クラスに基づく可変長記憶マルコフモデル

森 信介[†]

本論文では，クラスに基づく可変長記憶マルコフモデルとその学習アルゴリズムについて述べる．このモデルは，可変長記憶マルコフモデルの拡張であり，クラスに基づく確率的接尾辞木に基づいている．木の各ノードは，自動クラスタリングによって得られた単語クラスを持っている．実験では，我々が提案するクラスに基づく可変長記憶マルコフモデルと，単語 2-gram モデル，クラス 2-gram モデル，単語 3-gram モデル，可変長記憶マルコフモデルとを比較した．また，これらのモデルに基づく英語の品詞タガーと日本語の形態素解析器の精度を比較した．実験の結果，クラス 2-gram モデルとのモデル記述の記憶領域の比較を除いて，クラスに基づく可変長記憶マルコフモデルが，クロスエントロピーとモデル記述の記憶領域と解析精度において，他のモデルよりも優れていた．

Class-based Variable Memory Length Markov Model

Shinsuke Mori[†]

In this paper, we present a class-based variable memory length Markov model and its learning algorithm. This is an extension of variable memory length Markov model. Our model is based on class-based probabilistic suffix tree, whose nodes have an automatically acquired word-class relation. We experimentally compared our new model with word-based bi-gram model, word-based tri-gram model, class-based bi-gram model and word-based variable memory length Markov model. We also built part-of-speech taggers based on these models and compared their accuracies. The results show that a class-based variable memory length Markov model is better in cross entropy, tagging accuracy and model size than the other models mentioned above except for the comparison with the class-based bi-gram model in size.

1. はじめに

ある言語の文字列の出現確率を記述する確率的言語モデルは，音声言語を含めた自然言語処理のための言語モデルとして重要な位置を占めている．応用例としては，音声認識[1)]をはじめとして，文字認識[2)]，仮名漢字変換[3)]，中国語の入力[4)]，英語の品詞タグ付け[5)]，日本語の形態素解析[6)]，機械翻訳[7)]などがある．代表的な確率的言語モデルは，パラメータ推定や実装が容易な単語 n-gram モデル[8)]である．実際，多くの音声認識システムには，単語 n-gram モデルが利用されている．しかしながら，単語 n-gram モデルのパラメータの数は，語彙数の n 乗に等しいため，n の値が大きい場合には，現在利用可能なコーパスからパラメータを十分正確に推定することができない．結果として，十分な学習コーパスがあれば達成できるであろう予測力よりもかなり劣るモデルしか得られない．確率的言語モデルの予測力向上は，上述の様々な応用の精度を改善することが分かっており，予測力の向上を目的とした多くの改良が提案されている．クラス n-gram モデル[9)]や可変長記憶マルコフモデル[10)]などがその例である．

クラス n-gram モデルにおいては，各単語はクラスと呼ばれる単語のグループに属しており，単語列に対する統計より信頼性のあるクラス列に対する統計を通して予測される．さらに，モデルの記述に必要な記憶領域が，単語 n-gram モデルよりも小さいという利点もある．クラス n-gram モデルを構築する際の最大の問題は，単語予測という観点から最良と思われる単語のグループ分けを算出することである．この問題は，一般に単語クラスタリングによって解決される．単語クラスタリングにはいくつかの先行研究があり，自動的に獲得されたクラスに基づくクラス n-gram モデルが単語 n-gram モデルと比較して，予測力が同程度かそれ以上であり，必要記憶域が顕著に小さいと報告されている[9),11),12)]．

別の改良である可変長記憶マルコフモデル[10)]では，

† 日本アイ・ビー・エム株式会社東京基礎研究所
IBM Research, Tokyo Research Laboratory, IBM Japan, Ltd.

予測力の向上につながると期待される履歴が選択的に伸ばされる．たとえば，直前の単語が「この」であれば，可変長記憶マルコフモデルは，単語 2-gram モデルのように，その前の単語を区別せず，直前の単語が「の」であれば，単語 3-gram モデルのように，その前の単語を区別するということが可能である．さらにその前の単語も，次の単語の予測に有用な情報を保持していると見なされれば，単語 4-gram モデルのように予測に利用される．このように，予測に利用する履歴が必要に応じて伸ばされるので，ある n-gram モデルと同じ程度の記憶領域を要する可変長記憶マルコフモデルの予測力は，その n-gram モデルよりも高いことが期待される．また，ある n-gram モデルと同程度の予測力を達成するために必要な可変長記憶マルコフモデルの記憶領域は，その n-gram モデルよりも小さいことが期待される．

本論文では，可変長記憶マルコフモデルにクラスの概念を取り入れたクラスに基づく可変長記憶マルコフモデルを提案する．このモデルにより，たとえば「この日曜日の」,「この月曜日の」，…,「この土曜日の」のような類似の文脈を 1 つの文脈として扱うことが可能になり，その結果データスパースネスの問題が軽減され，より精度の高い言語モデルの構築が可能になると考えられる．クラスに基づく可変長記憶マルコフモデルの学習過程は，可変長記憶マルコフモデルとほぼ同じであるが，各ノードを展開するときに単語クラスタリングを含む点が異なる．したがって，クラスに基づく可変長記憶マルコフモデルの構築には，可変長記憶マルコフモデルの構築よりも時間がかかる．しかしながら，クラスに基づく可変長記憶マルコフモデルは，可変長記憶マルコフモデルよりも必要となる記憶領域が小さいのみならず，より予測力が高くなることが期待される．実験の結果，クラスに基づく可変長記憶マルコフモデルは，同じコーパスから学習した可変長記憶マルコフモデルや単語 3-gram モデルと比較して，クロスエントロピーと，単語分割や品詞付与の精度と，モデルの記述に必要となる記憶領域のすべてにおいて優れていた．

2. 単語に基づく確率的言語モデル

単語に基づく確率的言語モデルは，文を単語の列と見なし，文の生成確率を計算する．通常，それぞれの単語の生成確率は，先行する単語を履歴と見なして計算され，文の生成確率は，以下の式が示すように，それぞれの単語の確率の積で表される[8]．

$$P(\boldsymbol{w}) = \prod_{i=1}^{m} P(w_i|w_1 w_2 \cdots w_{i-1})$$

ここで $\boldsymbol{w} = w_1 w_2 \cdots w_m$ は，与えられた文の単語列を表す．この式中の確率 $P(w_i|w_1 w_2 \cdots w_{i-1})$ の値を正確に推定することは困難であり，様々な近似が提案されている．この章では，それらの近似の中から，単語 n-gram モデルとクラス n-gram モデルと可変長記憶マルコフモデルについて順に説明する．

2.1 単語 n-gram モデル

単語 n-gram モデルは，以下の式が示すように，直前の $k = n - 1$ 個の単語を履歴と見なし，各単語の予測に用いる．

$$P(\boldsymbol{w}) = \prod_{i=1}^{m} P(w_i|w_{i-k} w_{i-k+1} \cdots w_{i-1})$$

この式の確率 $P(w_i|w_{i-k} w_{i-k+1} \cdots w_{i-1})$ は，以下の式を用いて，学習コーパスから最尤推定される．

$$P(w_i|w_{i-k} w_{i-k+1} \cdots w_{i-1}) = \frac{N(w_{i-k} w_{i-k+1} \cdots w_i)}{N(w_{i-k} w_{i-k+1} \cdots w_{i-1})}$$

ここで $N(x)$ は，学習コーパス中での事象 x の頻度を表す．多くの場合，学習コーパスの大きさが不十分で，パラメータの推定値は必ずしも正確ではない（データスパースネス問題）．この問題に対処するため，次の式が示すような，パラメータ推定がより容易な n の値がより小さい n-gram モデルとの補間が提案されている[13]．

$$P'(w_i|w_{i-k} w_{i-k+1} \cdots w_{i-1}) = \sum_{j=1}^{k} \lambda_j P(w_i|w_{i-j} w_{i-j+1} \cdots w_{i-1}), \quad (1)$$

$$\text{where } 0 \le \lambda_j \le 1 \text{ and } \sum_{j=1}^{n} \lambda_j = 1$$

係数 λ の値は，確率値 P の推定に用いられるコーパスとは別に用意された比較的小さいコーパスを用いて最尤推定される．この方法では，確率値の推定に用いることができるコーパスの大きさが小さくなり，推定値の信頼性が少しではあるが低下するという問題がある．これに対処する方法として削除補間と呼ばれる方法がある．これは，パラメータ推定のためのコーパスを k 個に分割し，$k - 1$ 個の部分で確率値を推定し，残りの部分で補間の係数を推定するということをすべての組合せにわたって行い，その平均値をとるという方法である．

単語 n-gram モデルは，パラメータ推定が容易であ

り，自然言語の確率的性質をうまく表現することができるので，音声認識やタガーの言語モデルとして広く利用されている．それゆえに，単語 n-gram モデルを改良したモデルも多数提案されている．以下の節では，これら単語 n-gram モデルの改良のうち，クラス n-gram モデルと可変長記憶マルコフモデルについて説明する．

2.2 クラス n-gram モデル

上述のように，単語 n-gram モデルは，直前の $(n-1)$ 個の単語から次の単語を予測するのであるが，いくつかの単語は，n-gram モデルにおいては，類似した振舞いをすることが容易に想像される．そのような単語を区別することはモデルのパラメータの数を無駄に増加させるのみである．そのような単語の例としては，「日曜」,「月曜」，…,「土曜」などの曜日を指す単語があげられる．クラス n-gram モデル[9]では，各単語はクラスと呼ばれるグループに分類されており，以下の式が示すように，まず次のクラスを直前のクラス列から予測し，それから次の単語をその予測されたクラスから予測する．

$$P(\boldsymbol{w}) = \prod_{i=1}^{m} P(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1})P(w_i|c_i)$$

ここで c_i は i 番目の単語が属するクラスを表す．この式では，各単語が唯一のクラスに属することが仮定されている．この式中の確率値は，単語 n-gram モデルの場合と同様に，学習コーパスにおける頻度から推定され，補間も同様に行われる．

$$\begin{aligned} &P'(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1}) \\ &= \sum_{j=1}^{k} \lambda_j P(c_i|c_{i-j}c_{i-j+1}\cdots c_{i-1}), \qquad (2) \\ &\quad \text{where } 0 \le \lambda_j \le 1 \text{ and } \sum_{j=1}^{n} \lambda_j = 1 \end{aligned}$$

クラス n-gram モデルにおける最大の問題は，単語予測という観点から最良の単語とクラスの関係を自動的に見つけることである．この問題は，単語クラスタリングと呼ばれている．クラス n-gram モデルという観点から類似の振舞いをする単語を 1 つのクラスとして扱うことで，確率推定がより正確になり，かつモデルの記述に必要な記憶領域も減少する．その一方で，振舞いが異なる単語を 1 つのクラスとして扱うと，次のクラスを予測するための確率分布が学習コーパスの特徴をうまく表さなくなる可能性がある．クラス n-gram モデルに対して適切な単語とクラスの関係を見つけることを目的とした単語クラスタリングが提案されており，それによる単語 n-gram モデルの改善が報告されている[9),11),12)]．

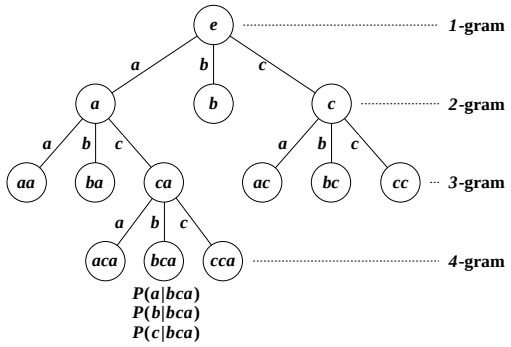


図 1 単語に基づく確率的接尾辞木
Fig. 1 A word-based probabilistic suffix tree.

2.3 可変長記憶マルコフモデル

単語 n-gram モデルの別の拡張として可変長記憶マルコフモデル[10]が提案されている．可変長記憶マルコフモデルは，確率的接尾辞木（PST；probabilistic suffix tree）で表現される．このモデルでは，履歴として参照される文脈の長さ（n の値）が，文脈が予測に有効かどうかに応じて変化する．

アルファベット Σ 上の確率的接尾辞木 T は，次数 $|\Sigma|$ の木である．木の各辺には，Σ の記号 1 つが付与されており，各内部節点からは，親節点への辺のほかには，アルファベットの各記号を持つ辺が出ている．節点は，その節点から根への経路の各辺の記号の連接に等しい文字列 s と，履歴が s である場合の次の記号の確率分布 $\gamma_s : \Sigma \to [0,1]$ の組 (s, γ_s) を持つ．各節点に付与された確率分布は，以下の条件を満たす．

$$\sum_{\sigma \in \Sigma} \gamma_s(\sigma) = 1$$

確率的接尾辞木は，無限長の文字列を生成することができるが，ここでは，有限の長さの文字列の生成確率について考える．確率的接尾辞木 T が文字列 $\boldsymbol{r} = r_1 r_2 \cdots r_m \in \Sigma^m$ を生成する確率は

$$P_T(\boldsymbol{r}) = \prod_{i=1}^{m} \gamma_{s^{i-1}}(r_i)$$

である．ここで $s^0 = e$ であり，$s^j (1 \le j \le m-1)$ は，T の根から $r_j r_{j-1} \cdots r_1$ に対応する辺をたどることで到達できる最も深い節点に対応する文字列を表す．たとえば，**図 1** の確率的接尾辞木による文字列 “$abcab$” の生成確率は $P(a|e)P(b|a)P(c|b)P(a|bc)P(b|bca)$ となる．

確率的接尾辞木は，次の記号の予測に際して，根から履歴をたどることで到達できる最も深い節点を計算

する必要があるので，これに基づく可変長記憶マルコフモデルは確率値の計算コストが n-gram モデルに比べて高い．しかし，確率的接尾辞木と等価な確率有限状態オートマトンが存在することが証明されており，次の記号の予測を行う節点へ直接遷移することが可能となる[10]．

可変長記憶マルコフモデルは，英語の言語モデル[14]や英語の品詞付与[15]に応用されている．これらの応用では，確率的接尾辞木のアルファベット Σ は，英語の品詞である．また，日本語の形態素解析[16]にも応用されており，この場合のアルファベット Σ は形態素である．

3. クラスに基づく可変長記憶マルコフモデル

この章では，我々が提案する新しい言語モデルについて述べる．これは，可変長記憶マルコフモデルの拡張であり，可変長記憶マルコフモデルをその特別な場合として含む．

3.1 確率的接尾辞木への単語クラスタリングの導入

すでに述べたように，単語 n-gram モデルにクラスの概念を導入することで，モデルの記述に必要となる記憶領域を確実に減少させると同時に，場合によっては予測力を向上させることができる．したがって，可変長記憶マルコフモデルにクラスの概念を導入することで，従来の単語に基づく可変長記憶マルコフモデルの記述に必要となる記憶領域を減少させ，予測力を向上させることが可能となると考えられる．ここでも，単語とクラスの関係としてどのようなものを利用するかが重要である．直観的にはまず品詞が考えられるが，品詞などの人間が定義したクラスに基づく n-gram モデルは，一般的に，単語 n-gram モデルよりも予測力が劣る．したがって，品詞をアルファベットとする可変長記憶マルコフモデル[14),15)]は，単語をアルファベットとする可変長記憶マルコフモデルよりも予測力が低いと考えられる．それゆえ，予測力の低下を招くことなくクラスの概念を可変長記憶マルコフモデルに導入するためには，クラス n-gram モデルのための単語クラスタリング[9),11),12)]に類似する方法を利用すべきである．これら単語クラスタリングの先行研究から，最適な単語とクラスの関係はその関係を利用するモデルに依存することが分かる．このことから，可変長記憶マルコフモデルにとって最適な単語とクラスの関係は，それが利用される文脈に依存すると考えられる．たとえば，2 つ前の単語に対する最適な単語とクラスの関係は，直前の単語が「する」である場合と「を」である場合では異なるであろう．これを実現するためには，文脈に応じて単語とクラスの関係を設ける必要がある．

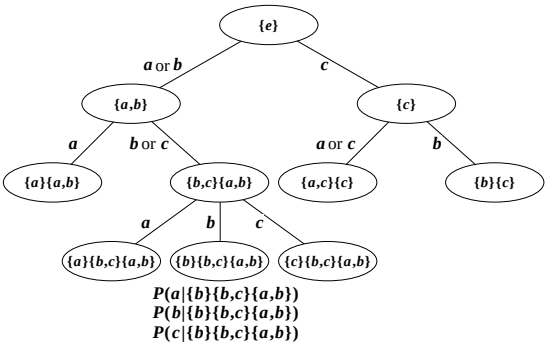


図 2 クラスに基づく確率的接尾辞木
Fig. 2 A class-based probabilistic suffix tree.

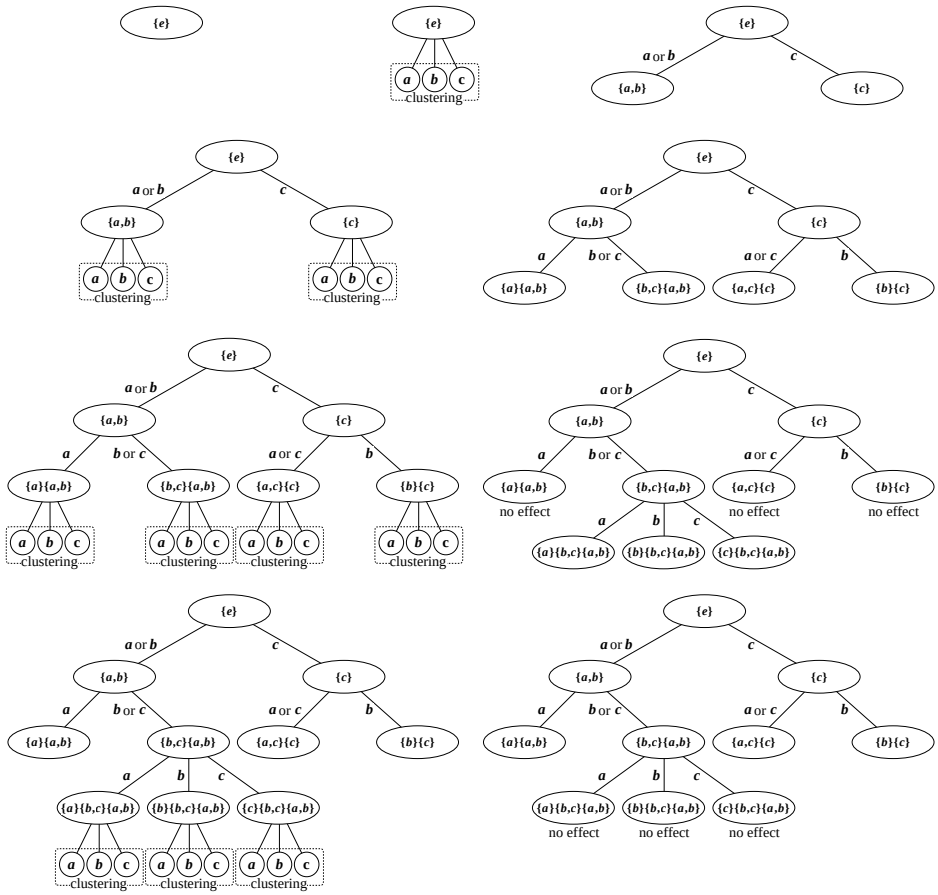
本論文で提案するクラスに基づく可変長記憶マルコフモデルは，クラスに基づく確率的接尾辞木によって表現される．木の各辺には，アルファベット Σ の部分集合に対応するクラスが付与されており，各内部節点からは，親節点への辺のほかに，アルファベットの各記号を含むクラスを持つ辺が出ている．各節点は，その節点から根への経路の各辺の記号集合の連接に等しい文字列集合 S と，履歴が S の要素である場合の次の記号の確率分布 $\gamma_S : \Sigma \to [0,1]$ の組 (S, γ_S) を持つ．各節点に付与された確率分布は，以下の条件を満たす．

$$\sum_{\sigma \in \Sigma} \gamma_S(\sigma) = 1$$

したがって，このモデルは，「先週の日曜」，「先週の月曜」，…，「先週の土曜」などの類似する文脈を 1 つの節点で表し，データスパースネス問題を，単語に基づく可変長記憶マルコフモデルよりもさらに緩和すると期待される．クラスに基づく確率的接尾辞木 T が Σ^m 上の文字列 $\boldsymbol{r} = r_1 r_2 \cdots r_m$ を生成する確率は以下の式で与えられる．

$$P_T(\boldsymbol{r}) = \prod_{i=1}^{m} \gamma_{S^{i-1}}(r_i)$$

ここで $S^0 = \{e\}$ であり，$1 \le j \le m-1$ に対して S^j は，T の根から $r_j r_{j-1} \cdots r_1$ に対応する辺をたどることで到達できる最も深い節点に対応する文字列集合を表す．たとえば，**図 2** のクラスに基づく確率的接尾辞木による文字列 “*abcab*” の生成確率は $P(a|\{e\})P(b|\{a,b\})P(c|\{a,b\}\{a\})P(a|\{b\}\{c\})P(b|\{b\}\{b,c\}\{a,b\})$ である．クラス n-gram モデルとの相違点として，次の単語は，クラスからではなく，履歴から直接予測されることに注意しなければなら



クラスに基づく確率的接尾辞木がアルゴリズムの実行に従って構築される様子（幅優先）を左から右に，そして上から下に向かって描いてある．

図 3 学習アルゴリズムの実行にともなう木の成長の様子
Fig. 3 An illustrative run of the learning algorithm.

ない．

従来の確率的接尾辞木と同様，クラスに基づく確率的接尾辞木も等価な確率有限状態オートマトンに変換することが可能であり，これにより各単語の予測に際して根から辺をたどることなく次の節点へ遷移することが可能になる．

3.2 学習アルゴリズム

クラスに基づく確率的接尾辞木は，以下のようにして構築される．初期状態では，木は空文字列のみからなる集合 $\{e\}$ をラベルに持つ唯一のノード（根）のみから構成される．アルゴリズムの実行に従って，**図 3** に示されるようにノードが再帰的に木に付加される☆．アルゴリズムは，以下の 2 つの処理で構成される．

☆ 誌面の都合から，図 3 では幅優先でノードが生成されていく様子を示しているが，再帰的なアルゴリズムでは深さ優先でノードが生成されていく．

- **clustering** ($node$)
  ノード $node$ のラベルで与えられる文脈に先行する記号に対して，最適の記号とクラスの関係を計算し，それを用いて特殊化することが有効である場合にはその関係を返し，そうでない場合には偽を返す．有効か否かを判断する関数については後で述べる．たとえば，図 2 のラベル $\{b,c\}\{a,b\}$ を持つノードでは，クラスタリングの対象は，ラベルで示される文脈 $(\{ba, bb, ca, cb\})$ の直前の記号がクラスタリングの対象である．換言すれば，この処理では，パターン $y\{b,c\}\{a,b\}x$ にマッチする学習コーパス中の文字列の変数 y を記号 x をより正確に予測することを目的に分類する．後述の実験に用いたクラスタリングのアルゴリズムは，文献 12) と同じである．初期状態では，各単語はその単語のみが属するクラスに分類されてい

expand(node)

```
return if (clustering(node) = false)
foreach child (nodes corresponding to classes)
    create child
    expand(child)
```

図 4 接尾辞木を構築する再帰的手続き
Fig. 4 The recursive process of tree creation.

る．この状態から，可能な 2 つのクラスの組合せに対してこれらを併合した場合の評価関数（後述）の変化を計算し，最適な 2 つのクラスの併合を選択し実行する．いかなる組合せも評価関数を改善しなくなればクラスタリングは終了する．

- **expand** (*node*)
  クラスタリングの処理を呼び出し，その結果得られたクラスに対応するノードを生成することで再帰的に *node* を展開する．**図 4** はこの処理の流れを示す．

3.3 基　　準

一般に，確率的言語モデルの予測力は，クロスエントロピーで表される．クロスエントロピーは，テストコーパスを $C_t = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_m\}$ とし，モデルによる単語列の生成確率を $P_{model}(\boldsymbol{w}_i)$ とすると，以下のように定義される．

$$H(C_t, P_{model}) = -\frac{1}{n}\sum_{i=1}^{m} \log_2 P_{model}(\boldsymbol{w}_i)$$

ここで，n はテストコーパスに含まれる文の総文字数である．クロスエントロピーは，モデルによるテストコーパスの記述長と見なすことができ，小さいほどモデルの予測が正確であることを意味する．

予測力は，確率的言語モデルの最も重要な評価基準であるので，単語クラスタリングやノードの展開の判断基準の要件は，この予測力を高めると期待され，現実的な時間で計算できることである．これまでの研究で提案されている基準としては，クラスに基づく n-gram モデルの構築を目的とした学習コーパスのエントロピー[9),11),15)]や，同じくクラスに基づく n-gram モデルの構築を目的とした平均クロスエントロピーと呼ばれるクロスエントロピーの模倣[12)]や，可変長記憶マルコフモデルの構築を目的とした KL-divergence[10)] などがある．単純な頻度も試みられているが，結果は良好ではないと報告されている[14)]．

これらの提案の中から我々は，クラスに基づく 2-gram モデルの構築のための単語クラスタリングに最適であると報告されている平均クロスエントロピー[12)]を採用した☆．この基準は，補間係数の推定に最適とされる削除補間法の拡張である．ある学習コーパスの平均クロスエントロピー $\overline{H}$ は，学習コーパスを k 個に分割することを前提に，以下のように定義される．

$$\overline{H} = \frac{1}{k}\sum_{i=1}^{k} H(C_i, M_i)$$

ここで，C_i は i 番目の部分コーパスであり，M_i は C_i を除いた学習コーパスから推定された確率分布であり，$H(C, M)$ はコーパス C のモデル M によるクロスエントロピーである．前述のとおり，クラスタリングのアルゴリズムは，文献 12) と同様，ボトムアップである．つまり，初期状態では，各単語はその単語のみが属するクラスに分類されている．次いで，クラスタリングアルゴリズムは，複数のクラスの併合を試みる．その際，併合前の平均クロスエントロピー $\overline{H}_b$ と併合後の平均クロスエントロピー $\overline{H}_a$ の差を計算する．クラスタリングの各段階で，平均クロスエントロピーの差 $\Delta\overline{H} = \overline{H}_a - \overline{H}_b$ が負となるクラスの組の併合が行われる．平均クロスエントロピーの差が負となる併合が複数ある場合には，差 $\Delta\overline{H}$ が最小となるクラスの組の併合が行われる．この考えを拡張して，確率的接尾辞木のノードを展開するか否かも，平均クロスエントロピーの差を基準とすることにした．しかし，判断には閾値を導入し，この符号ではなくその閾値との比較の結果を用いることとした．つまり，以下のように，クラスタリング前の平均クロスエントロピー $\overline{H}_\beta$ とクラスタリング後の平均クロスエントロピー $\overline{H}_\alpha$ の差が 1-gram モデルによる平均クロスエントロピーの t 倍より小さい場合のみ確率的接尾辞木のノードを展開する．

$$\overline{H}_\alpha - \overline{H}_\beta < t \times \overline{H}_{\text{1-gram}}$$

閾値を導入することにより，モデルのパラメータ数を増加させるだけで，効果が無視できる程度であるようなノードの展開や，場合によっては負の効果を持つノードの展開を避けることができる．閾値を決めるパラメータの値は，学習コーパスの一部を用いることで決定することができる．次章で述べる実験では $t = -0.0004$ とした．

4. 評　　価

前章までに説明したモデルを英語のコーパスと日本

☆ 記述長最小原理（MDL）を基準とすることも試みたが，良好な結果は得られなかった．

表 1 WSJ コーパス（英語）
Table 1 WSJ Corpus (English).

	文数	単語数	文字数	カバー率
学習	44,288	1,056,631	4,715,227	97.34%
テスト	4,920	117,135	523,455	95.83%

表 2 EDR コーパス（日本語）
Table 2 EDR Corpus (Japanese).

	文数	単語数	文字数	カバー率
学習	46,755	1,149,827	1,815,326	97.00%
テスト	20,780	509,261	802,576	95.44%

語のコーパスから構築し，予測力やモデルの記述に必要な記憶領域についての実験を行った．この章では，実験の結果とそれに基づくモデルの評価について述べる．

4.1 コーパス

実験に用いた英語のコーパス[17]は，Wall Street Journal（アメリカの経済新聞）の記事の文からなり，それぞれの単語には，45 ある品詞のいずれかが付与されている．コーパスの大きさは**表 1** のとおりである．日本語のコーパス[18]は，主に新聞や雑誌の記事の文からなる EDR コーパスであり，それぞれの単語には，15 ある品詞のいずれかが付与されている．**表 2** はこのコーパスの大きさである．日英それぞれの学習コーパスは，削除補間法による補間係数の推定などのために 9 個の部分コーパスに分割した．

それぞれのモデルの語彙は，9 個の部分コーパスの 2 個以上に出現する単語と品詞の組とした．この結果，英語のコーパスから推定されたモデルの語彙は，27,377 の単語と品詞の組であり，テストコーパスのカバー率は 95.83%であった．同様に，日本語のコーパスから推定されたモデルの語彙は，26,792 の単語と品詞の組であり，テストコーパスのカバー率は 95.44%であった．アルファベットは，語彙と各品詞の未知語を表す特別な記号と文区切り記号の和集合である．未知語の生成確率は，文字 2-gram モデルに基づく未知語モデル[19]によって与えられる．

4.2 比較対象のモデル

提案モデルの評価のために，単語 2-gram モデルと，単語 3-gram モデルと，クラス 2-gram モデルと，可変長記憶マルコフモデルとを同じ学習コーパスから推定し，同じテストコーパスに対して評価した．それぞれの言語モデルの構成の手順は以下のとおりである．

- 単語 2-gram モデル
  (1) 削除補間により式 (1) の補間の係数を推定.
  (2) すべての学習コーパスを対象に単語 2-gram と単語 1-gram を計数.
- 単語 3-gram モデル
  (1) 削除補間により式 (1) の補間の係数を推定.
  (2) すべての学習コーパスを対象に単語 3-gram と単語 2-gram と単語 1-gram を計数.
- クラス 2-gram モデル
  (1) 文献 12) の方法で単語クラスタリング.
  (2) 削除補間により式 (2) の補間の係数を推定.
  (3) すべての学習コーパスを対象にクラス 2-gram とクラス 1-gram を計数.
- 単語可変長記憶マルコフモデル
  (1) クラス可変長記憶マルコフモデルのクラスタリング部分を省略.

未知語モデルは，すべてのモデルに共通である．

なお，実験に利用した計算機の演算装置は Pentium III 933 MHz であり，クラス 2-gram モデルの推定に要した時間は 2 時間 8 分，クラス可変長記憶マルコフモデルの推定に要した時間は 7 時間 38 分であった．

4.3 評価基準

確率的言語モデルの最も重要な評価基準は予測力であり，これはテストコーパスに対するエントロピー（クロスエントロピー）で表される．クロスエントロピーは，モデルによるテストコーパスの記述長と見なすことができ，小さいほどモデルの予測が正確であることを意味する．

他の重要な評価基準として，モデルが必要とする記憶域の大きさがある．これは，モデルの非零のパラメータの数で測られる．非零のパラメータの数が少ないほどモデルに必要な記憶域が小さい．スパースネスを測るためにモデルの全パラメータの数も計数した．

モデルの応用例として，英語の品詞タガー（以下では単にタガーと呼ぶ）と日本語の形態素解析器を作成した．これは，品詞という概念を内包する確率的言語モデルを基にして，与えられた文字列 $\boldsymbol{x}$ に対する確率最大の表記と品詞の対の列 $\hat{\boldsymbol{w}}$ を計算することで実現される．これは，以下の式で表される（$\boldsymbol{w}$ の表記の連接は $\boldsymbol{x}$ に等しい）．

$$\begin{aligned}\hat{\boldsymbol{w}} &= \mathop{\mathbf{argmax}}_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{x}) \\ &= \mathop{\mathbf{argmax}}_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{x})P(\boldsymbol{x}) \\ &\qquad\qquad (\because P(\boldsymbol{x}) \text{ は } \boldsymbol{w} \text{ によらず}) \\ &= \mathop{\mathbf{argmax}}_{\boldsymbol{w}} P(\boldsymbol{x}|\boldsymbol{w})P(\boldsymbol{w}) \quad (\because \text{ベイズの公式}) \\ &= \mathop{\mathbf{argmax}}_{\boldsymbol{w}} P(\boldsymbol{w}) \quad (\because P(\boldsymbol{x}|\boldsymbol{w}) = 1)\end{aligned}$$

この式の最後の $P(\boldsymbol{w})$ が品詞という概念を内包する確率的言語モデルである．表記と品詞の対を単語と定

表 3 WSJ コーパスに対する結果
Table 3 Result of WSJ corpus.

言語モデル	クラス v-gram	単語 v-gram	クラス 2-gram	単語 2-gram	単語 3-gram
クロスエントロピー	1.7434	1.7467	1.7705	1.7846	1.7234
品詞付与の精度	95.17%	95.02%	94.93%	94.95%	94.95%
全パラメータの数	183.5 M	434.0 M	211.8 M	752.1 M	20.62 T
非零パラメータの数	312.8 K	470.7 K	182.1 K	349.1 K	726.9 K

v-gram = 可変長記憶マルコフモデル，$K = 1000$，$M = 1000^2$，$G = 1000^3$，$T = 1000^4$

表 4 EDR コーパスに対する結果
Table 4 Result of EDR corpus.

言語モデル	クラス v-gram	単語 v-gram	クラス 2-gram	単語 2-gram	単語 3-gram
クロスエントロピー	4.0090	4.0253	4.0789	4.1345	4.0117
形態素解析の適合率	92.22%	91.76%	92.19%	91.77%	91.77%
形態素解析の再現率	92.61%	92.48%	92.26%	92.38%	92.60%
全パラメータの数	139.2 M	2.190 G	52.09 M	718.7 M	19.27 T
非零パラメータの数	253.7 K	412.4 K	79.39 K	257.2 K	605.6 K

v-gram = 可変長記憶マルコフモデル，$K = 1000$，$M = 1000^2$，$G = 1000^3$，$T = 1000^4$

義とすることで，このモデルとして，単語 n-gram モデルやクラス n-gram モデルや可変長記憶マルコフモデルを用いることができる．

タガーの精度は，タガーによって正しい品詞が付与された単語の割合で測られる．付与された品詞が正しいか否かは，コーパスにあらかじめ付与されている品詞との比較により判別する．形態素解析器の精度は，以下の式で定義される再現率と適合率と呼ばれる値で測られる．

$$\text{再現率} = \frac{\text{正しい形態素の数}}{\text{テストコーパスの形態素の数}}$$

$$\text{適合率} = \frac{\text{正しい形態素の数}}{\text{形態素解析器の出力の形態素の数}}$$

ここで,「正しい形態素」とは，文頭からの文字数という意味でテストコーパスと同じ位置に出現し，テストコーパスと同じ品詞が付与された形態素のことである．

4.4 実験結果

表 3 は英語コーパスに対する実験の結果であり，**表 4** は日本語コーパスに対する実験の結果である．また，履歴長とその割合（頻度を考慮せず）を**表 5** と**表 6** に提示した．

タガーの精度は，実験条件の相違から正確には比較できないが，現在の最高水準（約 97%）[20] よりも低い傾向にある．この原因は，未知語モデルが単純であることと，語彙が少ないことによりテストコーパスにおけるカバー率が低いことであろう．英語のモデルと日本語のモデルの語彙とそのカバー率（表 1 と表 2 参照）はすでに述べたとおりであるが，学習コーパスに出現する単語を語彙とすれば，英語のテストコーパスのカバー率は 97.47%となり，日本語のテストコーパスのカバー率は 97.07%となる．これは，我々のモデルの語彙によるカバー率より，英語で 1.64%，日本語で 1.63%高い．これが，タガーや形態素解析の精度の絶対的な値が最高水準の値と比較して低い理由の 1 つであろう．もう 1 つの理由として，我々が用いている未知語モデルが単純であることがあげられる．我々が用いている未知語モデルは，文字 2-gram モデルに基づいているが，これは，特に英語において未知語の品詞決定に重要な情報を持つと考えられる長めの接尾辞や接頭辞の性質をとらえることができない．しかしながら，実験においては語彙や未知語モデルは各モデルに共通なので，以下で述べるモデル間の比較という意味では問題はない．

表 5 履歴の長さ（WSJ コーパス）
Table 5 History length (WSJ Corpus).

長さ	1	2	3
割合	46.7%	50.3%	3.0%

表 6 履歴の長さ（EDR コーパス）
Table 6 History length (EDR Corpus).

長さ	1	2	3
割合	36.0%	64.0%	0.0%

実験の結果，以下のことが示された．クラスに基づく可変長記憶マルコフモデルに必要な記憶領域は，単語 2-gram モデルに必要な記憶領域よりもわずかに小さく，さらに単語 3-gram モデルよりもはるかに小さいが，クラス 2-gram モデルよりは大きい．クラス 2-gram モデルは，実験した中では必要な記憶域が最小のモデルであるが，クラスに基づく可変長記憶マル

コフモデルと比較して，クロスエントロピーは高く，タガーや形態素解析器の精度はわずかながら低い．クラスに基づく可変長記憶マルコフモデルは，クロスエントロピーとタガーや形態素解析器の精度と必要な記憶域という観点で，単語に基づく可変長記憶マルコフモデルや単語 3-gram モデルよりも優れている．ただし，英語コーパスにおける単語 3-gram モデルとのクロスエントロピーの比較は例外である．確率的言語モデルの多くの応用においては，最も重要な評価基準は，クロスエントロピーで測られる予測力や，タガーや形態素解析器などの応用の精度である．したがって，クラスに基づく可変長記憶マルコフモデルが，実験において比較したモデルのなかで最も優れていると結論できる．

5. おわりに

本論文では，クラスに基づく可変長記憶マルコフモデルとそれを利用した確率的言語モデルについて述べた．クラスに基づく可変長記憶マルコフモデルは，可変長記憶マルコフモデルの拡張であり，各ノードの展開に際して単語クラスタリングが行われる．したがって，クラスに基づく可変長記憶マルコフモデルの構築には，従来の単語に基づく可変長記憶マルコフモデルの構築よりも多くの時間がかかる．その一方で，クラスに基づく可変長記憶マルコフモデルは，可変長記憶マルコフモデルに比べて，必要となる記憶領域が小さく，かつ予測力がより高くなることが期待される．このことは，実験の結果により傍証された．

実験では，英語のコーパスと日本語のコーパスから，クラスに基づく可変長記憶マルコフモデルと単語に基づく可変長記憶マルコフモデルと単語 3-gram モデルを構築し，それらのモデルに基づく英語の品詞タガーと日本語の形態素解析器を作成した．その結果，クラスに基づく可変長記憶マルコフモデルは，単語に基づく可変長記憶マルコフモデルや単語 3-gram に比べて，必要となる記憶領域が小さく，かつ予測力がより高かった．実験の結果として，記憶領域の大きさの差は顕著であるが，予測力の差は他の n-gram モデルの改良と同様に顕著ではなかった．したがって，本論文で提案したクラスに基づく可変長記憶マルコフモデルは，予測力の低下を招くことなく，モデルの記述に必要な記憶領域を縮小させたい場合に特に有効であろう．

参考文献

1) Bahl, L.R., Jelineck, F. and Mercer, R.L.: A Maximum Likelihood Approach to Continuous Speech Recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.6, No.2, pp.179–190 (1983).
2) Nagata, M.: Context-Based Spelling Correction for Japanese OCR, *Proc. 16th International Conference on Computational Linguistics* (1996).
3) 森 信介，土屋雅稔，山地 治，長尾 真：確率的モデルによる仮名漢字変換，情報処理学会論文誌，Vol.40, No.7, pp.2946–2953 (1999).
4) Chen, Z. and Lee, K.-F.: A New Statistical Approach to Chinese Pinyin Input, *Proc. 38th Annual Meeting of the Association for Computational Linguistics*, pp.241–247 (2000).
5) Jelinek, F.: Self-Organized Language Modeling for Speech Recognition, Technical Report, IBM T.J. Watson Research Center (1985).
6) 永田昌明：EDR コーパスを用いた確率的日本語形態素解析，EDR 電子化辞書利用シンポジウム，pp.49–56 (1995).
7) Brown, P.F., Cocke, J., Pietra, S.A.D., Pietra, V.J.D., Jelineck, F., Lafferty, J.D., Mercer, R.L. and Roossin, P.S.: A Statistical Approach to Machine Translation, *Computational Linguistics*, Vol.16, No.2, pp.79–85 (1990).
8) Shannon, C.E.: Prediction and Entropy of Printed English, *Bell System Technical Journal*, Vol.30, pp.50–64 (1951).
9) Brown, P.F., Pietra, V.J.D., deSouza, P.V., Lai, J.C. and Mercer, R.L.: Class-Based n-gram Models of Natural Language, *Computational Linguistics*, Vol.18, No.4, pp.467–479 (1992).
10) Ron, D., Singer, Y. and Tishby, N.: The Power of Amnesia: Learning Probabilistic Automata with Variable Memory Length, *Machine Learning*, Vol.25, pp.117–149 (1996).
11) Ney, H., Essen, U. and Kneser, R.: On Structuring Probabilistic Dependences in Stochastic Language Modeling, *Computer Speech and Language*, Vol.8, pp.1–38 (1994).
12) Mori, S., Nishimura, M. and Itoh, N.: Word Clustering for a Word Bi-gram Model, *International Conference on Speech and Language Processing* (1998).
13) Jelinek, F., Mercer, R.L. and Roukos, S.: Principles of Lexical Language Modeling for Speech Recognition, *Advances in Speech Signal Processing*, chapter 21, pp.651–699, Dekker (1991).
14) Niesler, T.R. and Woodland, P.C.: Variable-length category n-gram language models, *Computer Speech and Language*, Vol.13, pp.99–124 (1999).
15) Schütze, H. and Singer, Y.: Part of Speech

Tagging Using a Variable Memory Markov Model, *Proc. 32nd Annual Meeting of the Association for Computational Linguistics*, pp.181–187 (1994).
16) 春野雅彦，松本裕治：文脈木を利用した形態素解析，情報処理学会研究報告 NL112 (1996).
17) Marcus, M.P. and Santorini, B.: Building a Large Annotated Corpus of English: The Penn Treebank, *Computational Linguistics*, Vol.19, No.2, pp.313–330 (1993).
18) Japan Electronic Dictionary Research Institute, Ltd.: *EDR Electronic Dictionary Technical Guide* (1993).
19) Nagata, M.: A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm, *Proc. 15th International Conference on Computational Linguistics*, pp.201–207 (1994).
20) Thede, S.M. and Harper, M.P.: A Second-Order Hidden Markov Model for Part-of-Speech Tagging, *Proc. 37th Annual Meeting of the Association for Computational Linguistics*, pp.175–182 (1999).

(平成 13 年 3 月 5 日受付)
(平成 13 年 11 月 14 日採録)

森　信介（正会員）

1998 年京都大学大学院博士後期課程修了．同年日本アイ・ビー・エム（株）入社．東京基礎研究所において計算言語学の研究に従事．工学博士．1997 年本学会山下記念研究賞受賞．