．

たとえば $n$ 個の属性 $att_1, att_2, \cdots, att_n$ とおく．

**step 2** 訓練データから証拠とクラスの組の頻度を調べる．

訓練データ中のあるデータの属性 $att$ の値が $a$ であるとし，そのデータのクラスが $C$ だとする．その場合，$(att, a)$ という証拠と，クラス $C$ の組 $((att, a), C)$ の頻度に 1 を足す．これを訓練データ中の全データに対する全属性について行う．

**step 3** 証拠の判別力と分類クラスを導く．

$((att, a), C)$ の頻度が $f_C$ であった場合，$f_C$ の最大値を与える $\hat{C}$ が証拠 $(att, a)$ に対する分類クラスとなる．またそのときの判別力 $pw((att, a))$ は以下で定義される．

$$pw((att, a)) = \log \frac{f_{\hat{C}}}{\sum_{C \neq \hat{C}} f_C}$$

また $default$ という特別な証拠も設定する．これは訓練データ中で，クラス $C$ の頻度が $sum_C$ であった場合，$sum_C$ の最大値を与える $\hat{C}$ が $default$ に対する分類クラスであり，そのときの判別力は以下で定義される．

$$pw(default) = \log \frac{sum_{\hat{C}}}{\sum_{C \neq \hat{C}} sum_C}$$

**step 4** 判別力の順に並べる．

全ての証拠と分類クラスの組を判別力の大きい順に並べる．これによって作成できた表が決定リストである．ただし証拠 $default$ の判別力よりも小さなものは表から外す．

決定リストが与えるのは判別結果のクラスだけであり，クラスに属する確率は求まらない．ここでは，訓練データから得られる各証拠に対応するクラスの分布からその確率を求める．

決定リストの作成手順の step 3 において，証拠 $(att, a)$ とクラス $C$ の組の頻度 $f_C$ が求まる．この証拠が採用されたときには，$a$ がクラス $C$ に属する確率 $P(C|a)$ を以下の式により与える．

$$P(C|a) = \frac{f_C}{\sum_A f_A}$$

### 3.2 属性の設定

ある文字がどのクラスに属するかを判断する材料が属性である．本論文では基本的に前後の数文字だけを属性とした．文字列 $x_1x_2ay_1y_2$ の中の文字 $a$ の属性として，表 1 の 10 種類を用意する．

**表 1** 設定した属性
Table 1 Attributes

| 属性 | 値 |
|---|---|
| $att_1$ | 文字列 $x_1x_2a$ |
| $att_2$ | 文字列 $x_2ay_1$ |
| $att_3$ | 文字列 $ay_1y_2$ |
| $att_4$ | 文字列 $x_1x_2$ |
| $att_5$ | 文字列 $x_2a$ |
| $att_6$ | 文字列 $ay_1$ |
| $att_7$ | 文字列 $y_1y_2$ |
| $att_8$ | 文字 $x_2$ |
| $att_9$ | 文字 $a$ |
| $att_{10}$ | 文字 $y_1$ |

### 3.3 属性による判別力の重み

決定リストの各属性のクラスを決めるための判別力は同一の基準で決められている．これはクラスを分類するには妥当であろうが，ここで求めたいのはクラスに属する確率なので，各属性から得られた証拠を公平に評価して決定リストの順位をつけるよりも，属性の種類によって段階的に決定リストを適用した方がよいと考えられる．これは属性によって証拠に重みをつけることに対応する．

ここでは表 2 のように重みを設定し，新たな判別力をもとに決定リストの順位を更新した．

**表 2** 属性に対する重み
Table 2 Weight of attributes

| 属性 | 重み |
|---|---|
| $att_1$，$att_2$，$att_3$ | +1000 |
| $att_5$，$att_6$，$att_9$ | +100 |
| $att_4$，$att_7$，$att_8$ ，$att_{10}$ | +0 |

## 4. 実　　験

### 4.1 決定リストの作成

ここでは '94 年度版毎日新聞 1 年間分の記事を訓練コーパスとした．まずこのコーパスを「茶筌」により形態素解析し，その結果をもとに，各文字にここで設定した 2,552 種類のクラスを与えた．

次に 3.1 で述べた step 2 以降の手順に従って，決定リストを作成した．作成できた決定リストの一部を表 3 に示す．表 3 の文字 "□" は文末の1 つ後ろに仮想的に作った文字である．また文字 "■" は文末の2 つ後ろに仮想的に作った文字である．また品詞の最後に付けている, s，m，e，i の文字は，それぞれその品詞の始まり，中間，終わり，その文字自身がその品詞となっていることを示す．またクラス分布とは，その証拠が選ばれたときに，各クラスが選択される確率の分布である．確率はクラスの後に括弧内で示した．また記載のないクラスに関しては確率が 0 だと考える．表 3 から分かるように作成できた決定リストの大きさは 1,951,965 であった．

### 4.2 実　行　例

簡単な実行例を示す．今「古代中国の哲学者」という文の 3 番目の文字「中」にクラスを付与してみる．この文字に対する証拠は以下の 10 種類である．

($att_1$, 古代中)，($att_2$, 代中国)，($att_3$, 中国の)，
($att_4$, 古代)，($att_5$, 代中)，($att_6$, 中国)，
($att_7$, 国の)，($att_8$, 代)，($att_9$, 中)，($att_{10}$, 国)

決定リストの中でこれらの証拠の各々の順位を調べる．その中で最も順位の高い証拠は順位 174,724 の ($att_3$, 中国の) であった．この証拠に対するクラスの頻度分布から，「中」の文字に与えるクラスとその確率は以下の通りとなる．

| | |
|---|---|
| 名詞-固有名詞-地域-国-s | 0.9977 |
| 名詞-一般-s | 0.0014 |
| 名詞-固有名詞-地域-一般-s | 0.0004 |
| 名詞-固有名詞-地域-一般-e | 0.0004 |

### 4.3 「茶筌」との解析結果比較

'95 年度の毎日新聞から適当に選んだ 1,000 文に対して本手法及び「茶筌」を用いて形態素解析を行った．そしてそれぞれの結果の異なり部分を比較することで評価を行った．

まず単語分割であるが，以下のように集計した．ある文字列に対する単語分割が等しいとき，同一の判定に 1 を加える．また単語分割が等しくないとき，その部分を含む最小の長さの文字列で，しかも前後の単語分割の位置が等しい文字列を取りだす．その文字列に対して，異なりの判定に 1 を加える．たとえば，図 4 では，同一個所が 2 個で異なり個所が 2 個である．

結果は，23,550 個の単語分割が同一であり，307 個の単語分割に違いがあった．つまり全体の 98.7% が同じ単語分割である．違いの部分の内訳は以下の通りである．

(1)　本手法の単語分割の方が正しい( 59 個)．

**表 3** 構築した決定リスト
Table 3 Derived decision list

| 順位 | 証拠 | 判別力 | クラス分布 |
|---|---|---|---|
| 1 | ($att_3$,。□■) | 1023.170 | 記号-句点-i (1.0) |
| 2 | ($att_2$, た。□) | 1021.336 | 記号-句点-i (1.0) |
| ... | ... | ... | ... |
| 143 | ($att_3$, 0 0 円) | 1015.644 | 名詞-数-m (1.0) |
| ... | ... | ... | ... |
| 1,365,864 | ($att_9$,「 ) | 121.347 | 記号-括弧開-i (1.0) |
| 1,365,865 | ($att_5$, た。) | 121.336 | 記号-句点-i (1.0) |
| ... | ... | ... | ... |
| 1,658,816 | ($att_4$, 日午) | 17.261 | 名詞-副詞可能-e (1.0) |
| 1,658,817 | ($att_7$, 疑者) | 15.876 | 名詞-一般-s (1.0) |
| ... | ... | ... | ... |
| 1,951,964 | ($att_7$,、越) | -3.531 | 動詞-自立-サ変・スル-連用形-i (0.0786)<br>名詞-副詞可能-e (0.0786)<br>...<br>名詞-固有名詞-人名-姓-e (0.0112) |
| 1,951,965 | default | -3.540 | 名詞-一般-s (0.0791)<br>名詞-一般-e (0.0791)<br>...<br>動詞-自立-五段・ガ行-連用タ接続-e (0.0001) |

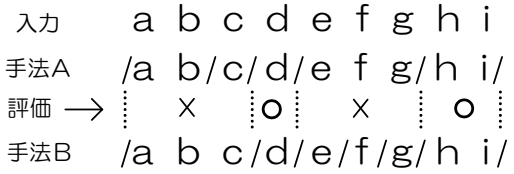



**図 4** 単語分割の評価方法
Fig. 4 Evaluation of word segmentation

これは多くの場合，未知語を本手法では正しく認識でき，「茶筌」ではできなかった結果である（27 個）．その他のものとしては様々である．たとえば，以下のような例がある☆．

| 本手法 | 優しく | 形容詞-自立-形容詞・イ段-連用テ接続 |
|---|---|---|
| 「茶筌」 | 優 | 名詞-一般 |
| | しく | 動詞-自立-五段・カ行イ音便-基本形 |

| 本手法 | 二十八 | 名詞-数 |
|---|---|---|
| | 日 | 名詞-接尾-助数詞 |
| 「茶筌」 | 二十八日 | 名詞-固有名詞-地域-一般 |

(2) 「茶筌」の単語分割の方が正しい（131 個）．
これは多くの場合，本手法が1 つの単語を過剰分割することが原因である．「茶筌」では辞書を利用しているために，低頻度の長い文字列（カタカナの地名や平仮名表記の単語など）からなる単語も1 単語として認識できるが，本手法ではそのような単語は過分割されることが多い．たとえば以下のような例がある．

| 本手法 | とび | 動詞-自立-五段・バ行-連用形 |
|---|---|---|
| | だ | 助動詞-特殊・タ-基本形 |
| | し | 動詞-自立-サ変・スル-連用形 |
| 「茶筌」 | とびだし | 動詞-自立-五段・サ行-連用形 |

| 本手法 | ソルト | 名詞-サ変接続 |
|---|---|---|
| | レークシティー | 名詞-一般 |
| 「茶筌」 | ソルトレークシティー | 名詞-固有名詞-地域-一般 |

(3) どちらも誤り（17 個）．
これはあまり使われない表現や単語の解析部分で生じている．たとえば，「防空ごう」や「もたれあい」といった単語の解析で本手法も「茶筌」も過分割している．

(4) どちらも正しい（100 個）．
単語分割は決定的にどちらかが正しいという判定を下せない場合も多い．たとえば，「情報処理」は「茶筌」では一単語として解析されているが，本手法では /情報/処理/ と分割されて

☆ ここに挙げた2 つの例は，現在の「茶筌 (ver 2.20)」では正しく解析できる．この実験を行った時期の「茶筌」のバージョンは 2.0b6 である．

いる．どちらが正しいかは複合語あるいは単語の定義に依存する．このようなタイプの違いはどちらも正しいと判断した．

ただし必ずしも本手法の方が分割数が多いというわけではないことを注記しておく．たとえば，「運輸省」は本手法では一単語だが，「茶筌」では/運輸/省/と分割されている☆．

単語分割に対しては，わずかだが，「茶筌」の方が精度がよい．ただしその差は非常に小さく，ほぼ同程度の精度を出せると見なせる．

次に品詞の付与の評価であるが，これは先の単語分割が同一のもの 23,550 個について付与した品詞が異なる部分について評価した．同一の品詞が付与された単語は，23,201 個 (98.5%) であった．本論文で使った品詞の種類が細かいことも考慮に入れると，品詞付けに対しても本手法はほぼ「茶筌」と同等の能力があることがわかる．

349 個が異なる品詞を付与された．異なる部分の評価の内訳は以下の通りである．

(1) 本手法の品詞付けの方が正しい（ 107 個）
(2) 「茶筌」の品詞付けの方が正しい（ 193 個）
(3) どちらも誤り（ 36 個）
(4) どちらも正しい（ 13 個）

ここでも「茶筌」の品詞付けの方がわずかに精度が良い．ただし，ここでは品詞が非常に細かく分類されているので，このような差が生じている．たとえば，「プレート」を本手法では「名詞-一般」に品詞付けているが，「茶筌」では「名詞-サ変接続」に品詞付けている．この場合は本手法の方が正しいが，このような差はあまり意味はない．また，「で」を助動詞「だ」の連用形ととるか，助詞ととるか，動詞の自立語か非自立語か，「と」は「助詞-格助詞-引用」か「助詞-並立助詞」などが目だった違いであり，品詞の大分類まで異なるような場合はほとんどなかった．

### 4.4 「茶筌」の解析結果の修正

単語分割で本手法が誤っているケースを観察すると，単語を過分割しているケースが目立つ．そこで本手法を「茶筌」の解析結果を修正するシステムとして位置づけてみる．この場合，「茶筌」の解析結果と本手法による解析結果を比較して，同一の単語分割であればそれを出力し，異なる単語分割の場合に，基本的には本手法の結果を出力する．ただし「茶筌」の解析結果を過分割した結果が本手法の結果となっている場合は「茶筌」の結果をそのまま採用する．つまり本手法により「茶筌」の解析結果を修正するという形で形態素解析を行ってみる．

上記の 1,000 文に対して，実験してみると，修正の個所は 141 個になり，内訳は以下の通りとなる．

(1) 修正は有効（ 53 個）
(2) 修正は悪影響（ 33 個）
(3) 修正前も後も誤り（ 10 個）
(4) 修正前も後も正解（ 45 個）

この結果から，「茶筌」の解析結果の修正のために本手法を利用した場合，「茶筌」の解析精度が上がることがわかる．

## 5. 考　　察

本手法による形態素解析の精度は「茶筌」と比較した場合，わずかに低かったが，ほぼ同程度と言える．一方，従来手法の精度をみると，文字ベースの HMM を利用した手法 [3] では，EDR コーパスの open テストで 95.9%，ATR 対話コーパスで 96.9% である．決定木を利用した手法 [5] では，ATR 対話コーパスの単語分割において 91.8%，品詞付けにおいて 91.6% である．k-近傍法を利用した手法 [5] では，EDR コーパスの open テストにおいて単語分割で 95.18%，品詞付けで 93.22% となっている．「茶筌」の精度は経験的に 95 〜 98% と思われるので，本手法は従来手法と比べても同程度以上には精度があると思われる．ただしこれらの精度は，当然，同列には比較できない．ここでは目安のために示した．

形態素解析システムの評価としては，テスト文に手作業で正解を付与し，その正解と解析結果との比較から評価する方法が一般的である．前述した従来の研究もそのような評価を行っている．しかしこの評価方法では，他手法，他システムとの比較が難しい．それぞれ利用している辞書や品詞体系が異なるからである．一方，「茶筌」は日本語形態素解析の実質的な標準システムと言える．その「茶筌」と同等の精度があったことは現実的に有効な手法であると言える．また「茶筌」の解析結果の修正として本手法を利用すれば，最終的に得られる結果は修正前のものよりも精度は高かったため，本手法の有用性はある．

また本手法は学習データとして「茶筌」の解析結果をそのまま利用している．この枠組みでは，「茶筌」の精度を越えることは原理的に困難であり，ここでの実験の結果もそれを示唆している．ただし，人間からの教示を必要としない点，つまり生のコーパスからの学習であることを考慮すれば，本実験の結果は十分な精

☆ 逆に「通産省」は「茶筌」では一単語だが，本手法では/通産/省/と分割されている

度を出したと言える．本論文の1 つの目的は形態素解析が分類問題として取り扱えることを示すことである．分類問題として取り扱うことができれば，今後の精度向上は期待できる．例えば，完全な正解を付与した学習データを用意することでも，「茶筌」の精度を越えられると考える．

本システムの処理時間に関して述べておく．本システムは perl で実装しているため，絶対的な処理時間そのもので「茶筌」には及ばない．たとえ C で記述したとしても，おそらく「茶筌」の方が速いであろう．ただ現実に使えないほど，処理時間がかかるわけではないことを注意したい．本手法は決定リストの検索に時間がかかるが，通常の形態素解析で必要とされる辞書引きを省くことができる．本システムで用いた決定リストのサイズ（200 万見出し程度）と辞書サイズ（数十万見出し程度）では，絶対的な検索時間が極端に変わるほどの差はでない考えられる．

本手法は分類問題の解法として決定リストを利用したが，決定リストは分類問題の手法としては，原始的である．決定木，あるいは最大エントロピー法などを利用することも可能である．ただしここで設定したクラスは数が多い．また属性のとり得る値も文字や文字列になるために，その種類も多い．そのために単純に決定木や ME 法が利用できるとは考えられない．何らかの工夫が必要であろう．たとえば，文字をグループ化することは，属性のとり得る値を減らすことに有効であろう [12]．

本手法は辞書や品詞間の接続表を解析時には利用していない．品詞列を求める際に辞書や品詞間の接続表を用いることで，更なる精度の向上も期待できる．また品詞間の接続のしやすさなども有効であろう．

本手法の長所として，未知語への対応がある．実験で利用した 1,000 文に対して，手作業によって未知語を確認したところ，カタカナ表記とアルファベット表記以外の未知語は 56 種類，78 個存在した☆1．そのうち「茶筌」で解析できたものは，当然，なかったが，本手法では 16 種類，27 個解析できていた．解析できたものは，漢字から構成される未知語がほとんどであった（13 種類，23 個）．それ以外のタイプの未知語はほとんど解析できていない．しかし平仮名だけからなる未知語の多くは，「精緻（せいち）」や「横尾忠則氏（よこお・ただのり）」のように漢字のよみを括弧内に付記したものが多い（24 種類中 14 種類）．この種の未知語の解析は括弧の存在を利用すべきであり，そのような情報も本手法では取り込めるであろう．

また過去，決定木 [5] を使った日本語形態素解析の研究はある．これは学習された分類器により，形態素解析を行うという点で本手法と類似している．ただしこの研究は単語分割を，文字列にクラスを付与する分類問題として扱っており，本手法のアプローチとは異なる☆2．さらに，本手法は文字ベースの HMM とも似ているが，その違いを述べておく．通常，HMM の状態遷移確率やシンボル出力確率は n-gram から学習させるが，本手法ではそれを様々な属性から学習させるのが大きな違いである．たとえば，山本らは日本語形態素解析を行うために，入力文の各文字に単語区切りの情報と品詞の情報を持たせたタグを付与するという手法を提案した [3]．これはクラスの設定方法が異なるが☆3，基本的に本手法と同様の処理を行っている．しかしそこでは形態素解析を HMM の枠組みで考えているために，状態遷移確率やシンボル出力確率を 3-gram から得ている．またこの研究の拡張として，n-gram を可変長の n-gram にした研究も行われている [2] ☆4．本手法は，その可変長よりもさらに自由な属性が利用できる枠組みを持つ．例えば，字種の情報やその文字が括弧内の文字かどうかなどを n-gram で利用するのは難しいが，分類問題の枠組みでは容易である☆5．あるいは，n-gram とは全く別のメタな情報（たとえば文書の種類，その文がタイトルかどうかなど）さえも利用できる．

また本手法が固有表現抽出の手法から応用されたことを注記しておく．たとえば固有表現の人名を抽出するのは，その人名の始まる単語列に人名の始まり，人名の中間，人名の終り，その単語自身が人名という 4 つのクラスを付与すればよい．設定した固有表現の種類ごとにこのクラスを用意してゆけば，固有表現抽出を分類問題として取り扱うことができる [14]．これを文字列に適用した研究 [15] を形態素解析に応用したものが本研究である．固有表現抽出のポイントの1 つは，当然ではあるが，未知の固有表現への対処である．固

☆1 ここでの未知語とは人名のように明らかに一単語として扱うべきだが，「茶筌」の辞書に未登録であるものである．

☆2 本論文の投稿中，最大エントロピー法を用いた形態素解析システムが提案された [13]．これも同様に，文字列にクラスに属する確率を与えるアプローチであり，本手法の文字にクラスを与えるというアプローチとは異なる．

☆3 [3] では，クラスの数は (品詞数)*2 になる．本論文のように更に細かく設定すべきかどうかの確認は今後の課題である．

☆4 ただし，[2] では品詞付けを行っていない．

☆5 前述したように平仮名からなる未知語の多くは，漢字のよみを括弧内に付記したものであるために，連接する平仮名文字どうし括弧内にあれば，繋がりやすいという規則が学習できると思われる．

有表現抽出の手法から本手法を派生させることにより，日本語形態素解析の課題である未知語処理への解決も意図している．

## 6. おわりに

本論文では日本語形態素解析を分類問題とみなして解くことを提案した．具体的には入力文の各文字に2,552種類のクラスのそれぞれのクラスに属する確率を付与する．そこからViterbiアルゴリズムによって最適なクラス列を求めることで形態素解析を行う．またここでは，分類問題の解法として決定リストを利用した．1年分の新聞記事から学習させた決定リストによる形態素解析は「茶筌」とほぼ同等の精度があることを示した．これは本手法の訓練データが「茶筌」の解析結果をそのまま用いていることも考えれば，十分に高い精度である．また本手法を「茶筌」による解析結果の修正に利用すれば，最終的に得られる解析結果の精度は改善された．このため本手法の有用性はある．本手法は分類問題に対する様々な学習戦略を適用できること，及び，未知語に対応した手法となっていることが長所である．

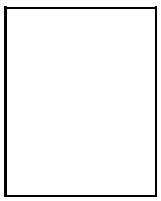
学習データを修正することで「茶筌」の精度を越えること，平仮名列の未知語を認識できる属性を増やし未知語認識の精度をあげること，および小規模の学習データから高精度の分類規則を学習できる手法を考案することなどを今後の課題とする．

**謝辞** 本研究は（財）栢森情報科学振興財団の研究助成金（K11研IV第71号）によって行われました．深く感謝します．

## 参考文献

1) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-$A^*$ N-Best Search Algorithm, *Proceedings of 15th International Conference on Computational Linguistics (COLING-94)*, pp. 201–207 (1994).
2) 小田裕樹, 北研二: $PPM^*$ モデルによる日本語単語分割, 技術報告 NL-128-2, 情報処理学会自然言語処理研究会 (1998).
3) 山本幹雄, 増山正和: 品詞・区切り情報を含む拡張文字の連鎖確率を用いた日本語形態素解析, 言語処理学会第3回年次大会, pp. 421–424 (1997).
4) Nagamatsu, K. and Tanaka, H.: A Stochastic Morphological Analysis for Japanese employing Character n-Gram and k-NN Method, *Proceedings of Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97)*, pp. 23–28 (1997).
5) Kashioka, H., Eubank, S. G. and Black, E. W.: Decision-Tree Morphological Analysis without a Dictionary for Japanese, *Proceedings of Natural Language Processing Pacific Rim Symposium 1997 (NLPRS'97)*, pp. 541–544 (1997).
6) Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer Verlag (1995).
7) Freund, Y., Robert Schapire (訳: 安倍直樹): ブースティング入門, 人工知能学会誌, Vol. 14, No. 5, pp. 771–780 (1999).
8) Quinlan, J. R.: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publisher (1993).
9) Ratnaparkhi, A.: Maximum Entropy Models for Natural Language Ambiguity Resolution, *PhD thesis*, University of Pennsylvania (1998).
10) Yarowsky, D.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, *32th Annual Meeting of the Association for Computational Linguistics*, pp. 88–95 (1994).
11) 松本裕治, 北内啓, 山下達雄, 平野喜隆: 日本語形態素解析システム『茶筌』version 2.0使用説明書, `http://cl.aist-nara.ac.jp/lab/nlt/chasen.html` (1999).
12) 小田裕樹, 森信介, 北研二: 文字クラスモデルによる日本語単語分割, 自然言語処理, Vol. 6, No. 7, pp. 93–108 (1999).
13) 内元清貴, 関根聡, 井佐原均: 最大エントロピーモデルに基づく形態素解析-未知語問題の解決策-, 自然言語処理, Vol. 8, No. 1, pp. 127–141 (2000).
14) Sekine, S., Grishman, R. and Shinnou, H.: A Decision Tree Method for Finding and Classifying Names in Japanese Texts, *Proceedings of the 6th Workshop on Very Large Corpora* (1998).
15) 新納浩幸: 拡張文字ベースのHMMを利用した固有名詞抽出, IREXワークショップ予稿集, pp. 151–157 (1999).

(平成12年7月21日受付)
(平成13年6月19日採録)



**新納 浩幸**（正会員）

昭和36年生．昭和60年東京工業大学理学部情報科学科卒業．昭和62年同大学大学院理工学研究科情報科学専攻修士課程修了．同年富士ゼロックス，翌年松下電器を経て，平成5年4月茨城大学工学部システム工学科助手にて着任．平成9年10月同学科講師，平成13年4月同学科助教授，現在に至る．博士（工学）．