誤り駆動型の素性選択による日本語形態素解析の確率モデル学習

北内 啓[†] 宇津呂 武仁[††] 松本 裕治[††]

自然言語処理において形態素解析は基本的かつ重要な技術であり，また応用範囲も広く，現在まで様々な形態素解析システムが開発されてきた．一方，近年大量の品詞タグ付きコーパスが利用可能になってきており，これを用いて形態素解析のパラメータを統計的に学習する方法がさかんに行われるようになってきた．しかし，パラメータ値を自動的に推定できるようになっても，その土台となる文法そのものは固定されているものが多い．たとえば，どういった品詞分類のもとでパラメータ推定を行えば高い精度が得られるのかということは，人手によって決めていることが多い．そこで本論文では，確率モデル学習の手法により，日本語形態素解析の精度を向上させるのに有効な品詞分類を自動的に学習する方法を提案する．その方法においては，解析誤りをもとに詳細化する品詞分類を素性として取り出し，品詞分類を段階的に細かくしていく．学習によって得られた品詞分類を用いて bi-gram のマルコフモデルに基づくパラメータ推定を行うことにより，形態素解析の精度を向上させた．実験により，人手で設定した品詞分類に比べ，より少ないパラメータ数でより高い精度を得ることができた．また，品詞分類を変化させることによって，パラメータ数や精度がどのように変化するかといった，品詞分類全体の性質をとらえることができた．

Probabilistic Model Learning for Japanese Morphological Analysis by Error-driven Feature Selection

Akira Kitauchi,[†] Takehito Utsuro[††] and Yuji Matsumoto[††]

Morphological analysis is the initial step of natural language processing, and thus is a fundamental and important technique. It is also an inevitable step in many application-oriented NLP systems, and various sorts of morphological analysis systems have been developed and implemented. Recently, very large part-of-speech tagged corpora have become available and corpus-based statistical techniques for improving morphological analysis have been intensively studied. However, most of those works mainly study parameter estimation of morphological analyzer with a fixed set of part-of-speech tags, which is predetermined by human intuition. This paper proposes a method of learning an optimal set of part-of-speech tags which gives the highest performance in Japanese morphological analysis. In our method, considering patterns of errors in the morphological analysis, candidates of more specific part-of-speech tags to be included in the model of morphological analyzer are generated. Then, the most effective candidate which gives the greatest decrease in errors is selected. In the experimental evaluation of the proposed method, we achieve a morphological analyzer of higher performance compared with a model with a hand-tuned set of part-of-speech tags, and with much smaller number of parameters.

1. はじめに

形態素解析は自然言語処理の中で最も基本的な技術であり，応用範囲も広い．従来から多くの形態素解析システムが開発され，実際に使用されてきた．しかし，その多くは文法や辞書の整備を人手で行うもので，メンテナンスに手間がかかる．特に，単語のコスト値や品詞どうしの連接のしやすさなどのパラメータを人手で調節するのは非常に困難である．

そのような手間を軽減するため，品詞タグ付きコーパスを用いて形態素解析のための言語的特徴を統計的に学習し，解析精度を向上させるという手法が確立されてきた[1),2)]．しかし，これらの手法の多くは，文法体系など，ある決められた枠組みの中でのパラメータ値を自動的に求めるというもので，品詞分類など文法体系そのものを動的に決定するという手法はあまり行われていない．たとえば，日本語の品詞分類は活用を持つため，活用型や活用形のすべての組合せを考えると品詞の種類は数百に及ぶ．さらに，助詞などは語彙

† 株式会社 NTT データ技術開発本部オープンシステムセンタ
Open System Center, NTT Data Corporation
†† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology

レベルでもコーパス中の分布が異なり，膨大な種類の品詞分類を考慮する必要がある．このように種類の多い品詞について高い精度が得られるような品詞分類を人手で決定するのは，非常に手間がかかる．

そこで本論文では，品詞タグ付きコーパスを用いた日本語形態素解析の統計的学習において，形態素解析の精度を向上させるのに有効な品詞分類を自動的に求める手法を提案する．本論文の手法では，1つ1つの品詞分類を素性と見なし，解析誤りをもとに素性集合を求めていくことで品詞分類を決定する．具体的には，まず初期状態としてかなり粗い品詞分類から学習を開始し，解析誤りの多い品詞や単語に注目して素性を抽出する．抽出した素性を一時的に追加した素性集合を用いてパラメータ推定を行い，解析精度が十分上がる場合はその素性を正式に素性集合に追加する．このように素性の抽出と追加の手順を繰り返すことにより，段階的に品詞分類を細かくしていき，最終的に精度がどのくらい向上するかを測定する．

本論文のパラメータ推定の方法は，基本的には bi-gram のマルコフモデルに基づいているため，前件と後件の品詞分類を掛け合わせてできる品詞分類での連接確率は求めることができる．しかし，詳細な分類の品詞どうしの連接がコーパス中に特徴的に出現する場合など，詳細な連接部分の連接確率とその連接以外の部分の連接確率を別々に求めたいことがある．本論文ではこの部分の連接確率も求められるようにすることで，解析精度を向上させた．

実験では，コーパスとして毎日新聞1995年度の記事15,000文の品詞タグ付きコーパスを使用し，訓練用として3種類のコーパス，評価用として2種類のコーパス，合計5種類のコーパス3,000文ずつを用いて精度とパラメータ数を測定した．形態素解析には日本語形態素解析システム「茶筌」[3]を使用し，IPAの文法体系[4]のもとで解析を行った．実験の結果，形態素解析の精度を向上させるような品詞レベルを自動的に求め，少ないパラメータ数で高い精度を得ることができた．

2. 誤り駆動型の確率モデル学習による日本語形態素解析

2.1 日本語形態素解析の確率モデル

本論文の確率モデルは，基本的にはN-gramのマルコフモデルに基づいている．N-gramのマルコフモデルにおける日本語形態素解析は

> 与えられた入力文 S に対する単語列 $W = w_1 \cdots w_n$ と品詞列 $T = t_1 \cdots t_n$ の同時確率 $P(W,T|S)$ を最大にするような，単語列と品詞列の組 $(\hat{W},\hat{T})$ を求める．

という問題に帰着され，$P(W,T|S)$ は次式によって与えられる．

$$P(W,T) = \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{1,i-1}) \qquad (1)$$

$P(w_i|t_i)$ が単語生成確率，$P(t_i|t_{1,i-1})$ が品詞連接確率である．本論文ではbi-gramのマルコフモデルに基づいており，品詞連接確率 $P(t_i|t_{1,i-1})$ は次のように近似される．

$$P(t_i|t_{1,i-1}) \approx P(t_i|t_{i-1}) \qquad (2)$$

式(2)では前件 t_{i-1} と後件 t_i の品詞分類は等しい．しかし，たとえば「歩く」のような活用語の場合，後件との連接には「歩か–ない」「歩い–た」のように活用形を区別する必要があるが，前件との連接には活用形の区別はあまり関係ない．また「歩か–ない」「食べ–ない」「読ま–ない」のように，活用型は区別せずに活用形のみ区別すればよい場合もある．そこで本論文では，前件の品詞分類を後件の品詞分類と関係なく自由に決定できるようにするために，前件の品詞列を $T' = t'_1 \cdots t'_n$ とした次式の確率モデルを用いる．

$$P(W,T',T) = \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t'_{i-1}) \qquad (3)$$

この確率モデルでは前件と後件の品詞分類が異なり，マルコフモデルの条件を満たしていない．しかし，この論文の目的は，前件と後件の品詞分類を柔軟に調整し，誤りが最も少なくなるようなモデルを見つけることであり，あえてマルコフモデルの条件を満たさない式(3)の確率モデルを採用した．

本論文では，パラメータ推定を行う際の品詞分類に基づく品詞の連接を素性と呼ぶ．すなわち，前件と後件の品詞分類を素性の集合ととらえることができる．品詞分類を決定することは，式(3)において前件の品詞 t'_{i-1} や後件の品詞 t_i の分類を決定することに相当する．解析誤りをもとに最適な素性集合を求め，その素性集合が表す品詞分類を用いて単語生成確率と品詞連接確率のパラメータ推定を行う．

2.2 パラメータ推定

パラメータ推定は，次式の最尤推定法によって単語生成確率と品詞連接確率を求める．式(3)の確率モデルにおいても，マルコフモデルの場合とまったく同じ要領でパラメータの最尤推定を行うことができる．C は出現回数を表し，たとえば $C(t_i)$ は品詞 t_i の出現回数，$C(w_i \wedge t_i)$ は品詞が t_i である単語 w_i の出現

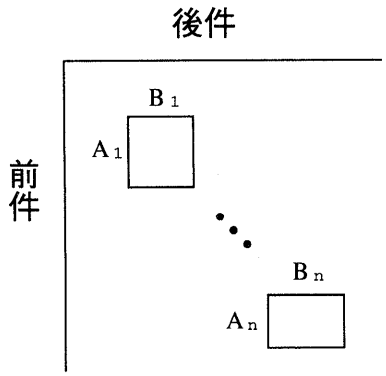


図1 連接パラメータ推定における n 個の領域の併合
Fig. 1 Merging n regions in bi-gram parameter estimation.

回数である．

$$P(w_i|t_i) = \frac{C(w_i \wedge t_i)}{C(t_i)} \tag{4}$$

$$P(t_i|t'_{i-1}) = \frac{C(t'_{i-1}, t_i)}{C(t'_{i-1})} \tag{5}$$

次に，詳細な分類の品詞どうしの連接に対するパラメータ推定について述べる．まず，一般的な場合として，**図1**のように互いに共通部分を持たない n 個の連接 $(A_1, B_1), \cdots, (A_n, B_n)$ に対し，これらを同一視し，

$$X = \bigcup_{i=1}^{n} (A_i \times B_i)$$

を1つの領域と考えて，連接確率を求める推定方法を述べる．

本論文の確率モデルの式(3)のある i において，t'_{i-1} を A_i, t_i を B_i とおくと，式(3)の一般項の式は次のようになる．

$$P(w_i|B_i)P(B_i|A_i) = \frac{P(w_i \wedge B_i)P(A_i, B_i)}{P(A_i)P(B_i)} \tag{6}$$

ここで，n 個の連接 $(A_1, B_1), \cdots, (A_n, B_n)$ を同一視することを，確率モデルの式(3)における役割に限定して同一視することに置き換えて考える．すなわち，一般項の式(6)のうち，単語に依存する部分をのぞいた $P(A_i, B_i)/P(A_i)P(B_i)$ の値が，同一視する n 個の領域の間で等しいという次式の制約条件を導入する．

$$\frac{P(A_1, B_1)}{P(A_1)P(B_1)} = \cdots = \frac{P(A_n, B_n)}{P(A_n)P(B_n)} = k \tag{7}$$

連接の総出現数を C_c，単語の総出現数を C_w とおく．各領域の同時確率 $P(A_i, B_i)$ の総和を考えると，次式のようになる．

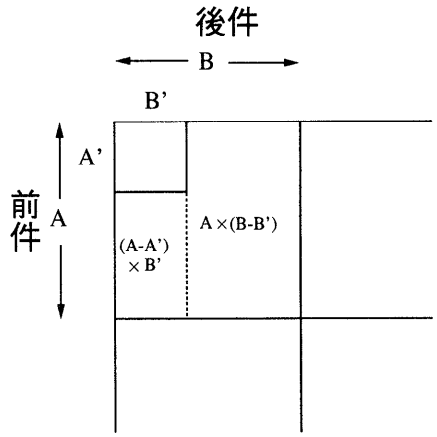


図2 連接パラメータ推定において例外的な連接の連接確率を別に求める場合
Fig. 2 Estimating an exceptional parameter in bi-gram parameter estimation.

$$\sum_{i=1}^{n} P(A_i, B_i) = \sum_{i=1}^{n} \frac{C(A_i, B_i)}{C_c} = \frac{1}{C_c} \sum_{i=1}^{n} C(A_i, B_i)$$

一方，式(7)より

$$\sum_{i=1}^{n} P(A_i, B_i) = k \sum_{i=1}^{n} P(A_i)P(B_i) = \frac{k}{C_w^2} \sum_{i=1}^{n} C(A_i)C(B_i)$$

一般に $C_c = C_w$ であるから，

$$k = \frac{C_w \sum_{i=1}^{n} C(A_i, B_i)}{\sum_{i=1}^{n} C(A_i)C(B_i)} \tag{8}$$

したがって，$P(B_i|A_i)$ の推定式は次のようになる．

$$P(B_i|A_i) = kP(B_i) = \frac{C(B_i) \sum_{j=1}^{n} C(A_j, B_j)}{\sum_{j=1}^{n} C(A_j)C(B_j)} \tag{9}$$

次に，**図2**のように粗い品詞分類の連接 (A, B) に対し，細かい品詞分類の連接 (A', B') だけを特別視して，(A', B') の部分の連接確率と，(A, B) から (A', B') を取りのぞいた部分 X の連接確率を別々に求める．

$$X = ((A - A') \times B') \cup (A \times (B - B'))$$

であるから，$(A-A')\times B'$ と $A\times(B-B')$ の2つの領域を同一視すると，式(8)より k の値は次式のようになる．

$$k=\frac{C_w(C(A-A',B')+C(A,B-B'))}{C(A-A')C(B')+C(A)C(B-B')}$$
$$=\frac{C_wC(X)}{C(A)C(B)-C(A')C(B')}$$

よって，$(A-A')\times B'$ と $A\times(B-B')$ の連接確率はそれぞれ以下のように求められる．

$$P(B'|A-A')=\frac{C(B')C(X)}{C(A)C(B)-C(A')C(B')}$$
$$P(B-B'|A)=\frac{C(B-B')C(X)}{C(A)C(B)-C(A')C(B')}$$

2.3 誤り駆動型の素性選択

本論文では，品詞分類を粗くした状態を初期状態とし，解析誤りをもとに詳細な品詞分類を素性として抽出，追加していくことにより品詞分類を細かくしていき，最適な素性集合を学習した．ここではその学習方法について説明する．この方法以外にも，細かい品詞分類を初期状態として，品詞分類を段階的に粗くしていくことにより素性集合を学習していく方法などが考えられる．

2.3.1 概　　要

学習には，3種類の訓練コーパスと2種類の評価コーパスを用いる．

訓練コーパス　素性を選択するときに使用する．

- パラメータ推定用コーパスA
  素性選択時，パラメータ推定を行うために用いる．
- 素性選択用コーパスB
  素性選択時，解析誤りから素性を抽出するために用いる．
- 素性評価用コーパスC
  素性選択時，素性集合に素性を追加して精度が向上するかどうかを評価するために用いる．

評価コーパス　訓練で得られた素性集合の解析精度を評価するときに使用する．

- パラメータ推定用コーパスD
  評価時，パラメータ推定を行うために用いる．
- 解析精度評価用コーパスE
  評価時，解析を行って精度を測定するために用いる．

学習は以下の手順で行う．

(1) 初期化
まず，初期の品詞分類のもとでコーパスAを用いてパラメータ推定を行い，コーパスBを解析する．解析時のパラメータ値には，コーパスAで学習した単語生成確率と品詞連接確率のほかに，コーパスA，B，C，D，Eに含まれるすべての単語にごく小さな単語生成確率を与えたものを使う☆．

(2) 素性候補の選択と追加
以下の手順を，適当な回数だけ，あるいは抽出する素性が1つもなくなるまで繰り返す．

(a) 素性候補の選択
コーパスBを解析した結果とコーパスBに付与されている品詞とを比較して得られた解析誤りをもとに，追加する素性集合の候補を抽出する．

(b) 素性の追加
素性候補の中から適当に取り出した素性集合を一時的に現在の素性集合に追加し，コーパスAでパラメータ推定を行い，コーパスCを解析する．解析時のパラメータ値には，コーパスAでパラメータ推定を行って求めた単語生成確率と品詞連接確率のほかに，コーパスA，B，C，D，Eに含まれるすべての単語にごく小さな単語生成確率を与えたものを使う．
解析の結果，十分に精度が上がっていれば一時的に追加した素性集合を正式に採用して現在の素性集合に追加する．

(3) 評価
学習の結果得られた素性集合を用いて，コーパスDでパラメータ推定を行い，コーパスEを解析し，最終的にどのくらい精度が向上したか測定する．
学習時と同様，解析時のパラメータ値には，コーパスDで学習した単語生成確率，品詞連接確率のほかに，コーパスA，B，C，D，Eに含まれるすべての単語にごく小さな単語生成確率を与えたものを使う．

2.3.2 初期の素性集合

学習を始めるにあたり，まず初期の素性集合，すなわち前件と後件の品詞分類を決めておく必要がある．素性集合は品詞細分類，活用型，活用形，語彙の4個の要素を組み合わせることで決定される．これらの4個の要素を組み合わせることで様々な素性集合を作り

☆ この論文では未知語の問題は取り扱わず，未知語が存在しないという理想的な状況で実験を行う．

解析文	コーパス				解析結果
さっぱり	副詞-助詞類接続	*	*	さっぱり	←(同左)
と	助詞-副詞化	*	*	と	助詞-格助詞-引用 * * と
おいしい	形容詞-自立	形容詞・イ段	基本形	おいしい	←(同左)
炒め	動詞-自立	一段	連用形	炒める	名詞-一般 * * 炒め物
物	名詞-接尾-一般	*	*	物	
です	助動詞	特殊・デス	基本形	です	←(同左)
.	記号-句点	*	*	.	←(同左)

図3 形態素解析結果の例
Fig. 3 Result of morphological analysis.

前件				後件			
副詞-助詞類接続	*	*	さっぱり	助詞-格助詞-引用	*	*	と
副詞-助詞類接続	*	*	さっぱり	助詞-副詞化	*	*	と
助詞-格助詞-引用	*	*	と	形容詞-自立	形容詞・イ段	基本形	おいしい
助詞-副詞化	*	*	と	形容詞-自立	形容詞・イ段	基本形	おいしい
形容詞-自立	形容詞・イ段	基本形	おいしい	動詞-自立	一段	連用形	炒め
形容詞-自立	形容詞・イ段	基本形	おいしい	名詞-一般	*	*	炒め物
動詞-自立	一段	連用形	炒め	名詞-接尾-一般	*	*	物
名詞-接尾-一般	*	*	物	助動詞	特殊・デス	基本形	です
名詞-一般	*	*	炒め物	助動詞	特殊・デス	基本形	です

図4 解析誤りから取り出された連接の例
Fig. 4 Bi-grams of morphemes extracted from errors in the morphological analysis.

出し，初期の素性集合を決定する．

2.3.3 素性候補の抽出

コーパスAを用いてパラメータ推定した後，コーパスBを解析し，その解析誤りをもとに素性候補を抽出する．まず形態素の誤りを前件・後件の組として取り出し，現在の品詞分類よりも細かい分類になるように素性の候補を抽出する．

形態素の誤りを連接として取り出すには，誤りのある形態素の前後のすべての連接を抜き出せばよい．たとえば，**図3**ではコーパスの2，4，5個目の形態素が解析結果では誤りとなっており，コーパスと解析結果の形態素をそれぞれ次のようにおくことができる．

コーパス $m_1\ m_2\ m_3\ m_4\ m_5\ m_6\ m_7$

解析結果 $m_1\ m_2'\ m_3\ m_4'\ m_6\ m_7$

この場合，次の9個の連接が取り出される．

- $(m_1, m_2), (m_1, m_2')$
- $(m_2, m_3), (m_2, m_3')$
- $(m_3, m_4), (m_3, m_4')$
- (m_4, m_5)
- $(m_5, m_6), (m_4', m_6)$

このようにして取り出されたすべての連接について，現在の品詞分類よりも細かい分類の素性の候補を抽出する．たとえば，現在の前件・後件の品詞分類が**図6**のようになっていた場合，**図4**のように取り出された連接のうち，3行目の連接（**図5**）に対して抽出する素性を考える．素性候補を抽出する際に品詞分類を詳

前件	助詞-格助詞-引用 * * と
後件	形容詞-自立 形容詞・イ段 基本形 おいしい

図5 素性候補のもととなる連接の例
Fig. 5 Bi-grams of morphemes for extracting candidates of features.

前件	助詞-格助詞-引用
後件	形容詞-自立 形容詞・イ段 *

図6 現在の品詞分類の例
Fig. 6 An example of a bi-gram of part-of-speech tags in the current model.

細化する方法は実験によって異なり，大きく分けて以下の3通りの方法がある．ここで，素性を抽出する前の前件の品詞分類を $A = \{a_1, \cdots, a_n\}$，後件の品詞分類を $B = \{b_1, \cdots, b_n\}$ とする．

(1) SP_s：連接の片方の分類を詳細化

連接の片方（たとえば前件）からある品詞を取り出すことによって品詞分類を細かくし，取り出した品詞と，もう片方（たとえば後件）のそれぞれの品詞との連接を素性候補として抽出する．たとえば，前件の品詞 a_i を詳細化して品詞 x を取り出した場合，$(x, b_1), \cdots, (x, b_n)$ の n 個の素性が素性候補の1つとして抽出され，前件の品詞分類は $A = \{a_1, \cdots, x, a_i - x, \cdots, a_n\}$ になる．

取り出される品詞は，現在の品詞分類より細かく，誤り形態素連接と同じかそれよりも粗い品

前件	後件
助詞-格助詞-引用 * * と	全品詞

前件	後件		
全品詞	形容詞-自立	形容詞・イ段	基本形
全品詞	形容詞-自立	形容詞・イ段	* おいしい
全品詞	形容詞-自立	形容詞・イ段	基本形 おいしい

図7 SP_s によって抽出された素性候補
Fig. 7 Candidates of features extracted by the method SP_s.

前件	後件
助詞-格助詞-引用 * * 全語彙	全品詞

前件	後件			
全品詞	形容詞の全細分類	全活用型	全活用形	
全品詞	形容詞の全細分類	全活用型	*	全語彙
全品詞	形容詞の全細分類	全活用型	全活用形	全語彙

図8 SP_m によって抽出された素性候補
Fig. 8 Candidates of features by the method SP_m.

詞分類である．現在の品詞分類が図6のようになっている状態で，解析誤りから 図5のような連接が取り出された場合，**図7** の4個の素性候補が抽出される．たとえば，1番目の素性候補においては，前件の助詞「と」(上の説明の a_i に相当) と，後件の各品詞 (b_i に相当) との間であらゆる可能な連接をすべて考え，これらの連接を要素とする素性集合が素性候補となっている．最終的に最も頻度の多い素性候補がこの4個の素性候補の中にあれば，それが最適素性として選択される．

(2) SP_m：連接の片方の品詞のすべての細分類・活用・語彙を詳細化
連接の片方のある品詞 (たとえば名詞，動詞，形容詞など) について，すべての細分類，活用型，活用形あるいは語彙を同時に詳細化する．たとえば，前件の品詞 a_i のすべての活用形 $x_1, \cdots, x_m$ を詳細化した場合，$(x_1, b_1), \cdots, (x_1, b_n), \cdots, (x_m, b_1), \cdots, (x_m, b_n)$ の mn 個の素性が素性候補の1つとして抽出され，前件の品詞分類は $A = \{a_1, \cdots, x_1, \cdots, x_m, \cdots, a_n\}$ になる．
現在の品詞分類について，前件の助詞はすべての細分類が詳細化され，後件の形容詞はすべての細分類と活用型が詳細化されている状態で，解析誤りから図5のような連接が取り出された場合，**図8** の4個の素性候補が抽出される．たとえば，1番目の素性候補においては，前件の引用助詞のそれぞれの語彙 (上の説明の x_i に相当) と，後件のそれぞれの品詞 (b_i に相当) との間であらゆる可能な連接をすべて考え，これらの連接を要素とする素性集合が素性候補となっている．最終的に最も頻度の多い素性候補がこの4個の素性候補の中にあれば，それが最適素性として選択される．

(3) SP_p：前件と後件のペアを素性として抽出
特殊な連接を前件と後件のペアの形で抽出し，その部分の連接確率だけをパラメータ推定によって求める．たとえば，前件の品詞 a_i と後件の品詞 b_i から (x, y) という連接を素性として抽出する．その場合，パラメータ推定において，(x, y) の連接確率と，(a_i, b_i) から (x, y) を取りのぞいた部分の連接確率を，式 (9) を用いて別々に求めることになる．
素性として抽出される連接は，少なくとも前件か後件のどちらかが現在の品詞分類よりも細かく，前件・後件とも誤り形態素連接と同じかそれよりも粗い品詞分類のものである．現在の品詞分類が図6のようになっている状態で，解析誤りから図5のような連接が取り出された場合，**図9** の素性候補が抽出される．前件と後件の両方が現在の品詞分類と同じペアは素性として抽出されないので，全部で7個の素性候補が抽出されることになる．SP_s や SP_m によって抽出される素性候補はそれぞれが素性の集合であるが，SP_p の場合はそれぞれが1つの素性となっている．最終的に最も頻度の多い素性候補がこの7個の素性候補の中にあれば，それが最適素性として選択される．

SP_s, SP_m, SP_p によって粗い連接から詳細な連接を抽出した場合，詳細な連接が優先され，粗い連接のうち詳細な連接の部分は無効になる．

2.3.4 最適素性の選択方法

解析誤りをもとに素性の候補を取り出した後，以下の手順によって最適素性を選択し，素性集合に追加していく．学習全体の手順とともに説明する．

(1) 初期化
はじめに，パラメータ推定用素性集合 F_p の初期値として，初期の素性集合 F_{pinit} を設定する．

(2) 素性候補の選択
品詞分類 F_p のもとでコーパスAを用いてパラメータ推定を行い，コーパスBを解析する．その解析結果とコーパスBに付与されている品詞とを比較して得られた解析誤りをもとに，追加する素性集合の候補 F_c を抽出する．

前件				後件			
助詞-格助詞-引用	*	*		形容詞-自立	形容詞・イ段	基本形	*
助詞-格助詞-引用	*	*		形容詞-自立	形容詞・イ段	*	おいしい
助詞-格助詞-引用	*	*		形容詞-自立	形容詞・イ段	基本形	おいしい
助詞-格助詞-引用	*	*	と	形容詞-自立	形容詞・イ段		
助詞-格助詞-引用	*	*	と	形容詞-自立	形容詞・イ段	基本形	*
助詞-格助詞-引用	*	*	と	形容詞-自立	形容詞・イ段	*	おいしい
助詞-格助詞-引用	*	*	と	形容詞-自立	形容詞・イ段	基本形	おいしい

図9 SP_p によって抽出された素性候補
Fig. 9 Candidates of features extracted by the method SP_p.

(3) 素性候補の追加
以下の処理を，停止条件を満たすまで繰り返す．

(a) F_c に含まれる素性候補の中でまだ検査していない素性候補のうち，最も頻度の多い素性候補を1つ取り出し，F とする．F は一般に素性の集合である．

(b) 素性集合 F_p に素性集合 F を一時的に追加し，コーパスAを用いてパラメータ推定を行い，コーパスCを解析する．

(c) 素性集合 F を追加する前と後で，誤り形態素数が減少した数 n を求める．減少した誤り形態素数 n が3個以上であり，かつ今回減少した誤り形態素数 n が前回減少した誤り形態素数 n_p の1/10よりも多かった場合☆，すなわち

$$n \geq 3 \text{ かつ } n > n_p/10$$

を満たしていた場合は精度が十分に上がったと見なし，素性集合 F を正式に素性集合 F_p に追加し(2)へ戻る．

停止条件 F_c に含まれる素性候補すべてについて検査を行った場合は学習を終了する．

3. 実験と考察

3.1 人手によって設定した素性集合

学習によって獲得した素性集合との比較のために，人手で素性集合をいくつか作成した．品詞分類の細かさを適当に調節し，4個の素性集合を用意した．

Fh_1 品詞大分類のみを区別した素性集合．品詞細分類，活用型，活用形は区別していない．

Fh_2 品詞細分類までを区別した素性集合．活用型，活用形は区別していない．

Fh_3 品詞細分類，活用型，活用形のすべてを区別した素性集合．

Fh_4 品詞細分類，活用型，活用形のすべてを区別し，さらに助詞については語彙まで区別した素性集合．

☆ 3個，1/10などは経験的に求めた閾値である．

3.2 評価基準

評価の際の基準としては，再現率，適合率，連接規則のパラメータ数によって，人手で設定した素性集合と学習で得られた素性集合を比較した．再現率，適合率は次式のように形態素単位で測定した．

$$\text{再現率} = \frac{\text{一致した形態素数}}{\text{コーパスの形態素数}}$$

$$\text{適合率} = \frac{\text{一致した形態素数}}{\text{学習システムが出力した形態素数}}$$

コーパスの形態素と学習したシステムの出力した形態素について，品詞，活用型，活用形，見出し語のすべてが一致していれば形態素が一致したと見なした．

3.3 実験

実験には，訓練コーパスA，B，Cおよび評価コーパスD，Eの5種類のコーパスそれぞれについて，毎日新聞1995年度の記事約35,000文（約3000記事，約96万形態素，約45万字）の品詞タグ付きコーパスから異なる3000文（約9万形態素）ずつを無作為に取り出したものを用いた．また，学習に用いるコーパス量について検討するため，3000文のうちの一部分を用いた実験も行った．

形態素解析には日本語形態素解析システム「茶筌」[3]を使用し，IPAの文法体系[4]のもとで解析を行った．IPAの文法体系は（名詞-固有名詞-人名-姓）のように，品詞分類が階層的な構造を持っているという特徴があり，品詞分類をどの程度細かくするかが問題となる．品詞の種類は品詞細分類，活用型，活用形を含め全部で約790種類ある．また，実験に用いた形態素辞書は，コーパスA，B，C，D，E中に出現した約26,000語の語彙が含まれている．

素性を抽出する方法として SP_s あるいは SP_m を用いて素性を抽出，追加する2通りの実験を行い，その後さらに SP_p を用いて素性を抽出，追加する実験を行った．また，精度を向上させるのに必要なコーパス量を検証するため，コーパスの量を変化させて学習する実験を行った．

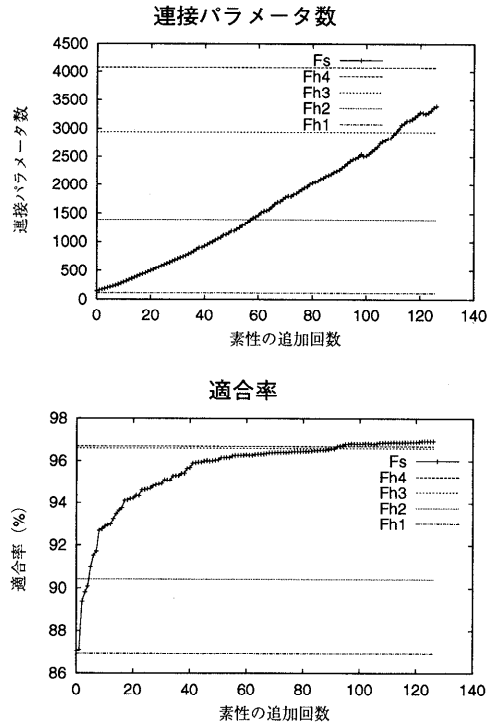


図 10 実験 E_s における連接パラメータ数・適合率の推移
Fig. 10 Changes of the number of bi-gram parameters and the precision in the experiment E_s.

3.3.1 品詞レベルで連接の片方の品詞分類を詳細化

はじめに，SP_s によって品詞レベルで連接の片方の品詞分類を詳細化し，素性を抽出，追加する実験 E_s を行った．ただし，助詞については語彙レベルまで詳細化した．

学習では，抽出した素性の中で最も頻度の多いものを素性集合に追加して，精度が上がるかどうか調べる．品詞分類が非常に粗い状態から，助詞以外の品詞についても語彙レベルまで詳細化すると，頻度は多いが精度が上がらないような素性が多く抽出されてしまい，なかなか精度が上がらない．そこで，実験 E_s では助詞のみ語彙レベルまで詳細化することにした．

初期の素性集合として，動詞と形容詞以外は品詞大分類のみを区別した品詞分類を用いた．品詞細分類と活用の両方を持つ動詞と形容詞については，抽出する素性候補があまりにも多くなってしまうため，学習の都合上（動詞-自立），（動詞-非自立）などのように品詞細分類までを区別した品詞分類を用いた．

学習で得られた素性集合を F_s とする．素性を追加するごとに，コーパス D でパラメータ推定しコーパス E を解析した．そのパラメータ数，適合率の変化を $Fh_1, \cdots, Fh_4$ とともに**図 10** に示す．再現率は適合率とあまり違いがないので省略した．なお，Fh_3，Fh_4 の適合率は差が少なく，適合率のグラフでは Fh_3 と Fh_4 はほとんど重なっている．

図 10 から分かるとおり，F_s の適合率は Fh_4 よりも高く，パラメータ数は Fh_4 より少ない．また，パラメータ数が Fh_3 より少ない段階で Fh_4 よりも高い適合率に達しているという結果が得られた．

実験 E_s では，品詞分類の詳細化を前件と後件合わせて 126 回行った結果，素性集合 F_s が得られた．詳細化された品詞分類すべてをここに載せるスペースはないので，それぞれの品詞が詳細化された回数の内訳を**表 1** に示す．

表 1 実験 E_s において各品詞が詳細化された回数
Table 1 Number of times each part-of-speech tag is specialized in the experiment E_s.

品詞	前件	後件	合計
名詞	15	10	25
記号	3	2	5
接頭詞	1	0	1
副詞	1	0	1
助詞	18	14	32
動詞	21	9	30
形容詞	4	6	10
助動詞	7	15	22
合計	70	56	126

3.3.2 品詞レベルで連接の片方の品詞のすべての細分類・活用・語彙を詳細化

次に，SP_m によって品詞レベルで連接の片方の品詞のすべての細分類・活用・語彙を詳細化し，素性を抽出，追加する実験 E_m を行った．ただし，助詞については語彙レベルまで詳細化した．

初期の素性集合として，品詞大分類のみを区別した品詞分類を用いた．

学習で得られた素性集合を F_m とする．素性を追加するごとに，コーパス D でパラメータ推定しコーパス E を解析した．そのパラメータ数，適合率の変化を $Fh_1, \cdots, Fh_4$ とともに**図 11** に示す．E_m は一度に多くの素性が抽出されるため，素性の追加回数が E_s（図 10）よりもかなり少なくなっている．再現率は適合率とあまり違いがないので省略した．なお，Fh_3，Fh_4 の適合率は差が少なく，図 11 の適合率のグラフでは Fh_3 と Fh_4 はほとんど重なっている．

図 11 から分かるとおり，F_m の適合率は Fh_4 よりも高く，パラメータ数は Fh_4 より少ない．また，パラメータ数が Fh_3 とほぼ同じ段階で Fh_4 よりも高い適合率が得られている．

実験 E_m で得られた素性集合 F_m の品詞分類を**表 2** に示す．詳細化されたものを○，詳細化されなかった

表2 素性集合 F_m の品詞分類
Table 2 Specialized part-of-speech tags of the feature set F_m.

品詞	前件				後件			
	細分類	活用型	活用形	語彙	細分類	活用型	活用形	語彙
名詞	○	—	—	—	○	—	—	—
記号	○	—	—	—	○	—	—	—
接頭詞	○	—	—	—	×	—	—	—
副詞	○	—	—	—	×	—	—	—
助詞	○	—	—	○	○	—	—	○
動詞	×	×	○	—	○	○	×	—
形容詞	×	○	○	—	○	×	○	—
助動詞	○	×	○	—	○	○	○	—

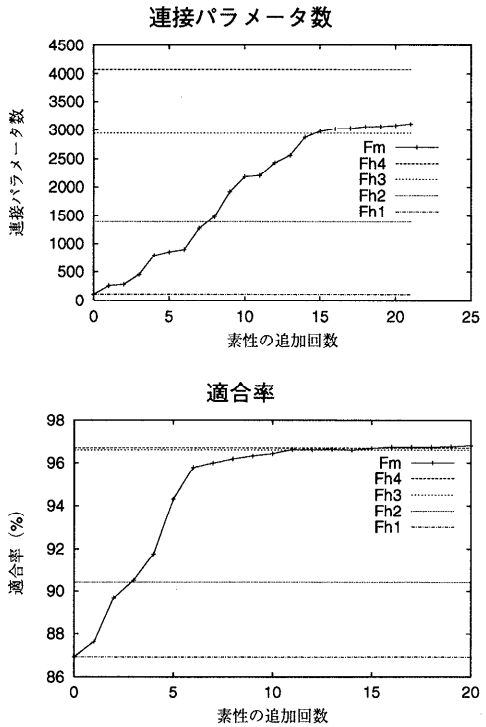


図11 実験 E_m における連接パラメータ数・適合率の推移
Fig. 11 Changes of the number of bi-gram parameters and the precision in the experiment E_m.

ものを×，詳細化できるか検証しなかったもの（具体的には助詞以外の語彙），あるいは活用を持たないために詳細化の候補が存在しないものを — で表した．なお，連体詞，感動詞，接続詞は細分類や活用を持たないので省略した．

細分類については，前件の動詞と形容詞，後件の接頭詞と副詞は詳細化されていない．活用形を持つ3個の品詞については，前件は3品詞とも活用形が詳細化されており，活用形によって後ろに接続する品詞が変化するだろうという予想どおりの結果となっている．しかし，後件については動詞が詳細化されていない．

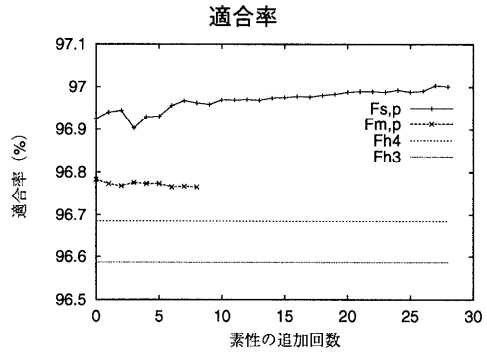


図12 実験 $E_{s,p}$ および $E_{m,p}$ における適合率の推移
Fig. 12 Changes of the precisions in the experiments $E_{s,p}$ and $E_{m,p}$.

また活用型については，動詞と助動詞は後件のみ詳細化され，形容詞は前件のみ詳細化されている．

3.3.3 語彙レベルを含む前件と後件のペアを素性として追加

(1)，(2)の学習によって得られた F_s，F_m を初期の素性集合として，さらに SP_p により語彙レベルで前件と後件のペアを素性として追加していく実験 $E_{s,p}$，$E_{m,p}$ を行った．E_s，E_m では助詞のみ語彙レベルまで詳細化したが，$E_{s,p}$，$E_{m,p}$ ではすべての品詞に対して語彙レベルの詳細化を行った．

実験 $E_{s,p}$，$E_{m,p}$ で得られた素性集合をそれぞれ $F_{s,p}$，$F_{m,p}$ とする．素性を追加するごとにコーパスDでパラメータ推定し，コーパスEを解析した．適合率の変化を図**12**に示す．

$F_{s,p}$ は F_s に比べ若干精度が上がっている．$F_{m,p}$ の精度は F_m とほとんど同じで，これ以上素性を追加してもパラメータ数が増えるだけで精度はほとんど向上しないことが予測できる．

実験 $E_{s,p}$ では28個の素性が追加された．そのうちの最初の10個の素性を図**13**に示す．また，実験 $E_{m,p}$ で追加された8個の素性一覧を図**14**に示す．

前件				後件			
名詞-固有名詞-人名	*	*		名詞-固有名詞-人名	*	*	
名詞-接尾	*	*		名詞-接尾-一般	*	*	
名詞-接尾	*	*		助詞-格助詞-一般	*	*	
名詞-固有名詞-人名	*	*		名詞-接尾-一般	*	*	
名詞-数	*	*	一	名詞-接尾-助数詞	*	*	
助詞-格助詞-連語	*	*	という	名詞-非自立-一般	*	*	
助詞-格助詞-引用	*	*		動詞-自立	サ変・スル	連用形	する
名詞-接尾	*	*		記号	*	*	
動詞-自立	サ変・スル	連用形	する	助動詞	不変化型	基本形	
動詞-自立	サ変・スル	未然形		助動詞	不変化型	*	よう

図13 実験 $E_{s,p}$ で追加された素性（全28個中，最初の10個）
Fig. 13 Features added in the experiment $E_{s,p}$ (first 10 out of the total 28 features).

前件				後件			
助詞-格助詞-一般	*	*	と	動詞-自立	サ変・スル	連用形	する
助詞-係助詞	*	*	は	形容詞-自立	形容詞・アウオ段	*	ない
動詞-自立	*	命令e		記号-読点	*	*	
助詞-連体化	*	*	の	名詞-数	*	*	一
名詞-固有名詞-地域-国	*	*	日	記号-括弧閉	*	*	
連体詞	*	*		名詞-接尾-副詞可能	*	*	後
名詞-固有名詞-地域-一般	*	*		名詞-接尾-一般	*	*	市
名詞-固有名詞-地域-一般	*	*		名詞-一般	*	*	県

図14 実験 $E_{m,p}$ で追加された素性（全8個）
Fig. 14 Features added in the experiment $E_{m,p}$ (the total 8 features).

図13，図14の連接は，詳細化される前の領域の連接に比べて，連接確率が小さいという特徴があるために詳細化されたものと，逆に連接確率が大きいという特徴があるために詳細化されたものがある．

3.3.4 コーパスの量を変化させて学習を行う実験

実験 E_s，E_m では5個のコーパス3000文ずつを用いて学習を行い，ある程度精度が向上した．ここではさらに，精度を向上させるのにどれだけの量のコーパスが必要かを検証するため，訓練コーパスA，B，Cと評価コーパスDの量を変化させて学習する2通りの実験を行った．

(1) 実験 Ec_{tr}：訓練コーパスA，B，Cの量を変化させて学習
実験 E_s，E_m において，素性を抽出するときに用いるコーパスA，B，Cの量を変化させて学習する実験 Ec_{tr} を行った．実験 E_s についてはコーパスA，B，Cをそれぞれ300，600，900，1200，1500，1800，2100，2400，2700，3000文の10通りに変化させて学習し，実験 E_m についてはコーパスA，B，Cをそれぞれ100，200，300，600，900，1200，1500，1800，2100，2400，2700，3000文の12通りに変化させて学習した．再現率と適合率は同じような変化だったので，適合率の変化のみを図**15**に示す．
実験 E_s に関しては，900文以降の精度の上昇は緩やかで，2100文以降ほとんど精度が上がっておらず，3000文よりもコーパスのサイズを大きくしてもほとんど精度が上がらないことが予測できる．また，実験 E_m については，1200文以降ほとんど精度が上がっておらず，学習に必要なコーパスのサイズは1200文で十分だといえる．

(2) 実験 Ec_{ts}：訓練時パラメータ推定用コーパスDの量を変化させて学習
実験 E_s，E_m において，評価時に用いるパラメータ推定用コーパスDの量を変化させて学習する実験 Ec_{ts} を行った．実験 E_s，E_m ともコーパスDを300，600，900，1200，1500，1800，2100，2400，2700，3000文の10通りに変化させて学習した．それぞれの適合率の変化を図**16**に示す．
E_s は1500文以降あまり精度が上がらず，2400文以降はほとんど変化がない．E_m は3000文まで精度が上昇し続けているが，2400文以降の精度向上はごくわずかで，これ以上コーパスの量を増やしても精度はほとんど上がらないことが予測できる．

表3 人手で設定した素性集合と誤り駆動の素性選択で得られた素性集合の比較
Table 3 Comparison of the feature sets selected by hand and those obtained by error-driven feature selection.

素性集合	再現率（誤り / 総形態素数）	適合率（誤り / 総形態素数）	パラメータ数	素性の抽出回数
Fh_1	86.747% (10349 / 78087)	86.904% (10208 / 77946)	111	
Fh_2	90.609% (7333 / 78087)	90.410% (7505 / 78259)	1389	
Fh_3	96.757% (2532 / 78087)	96.587% (2670 / 78225)	2942	
Fh_4	96.989% (2351 / 78087)	96.685% (2597 / 78333)	4073	
F_s	97.011% (2334 / 78087)	96.925% (2403 / 78156)	3390	1086 (126)
$F_{s,p}$	97.085% (2276 / 78087)	97.001% (2344 / 78155)	3400	3122 (28)
F_m	97.099% (2265 / 78087)	96.783% (2520 / 78342)	3112	40 (21)
$F_{m,p}$	97.097% (2267 / 78087)	96.766% (2534 / 78354)	3110	580 (8)

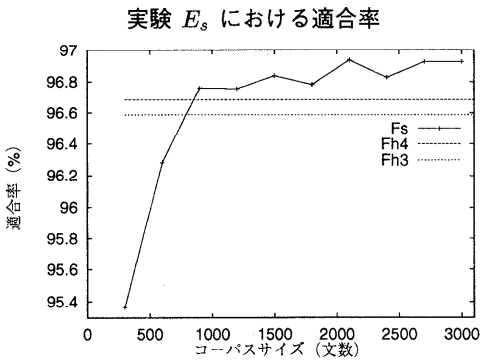


図15 実験 Ec_{tr} における適合率の推移
Fig. 15 Changes of the precisions in the experiments Ec_{tr}.

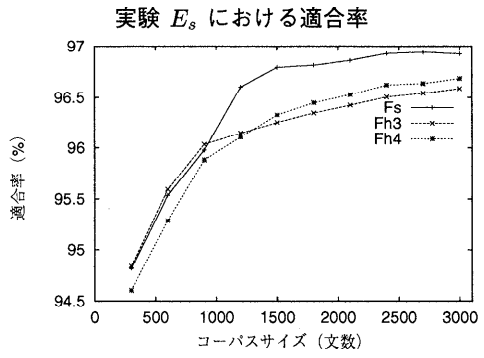


図16 実験 Ec_{ts} における適合率の推移
Fig. 16 Changes of the precisions in the experiments Ec_{ts}.

3.3.5 人手によって設定した素性集合との比較

人手によって設定した素性集合と，学習によって得られた素性集合の比較を**表3**に示す．それぞれの素性集合について，再現率，適合率，連接パラメータ数，素性を抽出した回数（カッコ内は素性を実際に追加した回数）を測定した．

学習で得られた F_s，F_m はともに Fh_4 よりも高い精度を示し，しかもパラメータ数は Fh_4 より少ないという結果が得られている．

素性を抽出した回数について F_s と F_m を比べると，F_m の方が F_s よりかなり少ない．これは素性抽出の方法の違いによるものである．F_m は F_s に比べて短い時間で学習を行うことができ，F_s よりわずかに低いがほとんど同じ精度を得られていることが分かる．

3.4 考　　察

学習による方法は人手で設定した素性集合に比べ，全般的に連接規則のパラメータ数が少なく，高い解析精度を得ることができた．人手によって設定した $Fh_1, \cdots, Fh_4$ を比べただけでは，単に Fh_4 の精度が最も高いということしか分からない．それに対し，本論文の手法によれば，精度だけに注目するならば，Fh_4 の精度はある程度限界に近く，これを大きく上

回る精度を達成することは困難であること，またパラメータ数まで考慮すれば，Fh_4 を少し上回る精度がより少ないパラメータ数で実現できることが分かる．すなわち，人手によって設定した $Fh_1,\cdots,Fh_4$ だけでは離散的な素性集合が点として存在するにすぎないが，本論文の手法により，可能な素性集合全体の空間を連続的に調べることが可能になるといえる☆．

また，コーパスの量を変えて学習を行うことにより，学習に必要な訓練コーパス，評価用コーパスの量をおおまかにつかむことができた．訓練コーパス，評価用コーパスとも 3000 文かあるいはそれよりも少ない量で，解析精度がほぼ上限に達することが分かった．

4. 関連研究

本論文の方法の特徴として 1. 品詞タグ付きコーパスを用いた統計的学習に基づく形態素解析を行う，2. 形態素解析の精度を上げるのに有効な素性を抽出，選択し，モデルに組み込んでいく，の 2 点があげられる．このような手法を用いた研究はほかにもいくつかある．

Brill[5]は，変形規則と呼ばれる操作を適用していくことにより英語の品詞タグ付けを行い，解析誤りを最も減らすような変形規則を追加していくことによって精度を向上させている．この研究は，品詞タグ付きコーパスを用いた学習を行い，解析誤りをもとに素性を追加していくという点では本論文の手法と似ているが，品詞タグ付けの方法は確率モデルに基づくものではない．また，品詞分類は最初から決められており，品詞分類の細かさをどう決定するかはこの論文では問題とされていない．

柏岡ら[6]は，形態素の構成の特徴，単語の分類体系上の特徴，文脈による特徴という 3 つの観点からとらえた属性を利用して学習した確率付決定木を用いて日本語形態素解析を行っている．形態素解析の手法は本論文と同様確率モデルに基づいているが，この手法では前後の単語の情報だけではなく，部分的な文字列の特徴や文頭の形態素の品詞など，一般的な N-gram モデルの形態素解析では扱われないような情報を扱っている．属性の選択の基準としてはエントロピーを用いており，本論文の誤り駆動の手法とは異なる．

Haruno ら[7]は，誤りが多い部分に注目してコーパスの分布を変えることを繰り返すことにより複数の確率モデルを学習し，それらのモデルを混ぜ合わせることで日本語形態素解析の精度を向上させている．確率モデル学習には文脈木を用いたマルコフモデル学習を行っている．この手法は，誤り率が高い単語のコーパス中の頻度に重み付けするという方法で誤りをモデルに反映させているが，本論文の手法では，追加素性の候補のうちで実際に誤りを減らすものを選択してモデルに追加するという直接的で確実な方法をとっている．また，本論文が bi-gram のマルコフモデルのみを扱っているのに対し，この研究では tri-gram 以上を含む可変長のマルコフモデルを用いている．この研究では後件の品詞分類は固定されており，前件の品詞分類のみについて，二階層の品詞分類（品詞大分類と品詞細分類）と語彙の詳細化を行っているのに対し，本論文では前件と後件の両方について，多階層の品詞分類，活用型，活用形，語彙の詳細化を行っている．また，詳細化する品詞を選択する方法も異なる．

5. おわりに

本論文では，誤り駆動の素性選択による日本語形態素解析の確率モデル学習の手法について述べた．解析誤りをもとに品詞分類を素性として抽出し追加していくことで，形態素解析の精度を向上させるのに有効な品詞分類を自動的に学習した．これにより，人手で品詞分類を調節するわずらわしさを解消することができた．学習によって得られた品詞分類を使って解析を行った結果，少ないパラメータ数で高い解析精度を得られることが分かった．また，品詞分類の細かさや分類の方法によってパラメータ数と解析精度がどのように変化するかといった，品詞分類の性質や特徴をつかむことができた．

今後の予定として，素性の抽出と追加の方法を改善すること，抽出する素性を様々に変化させて実験を行うこと，可変長の確率モデル学習を行うことなどを考えている．

☆ ただし，あらゆる可能な素性集合全体をしらみつぶしに探索することは，効率の点で実現不可能であり，本論文の手法においても，全探索空間のうちの一部に対してヒューリスティック探索を行っている．

参考文献

1) Church, K.: A Stochastic Parts Program and Noun Phrase for Unrestricted Text, *ANLP-88*, pp.136–143 (1988).
2) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proc. 15th COLING*, pp.201–207 (1994).
3) 松本裕治，北内 啓，山下達雄，今一 修，今村友明：日本語形態素解析システム『茶筌』version 1.0 使用説明書，Information Science Technical Report NAIST-IS-TR97007, Nara Institute of

Science and Technology (1997).
4) データベースワークショップテキストグループ：テキストデータベース報告書，技術研究組合新情報処理開発機構 (1995).
5) Brill., E.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, *Computational Linguistics*, Vol.21, No.4, pp.543–565 (1995).
6) 柏岡秀紀，Eubank, S.G., Black., E.W.: 確率付決定木を用いた日本語形態素解析，言語処理学会第3回年次大会論文集，言語処理学会，pp.433–436 (1997).
7) Haruno, M. and Matsumoto, Y.: Mistake-Driven Mixture of Hierarchical Tag Context Trees, *Proc. 35th Annual Meeting of ACL and the 8th Conference of EACL*, pp.230–237 (1997).

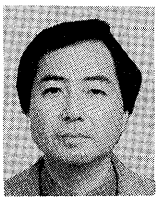
(平成10年8月7日受付)
(平成11年2月8日採録)

北内　啓

1996年京都大学理学部数学科卒業．1998年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．同年，(株) NTTデータ通信（当時）入社，現在に至る．自然言語処理の研究に従事．

宇津呂武仁（正会員）

1989年京都大学工学部電気工学第二学科卒業．1994年同大学大学院工学研究科博士課程電気工学第二専攻修了．京都大学博士（工学）．同年，奈良先端科学技術大学院大学助手，現在に至る．自然言語処理の研究に従事．人工知能学会，日本ソフトウェア科学会，言語処理学会，ACL各会員．



松本　裕治（正会員）

1955年生．1977年京都大学工学部情報工学科卒業．1979年同大学大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．1984～85年英国インペリアルカレッジ客員研究員．1985～87年（財）新世代コンピュータ技術開発機構に出向．京都大学助教授を経て，1993年より奈良先端科学技術大学院大学教授，現在に至る．京都大学工学博士．専門は自然言語処理．人工知能学会，日本ソフトウェア科学会，言語処理学会，認知科学会，AAAI，ACL，ACM各会員．