自然言語処理 123－1
（1998．1．19）

枝分かれ構造をもつ同時確率モデルによる形態素解析

藤本浩司，乾伸雄，小谷善行
東京農工大学工学部電子情報工学科
184 東京都小金井市中町2-24-16
{kfujim,nobu,kotani}@cc.tuat.ac.jp

統計手法を用いた形態素解析には，隠れマルコフモデルを連接確率モデルとして利用しているものが多い．しかし，品詞情報をそのままモデルの状態と見なすと必ずしも期待する確率予測値が得られないため，経験的にいくつかの形態素や分類グループをあらたな一状態として定義し，交互作用を拾う処置がなされる．

本報告では，品詞の枝分かれ構造を利用した連接確率モデルを提案する．1つの連接確率は，連接する符号の属するそれぞれの上位水準間の同時確率を要素とする束の形で表現される．このモデル（枝分かれ同時確率モデル）は，経験則を利用することなく，学習データから連接確率推定に最適な状態を選ぶことができる．日本語コーパス（RWCコーパス）を用いて，枝分かれ同時確率モデルの形態素解析評価実験を行い，隠れマルコフモデルより有意に高い精度が得られた．

Keyword：束　形態素解析　枝分かれ構造

Nested Joint Probability Model for Morphological Analysis

Koji FUJIMOTO, Nobuo INUI, Yoshiyuki KOTANI
Department of Computer Science
Tokyo University of Agriculture and Technology
2-24-16 Nakamachi, Koganei-city, Tokyo 184, Japan
{kfujim,nobu,kotani}@cc.tuat.ac.jp

Though hidden Markov model has been employed for morphological analysis, it is necessary to optimize the hidden state level among some tag categories heuristically in order to improve the accuracy. This paper presents a novel stochastic model based on a nested tag class structure. The Nested Joint Probability Model estimates morphemes transition probability by using the frequency lattice which is expressed by a direct product of two nested tag class hierarchies. For each pair of morphemes, some important frequency subset among lattice is selected and an optimized estimation formula is constructed automatically. The model effectiveness was tested, compared with a number of delivation models based on the hidden Markov model and some Markov Models on Japanese corpus (RWC corpus). The results showed that the Nested Joint Probability Model won the most accurate in open data and the highest noise robustness.

Keyword : lattice, morphological analysis, nested structure

1．はじめに

近年，形態素解析は，研究者の言語理論や経験的知識から作られた規則に基づく方法から，学習コーパスの統計情報を用いた確率モデルに基づく手法に移行しつつある．なかでも隠れマルコフモデル（Hidden Markov Model）[1),2),3),4),7)]を基本とした連接確率モデルを利用し，最尤な形態素候補列として最適解を得る多くの研究が報告されている．この方法は，学習コーパスと形態素辞書を用意すれば，なんら解析規則を与えることなしに，分かち書きとタグ付けの解が同時に得られる利点があるが，隠れ状態の設定方法に次にあげる問題がある．

・隠れ状態に割り当てる品詞・組成水準を粗くすると精密な連接確率が求めにくくなり，逆に詳細にすると遷移行列がスパースになる．

・各形態素の隠れ状態に割り当てる最適水準が，形態素によって，また連接する形態素の組によって一意に定められない．

前者の問題に対して森らは，2つの異なる水準のbi-gram（品詞bi-gramと形態素bi-gram）を用意し，重ね合わせる手法[13),14)]を提案している．また，川端，江原，白井らは，D-bigramや(n-1)-gramなど上位水準の確率情報からスパースなn-gramのスムージングを行っている（最大エントロピー法[8),9),10),16)]，二項分布に基づくback-off平滑法[15)]）．

後者については，竹内らは，連接する前後の品詞種別，活用形種別によって，隠れ状態に対応させる水準を上下させる経験則を使っている[10),11)]．例えば，原則として隠れ状態には品詞細分類を与えるが，先行品詞が助詞であれば，例外的に表層語を状態に割り当てる．春野らの研究では，n-gram の長さ nを最適な長さで選択する可変長隠れマルコフモデル[12)]が提案されている．Kempeの方法[6)]は，あらかじめ帰納的決定木アルゴリズムによって選別された上位水準の組の確率情報から，連接確率を推定するものである．

本報告で提案する枝分かれ構造をもつ同時確率モデル（枝分かれ同時確率モデル）は，枝分かれ分類をもつ事象間の同時確率が束（ラティス）構造をつくることに注目し，学習頻度情報から適切な推定式をラティスから導き出すものである．このモデルにより，連接する形態素対において，連接形態素の属する上位水準の同時出現頻度情報から有意に意味のある交互作用を取り出し，自動的に最適な形態素連接確率を推定することができる．この手法では，隠れ状態を設定する必要がなく，「状態遷移×出現確率」といった隠れマルコフモデル型の推定より強力な推定式が得られる．以降，2節にて枝分かれ同時確率モデルを定義し，3節で，枝分かれ同時確率モデルでの形態素連接確率の推定について述べる．4節では，隠れマルコフモデルなどとの比較実験結果を示す．

2 枝分かれ同時確率モデル

この節では，連接する2形態素の同時確率をあらわす枝分かれ同時確率モデルを提案する．まず，枝分かれ構造をもつ2事象に対し，その同時事象の分解表現式と度数ラティスを定義する．ひとつの同時事象に対し，分解表現式は複数定義されるが，その中から最適な分解表現式を選ぶ手順を与えることにより，枝分かれ同時確率モデルが定まる．

2.1 分解表現式と度数ラティス

今，事象A が枝分かれ構造をもつとは，A のとりうる属性値が，Ac_1, $c_1=1,\ldots,n_1$ の水準で分類されており，さらにおのおのの水準A_{c1} が，その下位層の水準Ac_1c_2, $c_2=1,\ldots,n_2$ によって階層的にくりかえし細分類されていることとする．2つの事象 A ,B が，それぞれ多層からなる枝分かれ構造をもつとし，A の第 i 層の水準$Ac_1c_2..c_i$ をあらためてA_i と記し，そのひとつ上位層，すなわち第 i-1 層のA_i の属する水準$Ac_1c_2..c_{i-1}$ をあらためてA_{i-1} とする．同様にB の第 j 層の水準を B_j , 第 j-1 層のB_j の属する水準を B_{j-1} とする． 水準 A_i, B_j の交互作用 $I_{A_iB_j}$ を

$$I_{A_iB_j}=\frac{P(A_iB_j\mid A_{i-1}B_{j-1})}{P(A_i\mid A_{i-1}B_{j-1})P(B_j\mid A_{i-1}B_{j-1})} \qquad (1)$$

と定める．ここで，P(X|Y) は事象Yが生起した場合の事象Xの条件付き生起確率である．交互作用 $I_{A_iB_j}$ は，水準A_i と水準B_j の最小上限水準$A_{i-1}\wedge B_{j-1}$ が生起した条件のもとでの，A_i と B_j が同時に生起する確率の独立性を示す．すなわち，$A_{i-1}\wedge B_{j-1}$ が生起したという前提において，A_i と B_j が独立に生起するならば，$I_{A_iB_j}$ は値1をとり，共起しやすければ1より大きく，共起しにくければ1より小さい値をとる．最小値0をとる場合は，同時に生起することがあり得ないことを意味する．

次に，この交互作用を使って最下位の同時確率を記述する．事象 A, B の層数をそれぞれ，m, n とし，ある観測値の属する各層の水準を，A_i $i=1,..,m$, B_j $j=1,\ldots,n$ とする．

水準A_m, B_n の同時確率 $P(A_mB_n)$ は，(1) より，

$$P(A_mB_n)=\frac{P(A_mB_{n-1})P(A_{m-1}B_n)}{P(A_{m-1}B_{n-1})}I_{A_mB_n} \tag{2}$$

とあらわされる．さらに右辺にあらわれる確率に，順次(1)を適用することにより，

$$\begin{aligned}P(A_mB_n) &= \frac{P(A_mB_{n-1})P(A_{m-2}B_n)}{P(A_{m-2}B_{n-1})}I_{A_{m-1}B_n}I_{A_mB_n}\\ &= \frac{P(A_mB_{n-2})P(A_{m-1}B_{n-1})P(A_{m-2}B_n)}{P(A_{m-1}B_{n-2})P(A_{m-2}B_{n-1})}\\ &\quad \cdot I_{A_mB_{n-1}}I_{A_{m-1}B_n}I_{A_mB_n}\\ &= \frac{P(A_mB_{n-2})P(A_{m-2}B_n)}{P(A_{m-2}B_{n-2})}\\ &\quad \cdot I_{A_{m-1}B_{n-1}}I_{A_mB_{n-1}}I_{A_{m-1}B_n}I_{A_mB_n}\\ &\vdots\\ &= P(A_m)P(B_n)\prod_{\substack{i=1,..,m\\ j=1,..,n}}I_{A_iB_j}\end{aligned} \tag{3}$$

と展開される．このようにして得られる交互作用と上位水準同時確率の積で表現される右辺の式を，$P(A_mB_n)$ の枝分かれ同時確率の分解表現式と呼ぶことにする．分解表現式は，任意の上位水準同時事象の集合 S に対し，式(4)で一意に与えられる．

$$P(A_mB_n)=\frac{\prod_{A_iB_j\in S_\vee}P(A_iB_j)}{\prod_{A_iB_j\in S^\wedge}P(A_iB_j)}\prod_{A_iB_j\in I}I_{A_iB_j} \tag{4}$$

where

$$S\subseteq\Sigma=\{A_iB_j\mid i=0,..,m;j=0,..,n\}$$

$$A_iB_j\le A_{i'}B_{j'}\Leftrightarrow i\ge i'\wedge j\ge j'$$

$$\min S=\left\{x\in S\mid{}^\forall s\in S,\neg(s<x)\right\}$$

$$S_\vee=\min\left(S\cup\{A_mB_0,A_0B_n\}\right)$$

$$S^\wedge=\min\left\{x\in\Sigma\mid{}^\exists s,s'\in S_\vee,s\ne s'\wedge s\le x\wedge s'\le x\right\}$$

$$I=\left\{x\in\Sigma\mid{}^\forall s\in S_\vee,\neg(s\le x)\right\}$$

また分解表現式は，図1に示すように束（ラティス）の上であらわすことができる．図1のラティスは，各格子点が対応する水準を示しており，上下が包含関係の順序を示している．ここでは，同時確率 $P(A_mB_n)$ は，太線で結ばれた4水準の確率と四角でマークされている5個の交互作用から構成できることを示しており，式(5)の分解表現式に相当する．

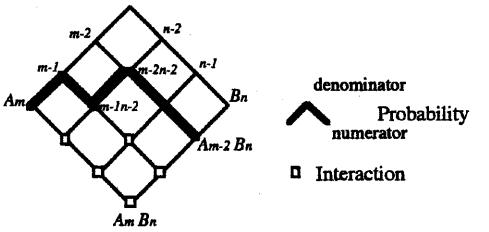

図1　枝分かれ効果をあらわすラティス

$$P(A_mB_n)=\frac{P(A_m)P(A_{m-1}B_{n-2})P(A_{m-2}B_n)}{P(A_{m-1})P(A_{m-2}B_{n-2})}\cdot I_{A_mB_{n-2}}I_{A_mB_{n-1}}I_{A_{m-1}B_{n-1}}I_{A_{m-1}B_n}I_{A_mB_n} \tag{5}$$

従って，もしいくつかの交互作用が無視できる，すなわち1と見なすことができるなら，$P(A_mB_n)$ は，それによって決まるいくつかの上位水準の生起確率より推定できるといえる．具体的には，無視する交互作用すべてを上部から被覆する格子結線を選び，その谷格子にあたる確率を分子に，峰格子にあたる確率を分母に与えればよい．

ここで，学習データから各水準A_iB_j ($i=0,1,...m$; $j=0,1,...,n$) の度数F_{ij} が得られているとする．なお，F_{00}, F_{i0}, F_{0j} はそれぞれ，全観測度数，水準A_i，水準B_j の観測度数をあらわす．与えられた観測値に対する $P(A_iB_j)$ の最尤推定量は，F_{ij}/F_{oo} であるから，$P(A_mB_n)$ の最尤推定量も度数F_{ij} を使ってあらわされる．これは分解表現式 (4) に完全に対応しており，$P(A_iB_j)$ の代わりにF_{ij} を与え，交互作用を1とすれば良い．図1のラティスについても，各格子の示す水準A_iB_j に度数F_{ij} を与えれば，度数で表現された最尤推定量の分解表現式に自然に対応づけられる．これを以後，度数ラティスと呼ぶ．

2.2　最適な分解表現式の選択

最下位の同時確率$P(A_mB_n)$ の最尤推定量は，1と見なせる上位水準の交互作用が決まれば，最適な分解表現式により決定することを述べた．次にどれらの交互作用が1と見なせるかを決定しなければならない．これは度数ラティスに対して，2元分割表の独立性の検定を逐次行うことにより判断することができる．独立性の検定には，χ^2 検定や尤度比検定，有意水準を用いないAIC基準による方法などがあるが，例えば χ^2

検定をもちいるなら，交互作用$I_{A_iB_j}=1$ なる帰無仮説に対する χ^2 値は，

$$\chi^2=\frac{F_{i-1j-1}(F_{ij}F_{\bar{i}\bar{j}}-F_{i\bar{j}}F_{\bar{i}j})^2}{F_{i-1j}F_{i-1\bar{j}}F_{\bar{i}j-1}F_{ij-1}}$$

$$\text{where}\quad F_{i\bar{j}}=F_{ij-1}-F_{ij}$$
$$F_{\bar{i}j}=F_{i-1j}-F_{ij}$$
$$F_{\bar{i}\bar{j}}=F_{i-1j-1}-F_{i-1j}-F_{ij-1}-F_{ij} \qquad (6)$$

とあらわされる．これは自由度1のχ^2 分布となることから，あらかじめ定められた危険率で有意かどうかが判定される．ただし，期待度数があまりに小さいと適用が困難になるため，フィシャーの直接法など別の方法をとるか，あえて棄却しないことが望ましい．

現在実装しているプログラムでは，危険率 0.1%の χ^2 検定を採用している．この小さい危険率は，「無意味に小さいが有意な差」に対する有意判定を避けるためと，多重比較による誤まりを小さくするためである．また観測度数 0 に対しては，期待度数の95%上側信頼限界が3[1]であることから，それにもとづき度数3が得られているものとした保守的な判定を行っている．さらに期待値が小さい場合は検定を行わず，交互作用がなかったものと同じ扱いをしている．

最適な分解表現式を構成する水準は，次の手順により，確定リストとして得られる．得られる確定リストは式(4) における同時水準の部分集合 S である．

Step-1　候補リスト，確定リストを空に初期化する．
Step-2　候補リストに最下位水準 A_mB_n を与える．
Step-3　候補リストからひとつの水準を抜きだし，交互作用の有無を検定する．
Step-4a　帰無仮説「交互作用無し」が棄却されたら，その水準を確定リストに加える．
Step-4b　棄却されなければ，その水準 A_iB_j のひとつ上位の水準 A_iB_{j+1}，$A_{i+1}B_j$ を候補リストに加える．
Step-5　候補リストが空であれば終了．そうでなければ，Step-3から繰り返す．

---

[1] Poisson 分布から，

$$\alpha=\sum_{x<0}e^{-upper\lambda}\frac{upper\lambda^x}{x!}\le\sum_{x=0}e^{-upper\lambda}\frac{upper\lambda^x}{x!}$$

を解いて求められる．(α は危険率，$upper\lambda$ は上側信頼限界である)

以上から，枝分かれ同時確率モデルは，与えられた度数ラティスに統計的検定を施して得られる最適な分解表現式を使って，最下位水準の同時確率を推定するモデルであるといえる．

3 枝分かれ同時確率モデルによる形態素連接確率

形態素の品詞分類は一般に，動詞・名詞などの大分類，格助詞・接続助詞などの第2分類，さらに詳細な分類というように階層構造であらわされる．末端の分類の下にまとめられるおのおのの形態素は，表記が同じであっても，文法上なんら共通する性質は仮定できないことから，形態素自身もこの品詞分類の最下位の階層（葉）として扱う．これらの分類は，文法規則の類似性を反映したものであるから，2つの連続する形態素の連接確率も枝分かれ同時確率モデルの上に記述することができる．

先行する形態素 A_3 を第3階層（末端階層）の水準とし，その上位階層に第2階層として品詞第2分類 A_2 を，第1階層として品詞第1分類 A_1 を定める．同様にして，対象形態素とその上位水準を B_3，B_2，B_1 とする．

	先行符号	対象符号	例	
形態素	A_3	→ B_3	日本	→ は
品詞第1分類	A_1	B_1	名詞	助詞
品詞第2分類	A_2	B_2	固有	格助詞

学習データにおける水準A_iB_jの同時観測度数を F_{ij} であらわすと，連接確率の推定値 $\hat{P}(B_3|A_3)=\hat{P}(A_2B_3)/\hat{P}(A_3)$ は，(3) より

$$\hat{P}(B_3|A_3)=\begin{cases}\dfrac{F_{32}F_{23}}{F_{30}F_{22}} & \text{if } I_{A_3B_3}=1\\ \dfrac{F_{32}F_{13}}{F_{30}F_{12}} & \text{if } I_{A_2B_3}=I_{A_3B_3}=1\\ \dfrac{F_{31}F_{22}F_{13}}{F_{30}F_{21}F_{12}} & \text{if } I_{A_3B_2}=I_{A_2B_3}=I_{A_3B_3}=1\\ \dfrac{F_{31}F_{13}}{F_{30}F_{11}} & \text{if } I_{A_3B_2}=I_{A_2B_3}=I_{A_2B_3}\\ & \quad =I_{A_3B_3}=1\\ \vdots & \\ \dfrac{F_{03}}{F_{00}} & \text{if all Interactions}=1\end{cases} \qquad (7)$$

などと，無視する交互作用に応じて計算できる．

このように，枝分かれ同時確率モデルは，多層からなる階層型品詞分類・素性分類をもつ２形態素連接確率の推定に応用できる．また，先行するｎ個の形態素の品詞列を，ｎ階層の枝分かれ水準としてモデル化することによって，可変長 n-gram にも利用できる．

次に，枝分かれ同時確率モデルによる推定確率計算例を示す．

例　先行形態素　動詞・サ変スル・「し」
　　対象形態素　助詞・格助詞・「の」
　の連接確率 P(の[助詞・格]| し[動詞・サ])
　を推定する．

この例は，「動詞・サ変スル」に属する形態素「し」に「格助詞」の「の」が連接するパスを示しており，実際は，文法上連接しない，連接確率０となるべきものである．対する度数ラティスを図２に与える．

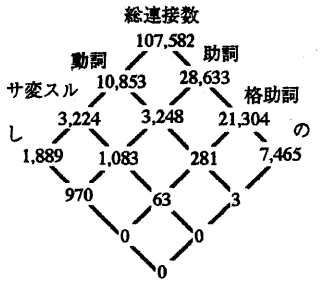

図２　「し」と「の」の連接の度数ラティス

度数ラティスの{し->の}自身の連接度数が０であることから，確かに約１１万件の学習データ中，この２形態素の連接は一度も観測されなかったことがわかる．品詞第２区分遷移を隠れ状態遷移とする隠れマルコフモデルの場合，連接確率の推定値は，

　P(の[助詞・格]| し[動詞・サ])
　＝P(の| [助詞・格]) P([助詞・格]|[動詞・サ])
　＝7,465/21,304　×　63/3,224 ＝0.68 %

となる．

枝分かれ同時確率モデルの場合，まず，検定を行わずに適当な分解表現式で推定する場合を考える．できるだけ，無視する交互作用は少ない方が望ましいだろうから，右下の度数を使うことになる．しかし{ [サ変スル]->の}，{ し->[格助詞] }，{ の->し} の３個の連接の度数は０となっており，直感的にこれらを除くことにして，０を囲む上位の度数を使って推定すると，

　P(の[助詞・格]| し[動詞・サ])
　＝3/281　×　63/1,083　×　970/1,889 ＝0.0319%

と，隠れマルコフモデルより約20分の1，確率が小さくなる．

次に枝分かれ同時確率モデルの定義にしたがい，検定結果から有意と判定された度数のみを使って再度推定してみる．先ほどの３つの０度数についてだが，{ し->[格助詞] } は，度数が有意に小さいが，これ以外の連接度数０は，有意ではない．よって{ し->[格助詞] }の度数に３を代用し，他の有意な度数を使って，再計算すると

　P(の[助詞・格]| し[動詞・サ])
　＝3/281　×　(3)/1,889 ＝0.0017 %

と十分，確率０に近づけることができる．(式中の(3)は，真の度数を推定度数３で代用していることをあらわしている．)

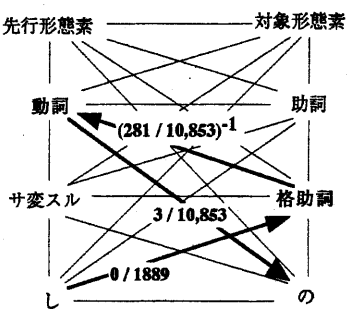

図３　「し」と「の」の連接推定値

この推定は，図３に見るように，{[動詞]->の}の連接確率と，{し->[格助詞]}の連接確率を両方利用するため，さらに{[動詞]->[格助詞]}の連接確率を割ることで調整していることがわかる．

このように，現在の枝分かれ同時確率モデルは，完全に連接確率を０とすることはないが，利用できる情報をすべて取り込むことで，より本来の値を推定していると考えられる．

4　評価実験

上述のアルゴリズムで枝分かれ同時確率モデルを実装し，日本語コーパス（ＲＷＣコーパス）を使って，クローズドデータおよびオープンデータにおける解析精度の評価を行った．

4.1 評価実験方法

行う形態素解析処理は，解析対象の平文を辞書引きして得られる形態素ラティスの各々の連接確率を確率モデルにより算出し，ビテルビアルゴリズム[5]によって最大確率をもつパスを見つけ解析結果とするものである．連接確率を求めるモデルとして，枝分かれ同時確率モデル (NEST) には，品詞第１区分，品詞第２区分に形態素自体を最下層に加えた３層の枝分かれ構造を採用した．比較対照には，

・品詞第１区分を隠れ状態とする隠れマルコフモデル (HMM1)
・品詞第２区分を隠れ状態とする隠れマルコフモデル (HMM2)
・形態素自体の遷移確率をあらわすマルコフモデル (MM3)
・品詞第２区分を隠れ状態とする隠れマルコフモデルと形態素遷移マルコフモデルおよび，品詞第２区分から形態素，形態素から品詞第２区分へ遷移するマルコフモデルの４つを重ね合わせたモデル[13),14)] (SP)

の４モデルを用意した．

HMM1 および HMM2 モデルは，学習データから，品詞遷移確率および形態素出現確率を算出し，形態素ラティスから最大確率のパスを求める．HMM1では，隠れ状態とする品詞として第１品詞区分を，HMM2では，第2品詞区分を割り当てている．

MM3モデルは，学習データから得られた形態素遷移確率をそのまま形態素連接確率とするものであり，学習データに現れなかった形態素連接は，接続可能性がないと見なされる．

SPモデルは，学習データから得られた形態素間遷移，形態素品詞遷移，品詞形態素遷移，品詞間遷移をそれぞれ，100,10,10,1 の割合で重みづけして保持している．この重みづけの効果により，形態素間遷移が受理されれば，ほぼそれが優先的に採用され，形態素間遷移が受理されない場合は，形態素品詞遷移がそれに代わって採用される．すなわちSPモデルは，その重みづけをうまく利用することで連接情報のない形態素遷移を品詞遷移が橋渡ししていくモデルである．このモデルは，森らの研究[13),14)]を参考にしているが，文献では品詞形態素遷移，形態素品詞遷移ではなく，付属語のみ語彙化した遷移（付属語形態素と自立語品詞の合併集合間遷移），自立語のみ語彙化した遷移（付属語品詞と自立語形態素の合併集合間遷移）を使っている点が異なる．

用いたデータは，遷移確率学習用，解析精度評価用ともに，RWCコーパスの通商白書を用いた．

・学習用コーパス
　RWCコーパス　通商白書Ｈ４（約１０万形態素）
・評価用コーパス
　RWCコーパス　通商白書Ｈ４から182文
　（クローズド　6,916形態素）
　RWCコーパス　通商白書Ｈ５から238文
　（オープン　9,015形態素）

形態素辞書は学習用および評価用コーパスから全単語を抽出し未知語はないものとした．ただし，出現確率は，あくまで学習用コーパスのみから得られるものとするため，評価用コーパスに含まれる形態素は，０度数としている．

評価基準には，推定された形態素の第１・２品詞分類がコーパスのそれに合致するものを正解として，形態素単位の再現率と適合率を用いた．すなわち，

再現率 = コーパスに一致したモデルの推定形態素数 / コーパスの全形態素数
適合率 = コーパスに一致したモデルの推定形態素数 / モデルの全推定形態素数

である．

4.2 評価実験結果と考察

実験結果として，各モデルの学習用コーパス（クローズドデータ），評価用データ（オープンデータ）に対する再現率 (%Recall) と適合率 (%Precision) およびそれぞれの９５％信頼限界 (LB95-UB95) を表１に示す．

クローズドデータでは，再現率，適合率ともにSPモデルがもっとも良い精度を示した(98.9%,99.3%)．次に，MM3 モデル，NESTモデルの順であった．一方オープンデータにおいては，NESTモデルがもっとも良

く(94.6%,95.8%)，次に良い SPモデル (94.0%,94.5%) に対し９５％信頼で有意に精度が高い．クローズドデータに対するオープンデータの精度低下を見ると，明らかにMM3モデルが最も大きい (20-30point) が，SPモデルにおける 5 point 低下も，NESTモデルやHMM2モデルの 2.5 point 前後の低下に比べ十分大きい．これは，NESTモデルやHMM2モデルにくらべ，MM3モデルやSPモデルが，ノイズを拾って学習データにオーバーフィッティングしたためと考えられる．

この結果から，枝分かれ同時確率モデルが，最も高い精度を有意に示し，学習データのノイズに対してもロバストなモデルであることがわかった．

表1　枝分かれ同時確率モデル評価実験結果

MODEL	Closed Data (#Morph=6,916) %Recall LB95 - UB95	%Precision LB95 - UB95	Open Data (#Morph=9,015) %Recall LB95 - UB95	%Precision LB95 - UB95	%Point Decrease (Closed-Open) %Rec.	%Pre.
HMM1	90.8	92.4	87.4	89.5	3.4	2.8
	90.1 - 91.5	91.7 - 93.0	86.7 - 88.1	88.9 - 90.2		
HMM2	94.9	96.3	92.0	94.0	2.9	2.3
	94.4 - 95.4	95.9 - 96.8	91.4 - 92.5	93.6 - 94.5		
MM3	97.2	97.4	68.2	75.0	29.0	22.4
	96.8 - 97.6	97.0 - 97.8	67.2 - 69.2	74.1 - 75.9		
SP	98.9	99.3	94.0	94.5	4.8	4.8
	98.6 - 99.1	99.1 - 99.5	93.5 - 94.5	94.1 - 95.0		
NEST	96.9	97.7	94.6	95.8	2.4	2.0
	96.5 - 97.3	97.4 - 98.1	94.1 - 95.0	95.3 - 96.2		

表2　枝分かれ同時確率モデルの選択した分解表現式

分解表現式	1と見なされた交互作用 (AB_i をij と記す)	全形態素連接候補数	推定連接数	誤答推定連接数	平均を100とした誤答率のINDEX
	sample size	1,297	326	174	
Total		100%	100%	100%	100
$F_{31}/F_{30}\cdot F_{22}/F_{21}\cdot F_{03}/F_{02}$	13,23,32,33	4%	4%	0%	0
$F_{32}/F_{30}\cdot F_{03}/F_{02}$	13,23,33	10%	20%	3%	14
$F_{31}/F_{30}\cdot F_{12}/F_{11}\cdot F_{03}/F_{02}$	13,22,23,32,33	4%	4%	1%	16
$F_{31}/F_{30}\cdot F_{13}/F_{11}$	22,23,32,33	2%	5%	1%	23
$F_{32}/F_{30}\cdot F_{23}/F_{22}$	33	2%	5%	2%	47
F_{33}/F_{30} =MM3	(none)	6%	21%	10%	48
$F_{31}/F_{30}\cdot F_{23}/F_{21}$	32,33	3%	5%	5%	105
$F_{22}/F_{20}\cdot F_{03}/F_{02}$	13,23,31,32,33	12%	4%	5%	115
$F_{32}/F_{30}\cdot F_{13}/F_{12}$	23,33	2%	6%	7%	116
$F_{21}/F_{20}\cdot F_{13}/F_{11}$	22,23,31,32,33	9%	7%	10%	153
F_{23}/F_{20}	31,32,33	9%	10%	29%	276
$F_{22}/F_{20}\cdot F_{13}/F_{12}$ =HMM2	23,31,32,33	4%	3%	3%	112
$F_{21}/F_{20}\cdot F_{12}/F_{11}\cdot F_{03}/F_{02}$	13,22,23,31,32,33	17%	2%	7%	304
$F_{31}/F_{30}\cdot F_{03}/F_{01}$	12,13,22,23,32,33	3%	1%	0%	0
$F_{11}/F_{10}\cdot F_{03}/F_{01}$ =HMM1	12,13,21,22,23,31,32,33	4%	1%	3%	468
$F_{21}/F_{20}\cdot F_{03}/F_{01}$	12,13,21,22,23,31,32,33	3%	1%	3%	562
$F_{12}/F_{10}\cdot F_{03}/F_{02}$	13,21,22,23,31,32,33	3%	1%	3%	562
F_{13}/F_{10}	21,22,23,31,32,33	2%	0%	4%	UNK
$F_{31}/F_{30}\cdot F_{22}/F_{21}\cdot F_{13}/F_{12}$	23,32,33	1%	0%	1%	UNK
F_{03}/F_{00}	11,12,13,21,22,23,31,32,33	0%	0%	1%	UNK

表2は，実際にどのような分解表現が推定に使われたかを，解析途中の一部から無作為抽出して調べたものである．集計対象は，1) 辞書引き直後の形態素ラティスを構成するすべての連接，2) 最適連接鎖として選ばれた連接，3) 推定誤りをおかした連接の３つで，それぞれの場合の分解表現の構成比を示した．表の上半分は，ある推定連接での構成比が4%以上の分解表現式を対象に，誤答INDEXで昇順にソートしてある．下半分は，サンプル数が十分でないので，参考までに推定連接構成比順で列記した．

この表から枝分かれ同時確率モデルでは，決してHMM2やHMM1と同等な分解表現式を多く選択しているわけではなく，頻度に多少の差はあるものの，種々の分解表現式が使われていることがわかる．また，強い傾向ではないが，１と見なす交互作用が５個以上となる分解表現式は多くは採用されず，誤りも多いようである．

5 まとめ

本報告では，形態素連接確率を推定する新たな手法として枝分かれ同時確率モデルを提案した．枝分かれ同時確率モデルは，多層の階層分類をもつ枝分かれ水準の同時事象により構成される束（ラティス）から，最適な同時確率推定式を自動的に選ぶことができる．また，このモデルを使うことにより，形態素連接推定に，1) 詳細だがスパースな状態遷移行列を利用せずとも，適切な推定が可能になること，2) 形態素種別により隠れ状態を調整する必要がないことを示した．

日本語コーパス（ＲＷＣコーパス）を使った評価実験において，枝分かれ同時確率モデルは，隠れマルコフモデル，形態素bi-gram ，bi-gram 重ね合わせモデルより有意に高い正解率（再現率および適合率にて評価）を得た．特に，学習データの正解率に対する評価データの正解率の低下においても，最も小さいという結果から，ノイズによるオーバーフィッティングを回避する能力においても高いことが示された．

参考文献

1) P.F.Brown, V.J.D.Pietra, P.V.deSouza, J.C. Lai, R.L.Mercer, Class-Based n-gram Models of Natural Language., Computational Linguistics Vol.18, No.4, pp.467-479, (1992)

2) K.W.Church, A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, Proceedings of 2nd Conference on Applied Natural Language Processing(ANLP-88), pp.136-143, (1988)

3) K.W.Church, R.L.Mercer : Introduction to the Special Issue on Computational Linguistics Using Large Corpora, Computational Linguistics, Vol.19, No.1,pp.1-24, (1993)

4) S.J.DeRose : Grammatical Category Disambiguation by Statistical Optimization, Computational Linguistics, Vol.14, No.1, pp.31-39, (1988)

5) G.D.Forney,Jr. :The Viterbi Algorithm, Proceedings of IEEE,Vol.61 No.3 pp.268-278, (1973)

6) A.Kempe : Probabilistic Tagging with Feature Structure, COLING-94 pp.161-165 (1994)

7) B.Merialdo, Tagging English Text with a Probabilistic Model, Computational Linguistics Vol.20, No.2, pp.155-171, (1994)

8) A.Ratnaparkhi,A maximum entropy model for part-of-speech tagging, proceeding of the Empirical Methods in Natural Language Processing Conference,(1996)

9) 白井清昭, 乾健太郎, 徳永健伸, 田中穂積 : 最大エントロピー法を用いた単語bigramの推定, 情報処理学会研究会報告, 自然言語研究会 No.NL-116-4 pp.21-28, (1996)

10) 竹内孔一, 松本裕治 : HMMによる日本語形態素解析システムのパラメータ学習, 情報処理学会研究会報告, 自然言語研究会 No.NL-108-3 pp.13-19, (1995)

11) 竹内孔一, 松本裕治: 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, 情報処理学会論文誌, Vol.38,No.13,pp.500-509 (1997)

12) 春野雅彦, 松本裕治 : 文脈木を利用した形態素解析, 情報処理学会研究会報告, 自然言語研究会 No.NL-112-5 pp.31-36, (1996)

13) 森信介, 長尾眞 :形態素bi-gramと品詞bi-gramの重ね合わせによる形態素解析, 情報処理学会研究会報告, 自然言語研究会 No.NL-112-6 pp.37-44, (1996)

14) 森信介, 長尾眞 :語彙化マルコフモデルによる英語品詞タグ付け, 信学技報 No.NLC95-78(1996-03) pp.31-38, (1996)

15) 川端豪, 田中信詞 :二項分布に基づくN-gram言語モデルのBack-off平滑化, 信学技報 No.NLC95-58(1995-12) pp.1-6, (1995)

16) 江原暉将 :最大エントロピー法を用いてバイグラム確率からｎグラム確率を求める, 情報処理学会研究会報告, 自然言語研究会 No.NL-113-5 pp.25-30, (1996)