隠れマルコフモデルによる日本語形態素解析のパラメータ推定

竹内孔一[†] 松本裕治[†]

本論文では日本語形態素解析システムにHMM（Hidden Markov Model）を適応する手法について提案する．日本語では英語と異なり，わかち書きがされていないため，HMMパラメータの初期確率を等確率にした単純な学習では精度が上がらない．よって以下の3つの手法に対するHMM学習の効果について実験を行った．1）初期確率の影響．2）文法制約の導入．3）スムージング．最初の実験から初期確率については少量であっても正確なタグ付きコーパスから獲得することがHMM学習に大きく効果があることを明らかにする．次に文法による制約と確率の再推定におけるスムージング化を行った場合，人手により整備されている日本語形態素解析システムと同等以上の解析精度が得られることを示す．

HMM Parameter Learning for Japanese Morphological Analyzer

KOICHI TAKEUCHI[†] and YUJI MATSUMOTO[†]

This paper presents a method to apply Hidden Markov Model to parameter learning for Japanese morphological analyzer. When we pursued a simple approach based on HMM for Japanese part-of-speech tagging, it gives a poor performance since word boundaries are not clear in Japanese texts. We especially investigate how the following two information sources and a technique affect the results of the parameter learning: 1) The initial value of parameters, i.e., the initial probabilities, 2) grammatical constraints that hold in Japanese sentences independently of any domain and 3) smoothing technique. The first results of the experiments show that initial probabilities learned from correctly tagged corpus affects greatly to the results and that even a small tagged corpus has an enough effect for the initial probabilities. The overall results gives that the total performance of the HMM-based parameter learning outperforms the human developed rule-based Japanese morphological analyzer.

1. はじめに

日本語形態素解析は自然言語処理を行ううえで最も基本的かつ重要な処理である．我々の研究室で開発している日本語形態素解析システムJUMANは形態素間の接続と単語に対してコストを用いることにより曖昧性の絞り込みを行っている．この形態素間の接続のコストと各単語に対するコストの2つのパラメータでJUMANの精度が決定されるが，これらは従来人手により与えられてきた．しかし，このパラメータは分野に依存する[8]ため，分野の違いにより人手で変更するのは容易ではない．

近年，コーパスから統計的手法を用いた形態素解析のパラメータ学習に関する実験が行われている．人手によるのではなく，コーパスからの学習によってパラメータが獲得できれば分野の違いに対してもコーパスを用意することで柔軟に対処することができる．Church[3]らは英語のタグ付けにおいてBrown Corpusを利用したtrigramを用いた統計モデルを用いて95%の精度を獲得した．Cutting[4]，Merialdo[6]，Elworthy[5]らはHMMを利用してタグのない大規模なコーパスによる学習を行い95～96%の高い精度を得ている．しかしながら，日本語では英語のような単語のわかち書きがなされていないことと，語順の自由度が大きいことから単純な応用では成功しない．

Chang[1]らは中国語のタグ付けにおいてHMMを用いたが，その際，わかち書きされた大量のコーパスを用いた．また，日本語においては，大量のタグ付きコーパスによる学習でNagata[7]が95%の精度を獲得した．現在RWCやEDRなどでタグ付きコーパスが整備されようとしているが，下記にあげる理由から我々は大量のタグ付きコーパスを仮定しない方針で実験を行った．

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Informatin Science Nara Institute of Science and Technology

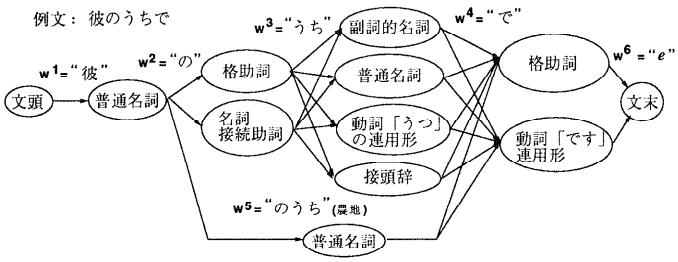

図1 日本語入力に対するHMMの状態遷移
Fig. 1 HMM state transition of Japanese input.

- 大規模なタグ付きコーパスを作成するのは大変なコストがかかる．また，使用される品詞集合や単語の定義が統一されていないために完全にデータを活用することができない場合が多い．
- 形態素間の接続に関する統計的性質はテキストの分野によって異なる[8]．そのため各分野に対してタグ付きコーパスを用意する必要がある．

本報告で我々は日本語の形態素解析システムのパラメータをHMMを用いて精度を向上させるためにどのような手法を用いればよいかを提案する．具体的には次の3つの手法の効果を実験によって確かめた．1）初期確率を獲得する手法の影響．2）文法制約の導入．3）スムージング．2）の文法制約とはどのような分野の文においても接続しない形態素もしくは品詞の接続関係を意味している．初期値に関する実験の結果から，初期確率値を獲得する方法の違いによって後のForward-Backwardアルゴリズムによる大量のテキストデータだけを用いた学習の結果が大きく異なることが明らかになった．最適な精度を獲得するためには，初期確率値には人手により作成させた誤りの少ない少量のタグ付きコーパスを利用して，2），3）の手法を適応する必要がある．このとき得られた精度は人手で整備されている日本語形態素解析システムJUMANの精度を超える結果となった．

1）の初期値に関する実験結果は英語のタグ付けにおけるMerialdo[6]とElworthy[5]も同様な結果を得ている．ただ，Merialdoは大量のタグ付けコーパスを初期値として必要とすると述べているが，我々の結果はElworthy[5]の場合と同様大変少ないタグ付きコーパスを初期値としても日本語の場合でも十分効果があることを示した．本手法を用いれば，日本語同様にわかち書きを持たない中国語などの言語にも同様の方法でアプローチすることができるはずである．

☆ 実際本論文でもEDRのタグ付きコーパスを使用する実験を行ったが，タグセットがあまりにも異なるために十分活用できなかった（3.1節参照）．

本論文の構成は次のとおりである．まず2章で日本語形態素解析システムJUMANとbigramのHMMの対応を示し，JUMAN-HMMシステムを構築する．その後，3章で初期値に関する実験ならびに文法制約とスムージングの効果について結果を明らかにし，4章でまとめる．

2. JUMAN-HMMシステム

2.1 JUMAN日本語形態素解析システム

JUMAN[9]は奈良先端科学技術大学院大学と京都大学によって開発されてきたルールベース型の日本語形態素解析システムである．JUMANでは2種のコストを用いて曖昧性を絞り込んでいる．1つは形態素間の接続に関するルールに対するコストで，他は各単語に対するコストである．文を解析する際，辞書引きによって得られた可能な解釈の組合せからコスト値を計算してもっともらしい解を選択する．解析結果は**図1**に示すように単語のラティス状の構造が出力される．曖昧性を残さずに最適なパスを計算した場合，そのわかち書きならびに品詞付けの精度は新聞の社説記事に対して約94%の精度を持っている．

このJUMANのラティス状の構造は1つの状態を1形態素と見なすとHMMの状態遷移構造と1対1に対応すると見なすことができる．その際，コスト値は確率値の逆と見なせばよい．つまり，HMMにおける確率値の対数の絶対値をとることによって確率のかけ算の値はコスト値の和に対応させることができる．また，英語の品詞タグ付けの場合[2),4)]と異なり，わかち書きの違いによる異なった単語列のパスが生じるため，それを扱えるようにHMMの構造を変更する必要がある．次章では本論文で用いたHMMの構造を明らかにする．

2.2 日本語形態素解析における隠れマルコフモデル

図1に示したように日本語の文に対する形態素解析を行った結果，各単語の品詞に対する曖昧性とわかち書きの違いによる曖昧性の2種類が存在する．これら

の異なるパスをすべて扱うために，我々はHMMに対していくつかの変更を行った．図1は日本語形態素解析の出力結果であるが，同時にHMMの状態遷移構造を示している．この図を用いて我々が使用するHMMの構造を説明する．まず状態遷移の最初と最後に“文頭”と“文末”という2つのダミーノードを設けた．ただし，“文末”ノードに遷移するためには空語“e”で状態遷移するものとする．これにより日本語入力文に対するすべての曖昧性のパスは“文頭”から始まって“文末”で終了することになる．

次に文の確率について定義する．日本語入力文字列Mは辞書引きの結果何通りかの単語列が得られる．わかち書きの違いによる曖昧性と品詞の曖昧性を含むすべての可能な単語列の集合をL_Mとする．前からk番目の単語の表記と品詞（状態）を(w_k, s_k)で表すと，L_Mの1要素は$(s_0)(w_1, s_1)..(w_k, s_k)..(w_{n+1}, s_{n+1})$と表現することができる．ここで$s_0, s_{n+1}, w_{n+1}$は“文頭”，“文末”，空語“$e$”をそれぞれ表す．入力文字列$M$に対する文の確率$P(M)$は，これら単語列の曖昧性をすべて足し合わせた確率と定義できる．つまり，図1において2つの異なるわかち書き|彼 の うち で|と|彼 のうち で|に対する品詞の曖昧性を含めた計18通りの単語列パスの確率をすべて足し合わせた確率が文の確率$P(M)$である．よって，$P(M)$を次のように定義する．

$$P(M) = \sum_{(s_0)..(w_{n+1},s_{n+1})\in L_M} \prod_{i=1}^{n+1} P(s_i|s_{i-1})P(w_i|s_i) \tag{1}$$

ここで$P(s_i|s_{i-1})$は状態遷移確率，$P(w_i|s_i)$は単語の生成確率を示しており，このモデルでは直前の状態だけを遷移確率の条件とみるbigramモデルとなっている．

Charniak[2]やCutting[4]らが仮定した前向き確率（forward確率）と後向き確率（backward確率）の計算も変更する必要がある．図1の単語“で”に注目すると，単語“うち”から来るパスと単語“のうち”から来るパスの2種類あることが分かる．このようにわかち書きの違いによるパスを扱えるようにする．以下の数式では状態の集合を$\{s^1, \cdots, s^i, \cdots, s^\sigma\}$と表し，辞書引きによって得られた単語の表記（シンボル）の集合を$\{w^1, \cdots, w^k, \cdots, w^E\}$とする．このとき，$k$は文頭からの順番を表すのではなく辞書引きによって得られた単語のシンボルに対して一意につけた番号を表す．w^Eは空語“e”を表す．また，w^{k-}は，シンボルw^kの直前に接続するシンボルの番号の集合を表し，w^{k+}は，w^kの直後に接続するシンボルの番号の集合を表す．たとえば，図1では，$w^{4-} = \{3, 5\}$で$w^{1+} = \{2, 5\}$である．

forward確率$\alpha_j(k)$はシンボルw^kを出力した状態s^jから“文頭”までの単語列の異なるすべてのパスの確率の総和を表し，backward確率$\beta_i(k)$はシンボルw^kを出力した状態s^iから“文末”までの同様の確率の総和を表す．

$$\begin{aligned}\alpha_j(k) &= \sum_{i=1}^{\sigma} \sum_{h\in w^{k-}} \alpha_i(h)P(s^i \xrightarrow{w^k} s^j) \\ &= \sum_{i=1}^{\sigma} \sum_{h\in w^{k-}} \alpha_i(h)P(s^j|s^i)P(w^k|s^j) \end{aligned} \tag{2}$$

$$\begin{aligned}\beta_i(k) &= \sum_{j=1}^{\sigma} \sum_{h\in w^{k+}} P(s^i \xrightarrow{w^h} s^j)\beta_j(h) \\ &= \sum_{j=1}^{\sigma} \sum_{h\in w^{k+}} P(s^j|s^i)P(w^h|s^j)\beta_j(h) \end{aligned} \tag{3}$$

よってこれらを用いて確率的な回数を定義することができる．ある状態s^iからs^jの遷移においてシンボルw^lを出力するパスの確率的回数を求めるとする．まず，s^iからs^jの遷移でシンボルw^lを出力するすべてのパスの確率を求める．これは，w^kの直後にw^lが出力する場合，forward確率とbackward確率を利用して$\alpha_i(k)P(s^i \to^{w^l} s^j)\sum_{h\in w^{k+}}\beta_j(h)$と書ける．これをすべての$k$について足し合わせて全体の確率$P(M)$で割ると確率的回数が求まる[10]．

$$C(s^i \xrightarrow{w^l} s^j) = \frac{1}{P(M)} \sum_{k=1}^{E} \alpha_i(k)P(s^i \xrightarrow{w^l} s^j) \sum_{h\in w^{k+}} \beta_j(h) \tag{4}$$

この確率的回数を用いて新たな状態遷移確率$P_e(s^j|s^i)$ならびに単語の生成確率$P_e(w^l|s^j)$を従来のHMMと同様に再計算によって求める．

$$P_e(s^j|s^i) = \frac{\sum_k C(s^i \to^{w^k} s^j)}{\sum_k \sum_j C(s^i \to^{w^k} s^j)} \tag{5}$$

$$P_e(w^l|s^j) = \frac{\sum_i C(s^i \to^{w^l} s^j)}{\sum_k \sum_i C(s^i \to^{w^k} s^j)} \tag{6}$$

HMMのパラメータは上記の数式を用いることに

よって大規模なコーパスから再推定することができる．これを組み込んだJUMAN-HMMシステムについては次節で説明する．

さて，HMMの状態 s^i は基本的に品詞であるが，助詞と助動詞については各単語の接続特徴を細かく観測するため1単語ごとに1つの状態を割り当てた．そのため状態数は143種類存在する．また，日本語において動詞や助動詞など活用語の活用形が直後の単語に影響を与えることが多い．そこで，bigramのHMMの状態遷移確率を考えるとき，活用する語が前に接続する場合，その活用語の活用型と活用形まで観測して状態遷移確率 $P(s^j|s^i)$ とする．そして，後ろ側に出現する場合は通常どおり品詞の状態とする．たとえば図1中の1状態「うち：動詞」の前後の遷移に注目すると，後方の接続に対しては P(格助詞 | 動詞 ∩ 子音動詞タ行 ∩ 基本連用形) の遷移確率を適応し，前方では P(動詞 | 名詞接続助詞∩"の") の遷移確率を適応する．よって形態素間の状態遷移の組合せ数を考えると，前側は活用型と活用形の組合せがあるため約700種類となるため約 700 × 143 通り存在する．

2.3 JUMAN-HMMシステム

図2 にJUMAN-HMMシステムを示す．このシステムは大きく分けて2つの部分からなる．1つは初期値獲得モジュールで他はHMM学習モジュールである．HMM学習モジュールではタグ付けされていないコーパスを入力としてJUMANの出力をHMMが受け取る流れとなっている．これはJUMANが生成するラティス状の構造がHMMの遷移構造と1対1に対応するため，いわばJUMANが辞書引きによる単語のグラフ構造を作る役割をして，HMMがそれを元に確率的数えあげを行う役割をしている．この際，JUMANの接続コストならびに単語の生成コストはHMMの持つ接続確率ならびに単語の生成確率とまったく対応した値とする（つまり，確率値の対数をとった絶対値をコスト値とする）．そして，コーパスを読み終えた後，再推定された連接確率ならびに単語の確率が辞書に格納される．2度目以降はこれの繰り返しにより学習を進める．

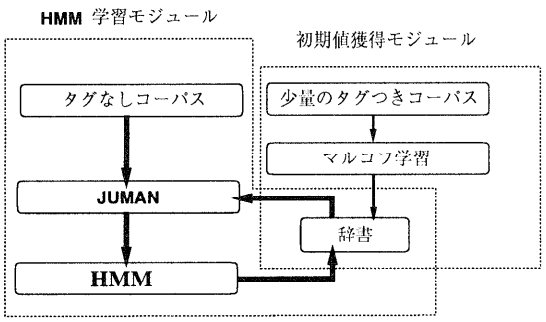

図2 JUMAN-HMMシステム
Fig. 2 JUMAN-HMM system.

初期値獲得モジュールではこの学習を始めるための最初の接続確率ならびに単語の生成確率を小さなタグ付きコーパスからマルコフ学習によって獲得する．ここで得られた確率値はHMMにはそのまま渡され，JUMANにはコスト値に変換して与えることにする．

3. HMMのパラメータ学習

我々は新聞記事によるHMM学習を行ったところ局所解に陥ってその解析精度は約80%にすぎなかった．これは英語のタグ付け[2),4)]における同様な手法を用いた場合の結果と比較すると高い精度ではない．これは図1で示した“のうち”のようにわかち書きされていないことによる曖昧性の増加と，「家で本を読む」「本を家で読む」のように語順の自由度が大きいことが主な原因である．これより単純な応用では十分精度が得られないことが分かる．

そこで精度を改善するためにHMM学習に以下に述べる3つの方法を導入し，その実験を行った．まず初期値が精度に重要な役割を果たすことを明らかにする．つまり，初期確率の獲得の方法によってHMM学習後の解析精度が大きな影響を受けることを明らかにする．次に文法制約が効果的であることを示す．文法制約とは日本語文ではありえない単語間の接続関係を記述することで，この制約により，文法的な誤りを減少させる．最後にスムージングに関する実験を行う．以上の3つの方法を導入した場合，ルールベースの日本語形態素解析の精度を超える結果が得られたことを示す．

3.1 初期値に関する実験

図2に示したように，HMM学習を実行するには最初に初期値が必要となる．初期値とは形態素間の接続に関する確率と単語の生成に関する確率である．大量のタグ付きコーパスが存在するなら問題は非常に簡単になる．つまり，タグ付きのコーパスがあればマルコフ学習によって単純な頻度の数えあげにより精度の高い結果が得られるからである．しかし，日本語では各分野に対して十分大きなタグ付きコーパスを獲得するのは困難である．また，特に日本語では一般的な単語の定義がなく，活用語の扱い方や助詞，助動詞の種類などが統一されていないためにタグ付きコーパスがあったとしてもそういった文法の違いによって十分活用することができないことが多い．現在大量のタグ

付きコーパスとしてEDR[11]のタグ付きコーパスが存在するが助詞の扱い，ならびに活用語の語尾の扱いが我々の文法とかなり異なるため，十分活用することができない．

そこで，我々はHMMパラメータ学習で良い結果を得るために大量のタグ付きコーパスは仮定しないが，良い初期値を求めることにする．使用する初期値の影響がHMM学習の結果にどのように反映されるかを観測するために以下のようなタグ付きコーパスを用意して実験を行った．

(1) **EDRタグ付きコーパスの利用**：我々の品詞集合とかなり異なるため単語のわかち書き情報を利用する．タグは現在のJUMANに付けさせる．

(2) **現在のJUMANによりタグ付けした朝日新聞の社説記事（65,000文）**：現在のJUMANでタグ付けした出力をコーパスとする．ただし5～7%の誤りを含む．

(3) **人手で作成したタグ付きコーパス（300文）**：朝日新聞の社説の記事に対して人手でタグを付けた．非常に少ないが誤りもほとんどない．

HMM学習用のコーパスとして朝日新聞の社説約200,000文を用いた．上記の(1)，(2)はタグ付きコーパスが大量に存在するがいくらか誤りを含んでいる．一方(3)は量は大変少ないがほとんど誤りを含まない．まず，最初の評価としてこれら各コーパスから得られた初期確率をコストに変換してJUMANに与え，そのJUMANの解析能力によって評価する．その結果を**表1**に示す．

表中の数字は誤り率を示している．これは，各初期値を元にしたJUMANが誤った品詞付けを行ったり単語のわかち書きを誤った場合の形態素数を，解析した全形態素数で割った値である．つまり，100%からこの誤り率を引くと適合率が求まる．また，括弧中の数字は助詞の細分類の違いを無視した場合の誤り率である．これはbigramのような局所的な情報だけでは，助詞の細分類を分けるのは難しいことが多いため設けた．タグ付きコーパス1は初期値と評価用の両方に用いている（約7,400形態素）．タグ付きコーパス2は同様に人手により作成されたタグ付きコーパスで評価用にだけ用いた（約6,600形態素）．両方とも社説92年5月の記事から順に取り出した文で，学習用のコーパス約200,000文中には含まれていない．当然insideデータによる実験が一番良い結果となった．表中の最後の行には比較として現在のJUMANの解析精度を示した．

表1 初期値を反映したJUMANの精度
Table 1 Error rates based on initial probabilities.

初期値のコーパス	tagged corpus 1 (300文)	tagged corpus 2 (300文)
1. EDR corpus	16.9 (16.0)	14.8 (13.6)
2. JUMAN corpus	9.9 (8.7)	7.6 (6.4)
3. tagged corpus 1	1.9 (inside)(1.7)	6.4 (5.5)
current JUMAN	7.6 (6.1)	5.5 (4.7)

この表の結果から，誤りを含んだ大量のタグ付きコーパスよりも誤りをほとんど持たない少量のタグ付きコーパスの方が大変有用であることが分かる．EDRコーパスの結果があまりにも悪い．これはEDRコーパスに出現する単語や記号の統計的性質が社説と異なったためである．よって，他の2つの初期値を利用して約200,000文（タグなし）を用いたHMM学習を行う．この結果を**表2**に示す．

この表からJUMANコーパスのように誤りを含む初期値ではHMM学習では改善されるどころかかえって悪化することが分かる．一方，誤りの少ないタグ付きコーパスならば少量であってもHMM学習により適合率がわずかであるが改善される．outsideデータの結果を現在のJUMANと比較すると精度的にはまだ劣っている．

この結果を得たHMM学習の学習回数は1回でそれ以上学習を進めると解析精度が落ちる傾向が観測された．これは英語のタグ付けにおけるMerialdo[6]やElworthy[5]らでも同様に観測される現象で，このことから単純なHMM学習だけではあまり大きな精度の改善は期待できないことが分かる．

3.2 文法制約の導入

本節では文法制約の導入がHMMの学習結果にどのように影響するか明らかにする．上記の学習後の獲得された接続ルールを観測すると文法的にまずありえない接続の確率が大きく獲得されて誤りの原因となることが観測された．たとえば動詞の語幹の直後に助詞が接続したり，接続詞の直後に助詞が来るなどである．これは未出現の接続関係に対しては低い確率値を与えているため数多く事例が出現するとその方向に学習が進むためである．これらの接続のうち，どのような分野でもありえない接続関係に対しては，あらかじめ確

表2 各初期値によるHMM学習後の結果
Table 2 Error rates of HMM trained results.

初期値のコーパス	tagged corpus 1 (300文)	tagged corpus 2 (300文)
2. JUMAN corpus	16.2 (15.4)	14.0 (13.3)
3. tagged corpus 1	3.8 (inside)(3.6)	6.0 (5.2)

（学習コーパス：新聞の社説約200,000文）

表3 文法制約を入れた場合のHMM学習の結果
Table 3 Error rates of HMM trained results with grammatical knowledge.

初期値のコーパス	tagged corpus 1 (300文)	tagged corpus 2 (300文)
tagged corpus 1	3.5 (inside)(3.3)	5.4 (4.7)
tagged corpus 2	6.9 (6.4)	3.3 (inside)(3.1)
current JUMAN	7.6 (6.1)	5.5 (4.7)

(学習コーパス：新聞の社説約200,000文)

表4 文法制約を入れた場合のHMM学習の結果
Table 4 Error rates of HMM trained results with grammatical knowledge.

初期値のコーパス	tagged corpus A (300文)	tagged corpus B (300文)
tagged corpus A	3.0 (inside)(2.8)	5.8 (5.2)
tagged corpus B	5.5 (4.9)	3.1 (inside)(2.9)
current JUMAN	7.2 (5.9)	6.3 (5.0)

(学習コーパス：新聞の社説約200,000文)

率を0として学習が進まないように制約を入れる．この実験では15個の文法制約ルールを用いた[10]．実験結果を**表3**に示す．

この表からルールベースのJUMANの精度をわずかながら超えていることが示されている．また，**表4**にはさきほどのタグ付きコーパス1と2を混合して2つに割った新たなタグ付きコーパスAとBを用いて同様な実験を行った結果を示している（A，Bとも約7,000形態素）．この結果からもHMM学習と文法制約の結果がルールベースのJUMANの精度を超えることを示している．

3.3 スムージング化とモデルの状態数の追加

本論文で使用しているHMMは状態数が143種類でbigramのHMMである．ただし，活用語に対しては接続の前側に来る場合に活用形が後接続の単語に影響を与えるので活用型と活用形の異なりまで観測する．よって状態遷移の組合せ数は約 700×143 種にも及ぶ．タグ付きコーパスの量は300文程度なのでこの組合せ数ではかなり未出現の状態遷移の組合せが多い．そこで，状態遷移確率に関してスムージング化を行う．特に，1つの語が1つの状態をなしている助詞，助動詞ならびに活用語に対する活用形に関してスムージングを行いその実験結果を以下に示す．

また，スムージング化するだけでなく助詞と助動詞と同様に1つの状態に割り当てる単語数を増やす実験も行った．スムージング化を行い，さらに状態に割り当てる単語数を追加することによって得られた結果，従来のHMM学習の結果を上回る精度を得た．

我々が提案している手法はコーパスさえ用意すれば実行することができる．そこで，以下の実験では分野を変えて，日経新聞の新聞記事を利用する．前実験と同様に初期値ならびに比較用としてタグ付きコーパス300文を2つ（tagged corpus 3（300文約8,600形態素），tagged corpus 4（300文約8,600形態素））を用意し，HMM学習用コーパスとして約20万文を用意した．これらは，コーパスから約201,000文を取り出し，そこからほぼ等間隔に1,000文取り出した中の600文をタグ付きコーパス，残りを学習用コーパスとして用意した．

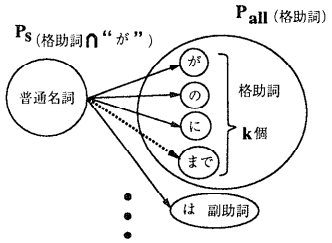

図3 スムージング化の例1
Fig. 3 Example of smoothing 1.

3.3.1 スムージングの方法

スムージングには後方接続に対するスムージングと前方接続に対するスムージングがあり，両者とも同様の方法で求めることができる．一言でいえば，統計的に意味のあるまとまりを状態と仮定して，その平均確率と元の確率との混合によりスムージング化を行う．

まず，後方接続から説明する．後方接続の場合，助詞と助動詞をスムーズ化する．**図3**は後方接続のスムージングの例で，「普通名詞」から各「格助詞」の語に状態遷移する場合が示されている．この図では，スムージングで使用する統計的に意味あるまとまりとして品詞細分類の「格助詞」という状態を作成している（ex. 他には，副助詞など☆）．また，図中の実線の矢印ではコーパスに出現した遷移を表しているが点線の矢印の"まで"はコーパスに出現しなかった遷移を表している．

このとき，たとえば P(格助詞∩"が"| 普通名詞) の確率を求めたいとする．まず，従来どおりコーパスに現れた遷移確率として P_s(格助詞∩"が"| 普通名詞) を求め，それと同時に「格助詞」としての状態遷移確率 P_{all}(格助詞 | 普通名詞) も求めておく．この P_{all} を未出現の単語も含めた格助詞の語の総数 k 個で割り平均確率 P_{ave} を求める．

$$P_{ave} = \frac{P_{all}}{k} \tag{7}$$

☆ 付録に使用した品詞を掲載する．

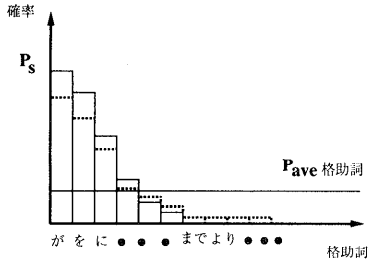

図4 スムージング化による分布の変化
Fig. 4 Graph of transition by smoothing.

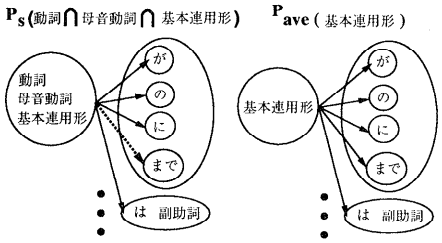

図5 スムージング化の例2
Fig. 5 Example of smoothing 2.

これを混合比 λ の割合で元の確率 P_s と混合してスムーズ化した確率値が求まる．

$$P = (1-\lambda)P_s + \lambda P_{ave} \quad (8)$$

これは助動詞の場合も同様に行うことができる．つまり，この式では元々の出現確率 P_s で分布に偏りがあった確率を，平均確率 P_{ave} を用いて，これより大きい確率は引き下げて，少ない確率は引き上げる役割を果たしている．**図4** はこの様子を図に示したもので，実線が元の確率を示し点線がスムージング後の確率を示している．

次に，前方接続に対するスムージング化を説明する．前方にスムージングする場合は活用語の活用形に対して行う．**図5** に動詞の連用形に関するスムージングを示している．方法としては後方の場合と同様で，意味のあるまとまりとして「基本連用形」という状態を前方に作成しコーパスからその遷移確率 P_{ave}(基本連用形) を計算する．同時に従来どおりの確率 P_s(動詞∩母音動詞∩基本連用形) を求めておく．この両確率を式 (8) に適応すれば同様にスムージングした確率値が獲得できる．これらスムージングに使用する式はすべて確率値を獲得してから計算し始めるのでマルコフ学習でも Forward-Backward アルゴリズムによる HMM 学習の結果に対してもスムージングを行うことができる．

3.3.2 HMM の状態数を追加する

助詞や助動詞のように1状態に1単語を登録する語をさらに追加してみる．登録する単語を決定するためにタグ付きコーパスを利用した．用意した tagged corpus 3,4 の 600 文を用いてその文中の単語の頻度を数え上げ，普通名詞，固有名詞，数詞，サ変名詞以外の単語について頻度の多い語を使用することにする．ここでは上位 20 語を使用することにした．内訳としては記号や接尾辞が多く，“来る” などのよく使われる動詞も入っている．使用した 20 語は付録に記載する．

表5 初期値におけるスムージングの結果
Table 5 Error rates of smoothed initial probability.

初期値のスムージング法	tagged corpus 3 (300 文)	tagged corpus 4 (300 文)
助詞と助動詞	1.4 (*inside*) (1.2) 6.1 (5.5)	6.0 (5.4) 1.5 (*inside*) (1.1)
活用形	1.4 (*inside*) (1.2) 5.6 (5.1)	5.3 (4.8) 1.4 (*inside*) (1.1)
上記の2つ	1.5 (*inside*) (1.2) 5.6 (5.1)	5.2 (4.7) 1.5 (*inside*) (1.1)
スムージングなし	1.1 (*inside*) (1.1) 5.2 (4.7)	5.1 (4.7) 1.4 (*inside*) (1.1)

3.3.3 初期値スムージングの実験結果

本論文で提案している JUMAN-HMM システムでは特に初期値を獲得するためのタグ付きコーパスは少ない量を仮定している．そこで，初期値獲得の際のマルコフ学習に対してスムージングを行った場合の HMM 学習に対する効果を観測してみた．まず，初期値を獲得した段階の JUMAN の解析精度でスムージングの評価を行ってみる．**表5** は助詞と助動詞のみ（つまり，後方のみ）スムージングした場合，活用形のみ（前方のみ）をスムージングした場合，混合した場合の結果を示している．このときのスムージング比率 λ は 0.3 とした．これは後の単語化した状態を追加した実験における最良の比率である．各欄の上段，下段はそれぞれタグ付きコーパス 3，4 を初期値として学習したものである．また，実験では文法制約をかけており，これ以降の実験ではすべて文法制約を行っている．表5 の結果は初期値の段階ではスムージングをいっさい行わない従来の結果（すなわち比率 $\lambda = 0$）のときが最も良い．しかし HMM 学習後の結果**表6** では初期値でスムージングを行った場合の精度が従来どおりの学習に比べて若干良くなろうとしている．特に活用形のみをスムージングした場合は良い．このように初期値獲得段階でのスムージングの効果は HMM 学習をした後に発揮されるため，初期値のスムージングで精度が悪くなってもそこで評価することはできない．

3.3.4 HMM 状態数の追加に関する実験

前項のスムージングに対してさらに HMM の状態数

表6 HMM学習後の各スムージングの結果
Table 6 Error rates of HMM learned results with smoothed initial probability.

初期値のスムージング法	tagged corpus 3 (300文)	tagged corpus 4 (300文)
助詞と助動詞	2.7 (*inside*) (2.4) 4.4 (3.6)	4.2 (3.7) 3.3 (*inside*) (2.9)
活用形	2.6 (*inside*) (2.4) 3.8 (3.3)	4.2 (3.6) 3.1 (*inside*) (2.7)
上記の2つ	2.7 (*inside*) (2.4) 4.1 (3.3)	4.3 (3.8) 3.4 (*inside*) (2.8)
スムージングなし	2.6 (*inside*) (2.3) 4.2 (3.6)	4.3 (3.6) 3.1 (*inside*) (2.7)

(学習コーパス：日経新聞の記事約200,000文)

表7 HMMの状態数を追加した場合の初期確率の精度
Table 7 Error rates based on initial probabilities with additional states of HMM.

初期値のスムージング法	tagged corpus 3 (300文)	tagged corpus 4 (300文)
助詞と助動詞	1.3 (*inside*) (1.2) 6.0 (5.6)	6.2 (5.7) 1.3 (*inside*) (1.0)
活用形	1.2 (*inside*) (1.2) 5.4 (5.0)	5.4 (4.9) 1.3 (*inside*) (1.1)
上記の2つ	1.3 (*inside*) (1.2) 5.3 (4.8)	5.2 (4.8) 1.5 (*inside*) (1.2)
スムージングなし	1.2 (*inside*) (1.1) 6.3 (5.8)	6.2 (5.7) 1.2 (*inside*) (1.0)

を追加した場合の学習効果について実験を行う．**表7**に3つのパターンのスムージングを行った場合とスムージングを行わなかった場合の各初期値の誤り率が示されている．スムージング比率λは0.3である．これは0.1ステップで実験して最も誤り率が低かったため採用した．前項の表5と比較すると状態数の追加により，スムージングをしなかった場合よりスムージングをした3つの方が良い初期値が獲得されている．特に活用形に関するスムージングを行うと効果的である．しかし，表7中では活用形ならびに助詞と助動詞をスムージングした場合が最も良い初期値となっているが，表5に示した状態数も追加せずスムージングもしない従来どおりの初期値の精度とほぼ並ぶ程度である．

では，これらの初期値を元にしてHMM学習した結果を**表8**に示す．この表中のタグ付きコーパス3に対しては活用形ならびに助詞と助動詞をスムージングした場合が良い結果を得ているが，タグ付きコーパス4に対しては優劣がはっきりしない．つまり，初期値ではかなり良かったスムージングした初期値も，HMM学習の結果スムージングの効果が現れないことがあることを示している．ただ，表6と比べてHMMの状態数を追加した効果を見ると，スムージングするしないにかかわらず0.1から0.4の値で誤り率が低くなる傾向があることからHMMの状態数を追加することは有効であった．

表8 HMMの状態数を追加した場合のHMM学習の結果
Table 8 Error rates of HMM trained results with additional states.

初期値のスムージング法	tagged corpus 3 (300文)	tagged corpus 4 (300文)
助詞と助動詞	2.6 (*inside*) (2.4) 4.1 (3.6)	4.2 (3.7) 3.0 (*inside*) (2.6)
活用形	2.6 (*inside*) (2.5) 3.9 (3.4)	4.1 (3.7) 3.0 (*inside*) (2.6)
上記の2つ	2.8 (*inside*) (2.5) 3.8 (3.3)	4.2 (3.8) 3.1 (*inside*) (2.6)
スムージングなし	2.5 (*inside*) (2.3) 4.2 (3.6)	4.1 (3.7) 2.9 (*inside*) (2.6)

(学習コーパス：日経新聞の記事約200,000文)

表9 HMM学習におけるスムージングの結果
Table 9 Error rates of smoothed HMM learning results.

HMM学習時のスムージング法	tagged corpus 3 (300文)	tagged corpus 4 (300文)
助詞と助動詞	2.6 (*inside*) (2.3) 4.1 (3.5)	4.1 (3.7) 3.0 (*inside*) (2.5)
活用形	2.5 (*inside*) (2.3) 3.9 (3.4)	4.0 (3.5) 3.0 (*inside*) (2.7)
上記の2つ	2.4 (*inside*) (2.2) 3.8 (3.2)	3.9 (3.5) 2.8 (*inside*) (2.4)
current JUMAN	5.6 (4.5)	6.1 (4.8)

(学習コーパス：日経新聞の記事約200,000文)
(HMM学習のスムージング率 $\lambda = 0.1$)

3.3.5 HMM学習のスムージングの実験結果

初期値をスムージングしてもHMM学習でその効果が薄れてしまうことがあることが前項で示された．そこで，さらなる精度向上のために初期値獲得の際にスムージングを行ったパラメータに対して，さらにHMM学習においてもスムージングを行った場合の効果を測定する．スムージングの方法は表7の3種の方法を使用し，初期値も同表の値を用いる．つまり初期値のスムージング比率は$\lambda = 0.3$でHMMの状態数は20語追加した場合である．

実験ではHMM学習時のスムージング比率は0.0から0.1ステップの刻みで0.4まで変化させた．そのときの各スムージングの方法で最も低い誤り率を示した結果について**表9**に示す．結局どのスムージング方法でもHMM学習時のスムージング比率は$\lambda = 0.1$のときが誤り率が最も低かった．

表9中の実験結果では活用形と助詞と助動詞をスムージングした場合が最も良い精度を得ている．まず，表8と比較してHMM学習におけるスムージングの効

果を比較すると 0.1 ～ 0.2% ほどタグ付きコーパス 4 で改善されている．またこの結果と表 6 のスムージングをしない従来のままの学習結果と比べるとタグ付きコーパス 3，4 どちらのアウトサイドデータに対しても 0.4% の精度の改善を得ることができた．初期値のスムージング，HMM の状態数の追加，HMM 学習時のスムージング，それぞれの効果はほんの 0.1 ～ 0.2% 程度の精度の改善であるが，これらの組合せにより適合率で 96% を超える精度を得た．この値は，表 9 の最下段に示す現在のルールベースの JUMAN の精度と比較して約 2% の精度の改善となっている．

4. まとめ

我々は日本語形態素解析システムに HMM を用いた学習システムを適応する方法について述べてきた．そこで，1）初期値の獲得方法による学習結果への影響，2）文法制約の導入，3）スムージングについて実験を行った．その結果，初期値の獲得のためには少量であっても誤りをほとんど持たない正確なタグ付きコーパスから獲得すると HMM 学習に大きく効果があることを示した．反対に誤りを多く含むと，いくら大量のコーパスがあっても HMM 学習による改善は見られなかった．また，文法制約を導入しスムージングと HMM の状態数の追加を行った結果，現行のルールベースの JUMAN に比べて約 2% の精度の向上を見ることができた．

本報告で提案した手法はでどんな分野に対しても少量で正確なタグ付きコーパスと大量のコーパスが存在すれば実行することができる．実際，本論文で新聞の社説と経済関連記事のコーパスを用いて実験を行った．このように，本手法を用いれば必要な分野に対して日本語形態素解析システムのパラメータを学習により獲得することができる．

本論文で使用している HMM は前後関係だけを観測する bigram の HMM を用いた．スムージングまで行った表 9 の HMM 学習の結果を観測してもインサイドの計算結果に対しても 3% 弱の誤りが存在する．これは bigram の限界と考えてよい．今後 trigram や n-gram などさらに遠い前後関係を観測した場合の学習法について実験していく必要がある．

謝辞 社説を使用させていただいた朝日新聞社，EDR コーパスを使用させていただいた日本電子化辞書研究所，新聞記事を使用させていただいた日経新聞社，各社に対して謹んで感謝の意を表します．

参考文献

1) Chang, C.-H. and Chen, C.-D.: HMM-based Part-of-speech Tagging for Chinese Corpora, *Proc. Workshop on Very Large Corpora*, pp. 40–47 (1993).
2) Charniak, E., Hendrickson, C., Jacobson, N. and Perkowitz, M.: Equations for Part-of-Speech Tagging, *Proc. 11th National Conference on Artificial Intelligence (AAAI-93)*, pp.784–789 (1993).
3) Church, K.: A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, *Proc. ACL 2nd Conference on Applied Natural Language Processing*, pp.136–143 (1988).
4) Cutting, D., Kupiec, J., Pedersen, J. and Sibun, P.: A Practical Part-of-speech Tagger, *Proc. 3rd Conference on Applied Natural Language Processing*, pp.133–140 (1992).
5) Elworthy, D.: Does Baum-Welch Re-estimation Help Taggers?, *Proc. ACL 4th Conference on Applied Natural Language Processing*, pp.53–58 (1994).
6) Merialdo, B.: Tagging English Text with a Probabilistic Model, *Computational Linguistics*, Vol.20, No.2, pp.155–171 (1994).
7) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proc. COLING-94*, pp.201–207 (1994).
8) Takeuchi, K. and Matsumoto, Y.: HMM Parameter Learning for Japanese Morphological Analyzer, *Proc. 10th Pacific Asia Conference Language, Information and Computation*, pp.163–172 (1995).
9) 松本裕治ほか：日本語形態素解析システム JUMAN 使用説明書 version 2.0, Technical Report, 奈良先端科学技術大学院大学（松本研究室），NAIST-IS-TR94025 (1994).
10) 竹内孔一：隠れマルコフモデルによる日本語形態素解析システムのパラメータ推定，修士論文，奈良先端科学技術大学院大学，NAIST-IS-MT9451067 (1995).
11) 日本電子化辞書研究所：EDR 電子化辞書使用説明書（第 2 版）(1995).

付　録

A.1 使用した品詞

- 括弧開，括弧閉，記号，空白，句点，読点
- 普通名詞，固有名詞，形式名詞，数詞，サ変名詞，地名，人名，副詞的名詞，時相名詞
- 名詞形態指示詞，連体詞形態指示詞，副詞形態指

示詞
- 動詞，形容詞，接続詞，感動詞，判定詞，連体詞，助動詞
- 副詞，発言副詞，様態副詞，程度副詞，量副詞，頻度副詞，時制相副詞，陳述副詞，評価副詞
- 格助詞，副助詞，名詞接続助詞，述語接続助詞，引用助詞，終助詞
- 名詞接頭辞，動詞接頭辞，イ形容詞接頭辞，ナ形容詞接頭辞
- 名詞性名詞接尾辞，名詞性名詞助数辞，名詞性述語接尾辞，動詞性接尾辞，形容詞性述語接尾辞，形容詞性名詞接尾辞

A.2 HMMの状態に追加した単語
- 読点“、”，句点“。”，記号“・”
- 括弧開“「”，“（”，括弧閉“」”，“）”
- 動詞“する”，“なる”，“ある”，“いう”
- 判定詞“だ”
- 形式名詞“こと”
- 動詞性接尾辞“いる”，“れる”
- 形容詞性述語接尾辞“ない”
- 名詞性名詞接尾辞“など”
- 名詞性名詞助数辞“円”，“年”，“%”

(平成8年8月22日受付)
(平成9年1月10日採録)

竹内 孔一（学生会員）

昭和43年生．平成7年奈良先端科学技術大学院大学情報科学研究科情報処理学専攻博士前期課程修了．同年同大学院博士後期課程入学．統計的手法に基づく自然言語処理に関する研究に従事．電子情報通信学会会員．

松本 裕治（正会員）

昭和30年生．昭和52年京都大学工学部情報工学科卒．昭和54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．昭和59～60年英国インペリアルカレッジ客員研究員．昭和60～62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，平成5年より奈良先端科学技術大学院大学教授，現在に至る．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．