形態素クラスタリングによる形態素解析精度の向上

森　信介[†]　長尾　眞[†]

本論文では，形態素クラスタリングと未知語モデルの改良による確率的形態素解析器の精度向上を提案する．形態素クラスタリングとしては，形態素 n-gram モデルをクロスエントロピーを基準としてクラス n-gram モデルに改良する方法を提案する．未知語モデルの改良としては，確率モデルの枠組の中で学習コーパス以外の辞書などで与えられる形態素を追加する方法を提案する．bi-gram モデルを実装し EDR コーパスを用いて実験を行なった結果，形態素解析の精度の向上が観測された．両方の改良を行なったモデルによる形態素解析実験の結果の精度は，先行研究として報告されている品詞 tri-gram モデルの精度を上回った．これは，我々のモデルが形態素解析の精度という点で優れていることを示す結果である．これらの実験に加えて，品詞体系と品詞間の接続表を文法の専門家が作成した形態素解析器との精度比較の実験を行なった．この結果，確率的形態素解析器の誤りは文法の専門家による形態素解析器の誤りに対して有意に少なかった．形態素解析における確率的な手法は，このような人間の言語直感に基づく形態素解析器と比較して，現時点で精度がより高いという長所に加えて，今後のさらなる改良にも組織的取り組みが可能であるという点で有利である．

キーワード：　形態素解析，*n-gram* モデル，コーパス，クラスタリング，未知語

An Improvement of a Morphological Analysis by a Morpheme Clustering

SHINSUKE MORI[†] and MAKOTO NAGAO[†]

This paper proposes improving a stochastic Japanese morphological analyzer through a morpheme clustering and an amelioration of the unknown word model. As a morpheme clustering, we propose a method which ameliorates a morpheme-based n-gram model into a class-based n-gram model with cross entropy criterion. As an amelioration of the unknown word model, we propose a method to incorporate a given morpheme set, such as dictionary, into it. As the result of experiments on the EDR corpus, we observed improvements of the accuracy. The analyzer adopting both methods marked a higher accuracy than an anteriorly reported part-of-speech-based tri-gram model. This result tells us that our morphological analyzer is better than the previous one in terms of accuracy. In addition to these experiments, we compared our analyzer with a grammarian's intuition-based analyser. The experimental results have shown the error rate of the stochastic analyzer was meaningfully smaller than that of the heuristic analyzer. The stochastic approach to Japanese morphological analysis is of great advantage to the ad-hoc method in higher accuracy, as well as in facility of further organized improvements.

KeyWords:　*morphological analysis, n-gram model, corpus, clustering, unknown word*

† 京都大学工学研究科 電子通信工学専攻, Department of Electrical Engineering, Kyoto University

1 まえがき

日本語には単語間に明示的な区切りがないので，入力文を単語に分割し，品詞を付加する形態素解析は日本語処理における基本的な処理である．このような視点から，今日までに多くの形態素解析器が人間の言語直観に基づき作成されている．一方，英語の品詞タグ付けではいくつかのコーパスに基づく方法が提案され，非常に高い精度を報告している (DeRose 1988; Church 1988; Brill 1992; Cutting, Kupiec, Pedersen, and Sibun 1992; Dermatas and Kokkinakis 1995; Charniak, Hendrickson, Jacobson, and Perkowitz 1993; de Marcken 1990; Weischedel, Meteer, Schwartz, Ramshaw, and Palmucci 1993; Merialdo 1994; Brill 1994; Dermatas and Kokkinakis 1995; Brill 1995; Franz 1997)．今日，多くの研究者が，英語の品詞タグ付けに関してはコーパスに基づく手法が従来のヒューリスティックルールに基づく手法より優れていると考えるに至っている．

日本語の形態素解析に対しては，コーパスに基づく手法が従来のルールに基づく手法より優れていると考えるには至っていないようである．これは，コーパスに基づく形態素解析の研究には，ある程度の規模の形態素解析済みのコーパスが必要であり，日本語においてはこのようなコーパスが最近になってようやく簡単に入手可能になったことを考えると極めて自然である．実際，コーパスに基づく形態素解析に関しては現在までのところ少数の報告がなされているのみである (丸山, 荻野, 渡辺 1991; Nagata 1994; 永田 1995; 竹内・松本 1995)．これらの研究で用いられているモデルはすべてマルコフモデル (n-gram モデル) であり，状態に対応する単位という観点から以下のように分けられる．

- 単語 (列) が状態に対応する (丸山他 1991)．
- 品詞 (列) が状態に対応する (Nagata 1994; 永田 1995; 竹内・松本 1995)

確率的言語モデルという観点からは，単語を単位とすることは過度の特殊化であり，品詞を単位とすることは過度の一般化である．これらは，未知コーパスの予測力を低下させ，形態素解析の精度を下げる原因になっていると考えられる．

我々は，この問題に対処するために，予測力を最大にするという観点よって算出したクラスと呼ばれる単語のグループを一つの状態に対応させ，基礎となる確率言語モデルを改良し，結果として形態素解析の精度を向上する方法を提案する．確率言語モデルとしてのクラス n-gram モデルは，最適なクラス分類を求める方法 (以下，クラスタリングと呼ぶ) とともにすでに提案されている (Brown, Pietra, deSouza, Lai, and Mercer 1992; Ney, Essen, and Kneser 1994; Kneser and Ney 1993)．しかし，これらの文献で報告されている実験では，クラスタリング結果を用いたクラス n-gram モデルの予測力は必ずしも向上していない．これらに対して，文献 (提出中) では削除補間 (Jelinek and Mercer 1980) を応用したクラスタリング規準とそれを用いたクラスタリングアルゴリズムを提案し，クラス n-gram モデルの予測力が有意に向上したことを報告している．本論文では，この方法を応用することで得られるクラス n-gram モデルを

基礎にした確率的形態素解析器による解析精度の向上について報告する．また，未知語モデルに確率モデルの条件を逸脱することなく外部辞書を追加する方法を提案し，この結果として得られる未知語モデルを備えた確率的形態素解析器による解析精度の向上ついても報告する．さらに，上述の改良の両方を施した確率的形態素解析器と品詞体系と品詞間の接続表を文法の専門家が作成した形態素解析器との解析精度の比較を行なった結果について述べる．

2　確率的形態素解析

日本語に対する形態素解析とは，日本語の文 (文字列) を入力とし，これを表記と品詞の直積として定義される形態素に分割する処理である．この節では，これを実現する手法の一つとしての確率的形態素解析とその基礎となる確率言語モデルと解の探索方法について述べる．

2.1　形態素解析の問題の定義

日本語の形態素解析は，日本語のアルファベット $\mathcal{X}$ のクリーネ閉包に属する文 $x \in \mathcal{X}^*$ を入力として，これを表記 $\mathcal{W} = \mathcal{X}^*$ と品詞 $\mathcal{T}$ の直積として定義される形態素 $\mathcal{M} = \{(w,t)|w \in \mathcal{W} \wedge t \in \mathcal{T}\}$ の列 m に分解して出力することと定義できる．このとき，出力される形態素列の表記の連接は，入力のアルファベット列に等しくなければならない．つまり，入力のアルファベット列 (長さ l) を $x = x_1 x_2 \cdots x_l$ とし，出力の形態素列 (要素数 h) を $m = m_1 m_2 \cdots m_h$ とすると以下の式が成り立つ必要がある．ただし，$w(m)$ は形態素 m の表記を表し，$w(m)$ は形態素の連接 m の表記の連接を表わすものとする．

$$w(m) = w(m_1)w(m_2)\cdots w(m_h) = x_1 x_2 \cdots x_l = x \tag{1}$$

一般に，これを満たす解は一意ではない．形態素解析の問題は，可能な解の中から人間の判断 (正解) に最も近いと推測される形態素列 (単語分割と品詞割り当て) を選択し出力することである．この選択の基準としては，文法の専門家が自身の言語直観を頼りにした規則に基づく方法と大量の正解例 (形態素解析済みコーパス) からの推定を規準にする方法がある．以下では，後者の一つである確率的形態素解析について説明する．

2.2　確率的形態素解析

確率的形態素解析器は，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 x に対する確率最大の形態素列 $\hat{m}$ を計算し出力する．これは，以下の式で表される．

$$\hat{m} = \underset{w(m)=x}{\operatorname{argmax}} P(m)$$

この式の最後の $P(m)$ が品詞という概念を内包する確率的言語モデルである．

このような確率的言語モデルには様々なものが考えられる．これらの良否の尺度としては，クロスエントロピーが一般的である．これは，確率的言語モデルを M とし，テストコーパスを $S = \{s_1, s_2, \ldots, s_k\}$ とすると以下の式で与えられる[1]．ただし，$|s|$ は文 s の長さ (文字数) を表わす．

$$H(M,S) = -\frac{\sum_{i=1}^{k} \log M(s_i)}{\sum_{i=1}^{k} (|s_i| + 1)}$$

1 式の分母の+1 は文末記号に対応する．これは，s_x, s_y と $s_x s_y$ を区別するために必要である．

この値は，コーパス Sをモデル Mで符合化した時の文字あたりの平均符合長の下限であり，Sとして無作為に抽出された十分多数の文を選択すれば，複数のモデルの良否を比較するための尺度となる．定義から明らかなように，この値がより小さいほうがより良い言語モデルである．クロスエントロピーの意味で良い言語モデルを用いる方が形態素解析の精度が良いと考えられる．

形態素 n-gram モデル

確率的言語モデル $P(m)$ は，形態素を一つずつ予測することを仮定すると，以下のように書き換えられる．

$$\begin{aligned} P(m) &= P(m_1 m_2 \cdots m_{h+1}) \\ &= \prod_{i=1}^{h+1} P(m_i | m_1 m_2 \cdots m_{i-1}) \end{aligned} \tag{2}$$

ここで m_{h+1}は，文末に対応する特別な記号である．これを導入することによって，すべての可能な形態素列に対する確率の和が 1 となることが保証される (Fu 1974)．

式 (2) は，ある時点 i での形態素 m_iの出現確率は最初の時点から時点 $i-1$ までの全ての形態素に依存することを表しているが，実装の簡便さなどを考慮して，時点 $i-k$から時点 $i-1$ までの連続する k個の形態素の履歴にのみ依存する k重マルコフ過程であると仮定する．この仮定は，以下の式で表される近似である．

$$P(m_i | m_1 m_2 \cdots m_{i-1}) \approx P(m_i | m_{i-k} m_{i-k+1} \cdots m_{i-1})$$

ここで m_j $(j \leq 0)$ は，文頭に対応する特別な記号である．これを導入することによって式が簡便になる．

一般に，確率 $P(m_i | m_{i-k} m_{i-k+1} \cdots m_{i-1})$ の値はコーパスから最尤推定することで得られる．これは，$N(m)$ を形態素列mのコーパスにおける頻度として，以下の式で与えられる．

$$\begin{aligned} P(m_i | m_{i-k} m_{i-k+1} \cdots m_{i-1}) &= \frac{N(m_{i-k} m_{i-k+1} \cdots m_i)}{N(m_{i-k} m_{i-k+1} \cdots m_{i-1})} \\ &= \frac{N(m_{i-k} m_{i-k+1} \cdots m_i)}{\sum_m N(m_{i-k} m_{i-k+1} \cdots m_{i-1} m)} \end{aligned}$$

このように，このモデルは連続する $n = k+1$ 個の形態素列の頻度に基づいているので，形態素 n-gram モデルと呼ばれる．

形態素 n-gram モデルにおいて場合に問題となるのは，状態に対応する形態素 (既知形態素) の選択である．一般的には，頻度の高い形態素を既知形態素とすることで高い予測力が実現できる．しかし，どのような形態素の集合を選択したとしても，テストコーパスに出現する可能性のあるすべての形態素が既知形態素であることは望めない．このため，未知形態素の扱いが避けられない問題となる．この問題に対処するため，未知形態素に対応する特別な記号を用意

し，既知の形態素以外はこの記号から次節で述べる未知語モデルにより与えられる確率で生成されることとする．未知形態素に対応する特別な記号は，かならずしも唯一である必要はなく，品詞などの情報を用いて区別される複数の記号であっても良い．我々の目的は形態素解析であるから未知形態素であっても品詞の推定が可能でなければならない．よって，各品詞に対して未知形態素に対応する記号を設ける．

以上に述べた形態素 n-gram モデル M_m による，形態素列の出現確率は以下の式で表される．ただし $\mathcal{M}_{in}$ は既知形態素の集合を表わし，t は m_i の品詞を表わす．また，$\mathtt{UM}_t$ は品詞 t に属する未知形態素に対応する特別な記号である．

$$M_m(m_1 m_2 \cdots m_h) = \prod_{i=1}^{h+1} P_m(m_i | m_{i-k} \cdots m_{i-2} m_{i-1}) \tag{3}$$

$$P_m(m_i | m_{i-k} \cdots m_{i-2} m_{i-1}) = \begin{cases} P(m_i | m_{i-k} \cdots m_{i-2} m_{i-1}) & \text{if} \quad m_i \in \mathcal{M}_{in} \\ P(\mathtt{UM}_t | m_{i-k} \cdots m_{i-2} m_{i-1}) M_{x,t}(m_i) & \text{if} \quad m_i \notin \mathcal{M}_{in} \end{cases} \tag{4}$$

この式の中の $M_{x,t}$ は，次項で述べる未知語モデルであり，品詞が t であることを条件として，引数で与えられる文字列の生成確率を値とする．

確率値の最尤推定においては，まず既知形態素集合を定義し，学習コーパスの未知形態素を未知形態素に対応する特別な記号に置き換えて頻度を計数する．

未知語モデル

未知語モデルは，表記から確率値への写像として定義され，既知形態素以外のあらゆる形態素の表記を 0 より大きい確率で生成し，この確率をすべての表記に渡って合計すると 1 以下になる必要がある．このような条件を満たすモデルの一つとして，文字 n-gram モデルがある．日本語の表記に用いられる文字は有限と考えられるので，形態素 n-gram モデルのときの未知形態素のような問題は起こらない．しかし，形態素 n-gram モデルの場合と同様に，文字を既知文字と未知文字に分類し，未知文字はこれを表わす特別な記号から生成されるものとすることもできる．文字の使用頻度には大きな偏りがあることが予測されるので，これらを一つのグループとみなすことで，モデルが改善されると考えられる．文字集合は有限であるから，未知形態素モデルの場合と異なり，各未知文字の生成確率を等確率とすることができる．このようにして構成される未知語モデルは以下の式で表わされる．ただし $\mathcal{X}_{in,t}$ は品詞が t である未知語モデルの既知文字の集合を表わし，$\mathtt{UX}_t$ は品詞 t の未知語モデルの未知文字に対応する特別な記号である．また，$w(m) = x_1 x_2 \cdots x_h$ としている．

$$M_{x,t}(w(m)) = M_{x,t}(x_1 x_2 \cdots x_h) \tag{5}$$

$$= \prod_{i=1}^{h+1} P_{x,t}(x_i | x_{i-k} \cdots x_{i-2} x_{i-1}) \tag{6}$$

$$P_{x,t}(x_i|x_{i-k}\cdots x_{i-2}x_{i-1}) = \begin{cases} P(x_i|x_{i-k}\cdots x_{i-2}x_{i-1}) & \text{if} \quad x_i \in \mathcal{X}_{in,t} \\ P(\mathrm{UX}_t|x_{i-k}\cdots x_{i-2}x_{i-1})\frac{1}{|\mathcal{X}-\mathcal{X}_{in,t}|} & \text{if} \quad x_i \notin \mathcal{X}_{in,t} \end{cases} \tag{7}$$

この式の中の x_j $(j \leq 0)$ は，語頭に対応する便宜的な記号である．また，x_{h+1} は，語末に対応する特別な記号であり，形態素に対するモデルの場合と同様に，すべての可能な文字列に対する確率の和が 1 となるために導入されている．

以上で説明した未知語モデルは，未知文字を等確率で生成するモジュールを「未知文字モデル」と考えると，形態素 n-gram モデルと相似の構造である．文字 n-gram モデルの確率値は，形態素 n-gram モデルの場合と同様に，既知文字を定義した後，未知形態素の実例における文字列の頻度から推定される．

低頻度事象への対処

上述したように，形態素 n-gram モデルのパラメータ推定には，出現頻度を基にした最尤推定が用いられる．しかし，対象とする事象の頻度が低い場合には，推定値の信頼性は低くなるという問題がある．この問題に対処する方法として，補間と呼ばれる方法が用いられる (Jelinek, Mercer, and Roukos 1991)．これは，次の式で表されるように，より信頼性が高いことが期待される，より低次のマルコフモデルの遷移確率を一定の割合で足し合わせるという操作を施すことを言う．

$$P'(m_i|m_{i-k}m_{i-k+1}\cdots m_{i-1}) = \sum_{j=0}^{k} \lambda_j P(m_i|m_{i-j}m_{i-j+1}\cdots m_{i-1}) \tag{8}$$

$$\text{ただし } 0 \leq \lambda_j \leq 1, \sum_{j=0}^{k} \lambda_j = 1$$

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 (Jelinek and Mercer 1980) 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全ての組合せ (k通り) に渡って行ない，その平均値をとるという方法である．

解の探索アルゴリズム

形態素 n-gram モデルによる形態素解析器は，入力として文字列xを受けとり，式 (3)(4)(6)(7)(8) を用いて計算される確率が最大の形態素列$\hat{m}$を式 (1) で表わされる条件の

下で計算し出力する．解の探索には動的計画法を用いることができ，入力の文字数 n に対して計算時間のオーダーが $O(n)$ となるアルゴリズムが提案されている (Ney 1984) (Nagata 1994)．

3 未知語モデルの改良

この章では，確率的形態素解析の精度を向上させる方法として，未知語モデルに外部辞書を付加する方法を提案する．これは，確率的言語モデルの予測力を改善する方法であり，確率的形態素解析の精度向上を直接の目的としているわけではないが，確率的言語モデルの予測力の改善は，結果としてそれに基づく確率的形態素解析器の解析精度を向上させる．また，予測力の高い未知語モデルを推定するための未知形態素の実例の収集方法についても述べる．

3.1 外部辞書の付加

前章で述べた未知語モデル $M_{x,t}$は，未知形態素だけでなく既知形態素の表記も 0 より大きい確率で生成する可能性がある．この場合には，以下の式が示すように，未知形態素の生成確率の合計は 1 未満となる．以下の説明では品詞 t を省略してある．また，形態素の集合を表わす記号 $\mathcal{M}_{in}$をその表記の集合を表わすとしている．

$$\sum_{m\in\mathcal{X}^*-\mathcal{M}_{in}} M_x(m)+\sum_{m\in\mathcal{M}_{in}} M_x(m)=\sum_{m\in\mathcal{X}^*} M_x(m)=1$$
$$\Leftrightarrow \sum_{m\in\mathcal{X}^*-\mathcal{M}_{in}} M_x(m)=1-\sum_{m\in\mathcal{M}_{in}} M_x(m)<1 \quad (\because M_x(m)>0,\ \exists m\in\mathcal{M}_{in})$$

これは，言語モデルとしての条件を満たしてはいるが，クロスエントロピーという点で改善の余地がある．つまり，既知形態素の生成確率を何らかの方法で未知形態素に分配することで，未知形態素の生成確率が大きくなり，テストコーパスにそのような未知形態素が出現した場合に，テストコーパスの出現確率が大きくなる．

既知形態素の生成確率の分配には，様々な方法が考えられるが，以下の式が表すように，すべての未知形態素にその生成確率に比例して分配する方法が一般的であろう．

$$M'_x(x_1x_2\cdots x_h)=\begin{cases}0 & \text{if} \quad m\in\mathcal{M}_{in}\\ \frac{1}{1-\sum_{m\in\mathcal{M}_{in}} M_x(m)} M_x(x_1x_2\cdots x_h) & \text{if} \quad m\notin\mathcal{M}_{in}\end{cases} \tag{9}$$

これに対して我々は，辞書の見出し語などとして与えられる形態素の部分集合に等しく配分することを提案する．つまり，ある形態素の集合が与えられたとして，ここから既知形態素を除いた集合を $\mathcal{M}_{ex}$ $(\mathcal{M}_{ex}\cap\mathcal{M}_{in}=\phi)$ として，この要素の生成確率を文字 n-gram モデルによる確率と既知形態素の生成確率の合計を $\mathcal{M}_{ex}$ の要素数で割った値の和とする．

$$M'_x(m)=M_x(m)+\frac{1}{|\mathcal{M}_{ex}|}\sum_{m\in\mathcal{M}_k} M_x(m) \quad (m\in\mathcal{M}_{ex}) \tag{10}$$

これは，既知形態素の生成確率を，学習コーパスには現れないが辞書などから形態素であると考

えられる文字列に優先的に分配し，それらの生成確率を相対的に高くすることを意味する．このような文字列の集合を外部辞書と呼ぶ．形態素解析が目的なので，外部辞書には文字列のほかにその品詞が記述されている必要がある．この方法により，確率言語モデルの枠内で，コーパスから推定された確率言語モデルに辞書などの異なる情報源の情報を付加できる．

以上に述べた外部辞書を備えた未知語モデル M'_x による文字列 $m = x_1x_2\cdots x_h$ の出現確率は以下の式で表される．

$$M'_x(x_1x_2\cdots x_h) = \begin{cases} 0 & \text{if} \quad m \in \mathcal{M}_{in} \\ M_x(x_1x_2\cdots x_h) + \frac{1}{|\mathcal{M}_{ex}|}\sum_{m\in\mathcal{M}_{in}} M_x(m) & \text{if} \quad m \in \mathcal{M}_{ex} \\ M_x(x_1x_2\cdots x_h) & \text{if} \quad m \notin \mathcal{M}_{in} \wedge m \notin \mathcal{M}_{ex} \end{cases}$$

これを式 (5) の代わりに用いる未知語モデルを外部辞書を備えた未知語モデルと呼ぶ．

3.2　未知形態素の実例の収集方法

文字 n-gram モデルの確率値は，形態素 n-gram モデルの場合と同様に，アルファベットを定義してから，未知形態素の実例における文字列の頻度から推定される．未知形態素の実例の収集の方法としては，学習コーパスに含まれるすべての形態素とすることや，学習コーパスにおける頻度が 1 である形態素とする (永田 1996) などが考えられる．我々は，削除補間法を応用した以下の方法を提案する．

> 学習コーパスを k個の部分コーパスに分割し，i 番目の部分コーパスの未知形態素の実例を，i 番目の部分コーパス以外を学習コーパスとし i 番目の部分コーパスをテストコーパスと見た場合の未知形態素とする．

我々が提案する方法は，削除補間法を応用して，実際のテストコーパスにおける未知形態素と類似した実例を得ているので，他の方法よりも優れていると予測される．実際に，予備実験としてこれらの方法を実装し，予測力という規準で比較した．その結果，我々が提案する方法が最良であった．したがって，実験にはこの方法を用いた．

4 形態素クラスタリング

この章では，形態素 n-gram モデルの一般化の一つであるクラス n-gram モデルを説明し，文献 (提出中) を応用して形態素解析のためのクラスを自動的に学習する方法を提案する．前章と同様に，確率的言語モデルの予測力の改善を目的としているが，学習されたクラス n-gram モデルに基づく確率的形態素解析器の解析精度は，形態素 n-gram モデルや人間の言語直観による品詞をクラスとした場合の品詞 n-gram モデルに基づく確率的形態素解析器の解析精度より高くなると考えられる．

4.1 クラス n-gram モデル

クラス n-gram モデル (Brown et al. 1992) では，あらかじめ形態素をクラスと呼ばれるグループに分類しておき，先行するクラスの列を直前の事象とみなして分類する．このモデルでは，次の形態素を直接予測するのではなく，次のクラスを予測した上で次の形態素を予測する．以下の式で，C_{in}は既知形態素に対応するクラスであり，これを品詞とすれば，品詞 n-gram モデルとなる．

$$P(m) = \prod_{i=1}^{h+1} P_c(m_i|c_{i-k}c_{i-k+1}\cdots c_{i-1}) \tag{11}$$

$$P_c(m_i|c_{i-k}\cdots c_{i-2}c_{i-1}) = \begin{cases} P(c_i|c_{i-k}\cdots c_{i-2}c_{i-1})P(m_i|c_i) & \text{if} \quad \exists c_i \in C_{in},\ m_i \in c_i \\ P(\mathrm{UM}_i|c_{i-k}\cdots c_{i-2}c_{i-1})M_{x,t}(m_i) & \text{if} \quad \forall c_i \in C_{in},\ m_i \notin c_i \end{cases} \tag{12}$$

この式の中の c_j $(j \leq 0)$ は，文頭に対応する特別な記号である．これを導入することによって式が簡便になる．また，c_{h+1}は，語末に対応する特別な記号であり，これを導入することによって，すべての可能な文字列に対する確率の和が 1 となる (Fu 1974)．

形態素に基づくモデルの場合と同様に，確率 $P(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1})$ の値および，確率 $P(m_i|c_i)$ の値は，コーパスから最尤推定することで得られる．

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

$$P(m_i|c_i) = \frac{N(m_i, c_i)}{N(c_i)}$$

この式において，クラスを品詞とすれば品詞 n-gram モデルが得られ，形態素からクラスへの写像が全単射であれば，形態素 n-gram モデルと等価になることが分かる．また，これをマルコフモデルと考えると，状態はクラスに対応する．

形態素 n-gram モデルと同様に，データスパースネスの問題に対処する方法として，補間を用いることができる (Brown et al. 1992)．これは，以下のように式 (8) において形態素 m をク

ラス c と読み変えれることで容易に得られる．

$$P'(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1}) = \sum_{j=0}^{k} \lambda_j P(c_i|c_{i-j}c_{i-j+1}\cdots c_{i-1}) \tag{13}$$

$$\text{ただし } 0 \leq \lambda_j \leq 1, \sum_{j=0}^{k} \lambda_j = 1$$

4.2　形態素クラスタリング

確率言語モデルの形態素クラスタリングの課題は，クロスエントロピーが最も低くなる形態素とクラス (図1の中の $c_1, c_2, \cdots, c_x$) の対応関係を算出することである．このようなクラスを用いて構築されたクラス n-gram モデルに基づく確率的形態素解析器の解析精度は，品詞 n-gram モデルや形態素 n-gram モデルに基づく確率的形態素解析器の解析精度よりも高くなることが期待される．従って，形態素クラスタリングの目的関数は，削除補間を応用することでクロスエントロピーを模擬すると考えられる以下のような値とした．

$$\overline{H} = \frac{1}{k}\sum_{i=1}^{k} H(M_i, S_i) \tag{14}$$

ここで，M_i は i 番目以外の $k-1$ の部分コーパスから推定された確率言語モデルであり，S_i は i 番目の部分コーパス (文の列) を表す．ここで問題としているのは，確率的言語モデルとしてクラス n-gram モデルを用いた場合の形態素のクラスタリングである．この場合，コーパスは一定であり，確率的言語モデル M は形態素とクラスの関係 F にのみ依存する．従って，上式の平均クロスエントロピーは，形態素とクラスの関係の関数とみなすことができる．この値がより小さいほうが，未知のコーパスに対してより良い言語モデルであることが予測される．よって，クラスタリングの目的は，式 (14) で定義される平均クロスエントロピーを最小化する形態素とクラスの関係を求めることである．

形態素とクラスの対応関係としては，ある形態素が一定の確率で複数のクラスに属するという確率的な関係も考えられるが，解空間が広大になるので，本研究では形態素は唯一のクラスに属することを仮定した．よって，クラスの集合は形態素の集合の直和分割となる．形態素とクラスの対応関係 F は，$\mathcal{M}, \mathcal{C}$ をそれぞれ形態素の集合とクラスの集合とすると，関数 $f : \mathcal{M} \mapsto \mathcal{C}\ (= 2^{\mathcal{M}})$ を用いて表すことができ，この関数は以下の条件を満たす[2]．

$$M = \bigcup_{m \in M} f(m)$$

$$\forall m \in M \text{ に対し } m \in f(m)$$

$$f(m_1) \neq f(m_2) \Rightarrow f(m_1) \cap f(m_2) = \phi$$

2 f の値は形態素の集合である (例: $f(m_1) = \{m_1, m_2, m_3\}$)．

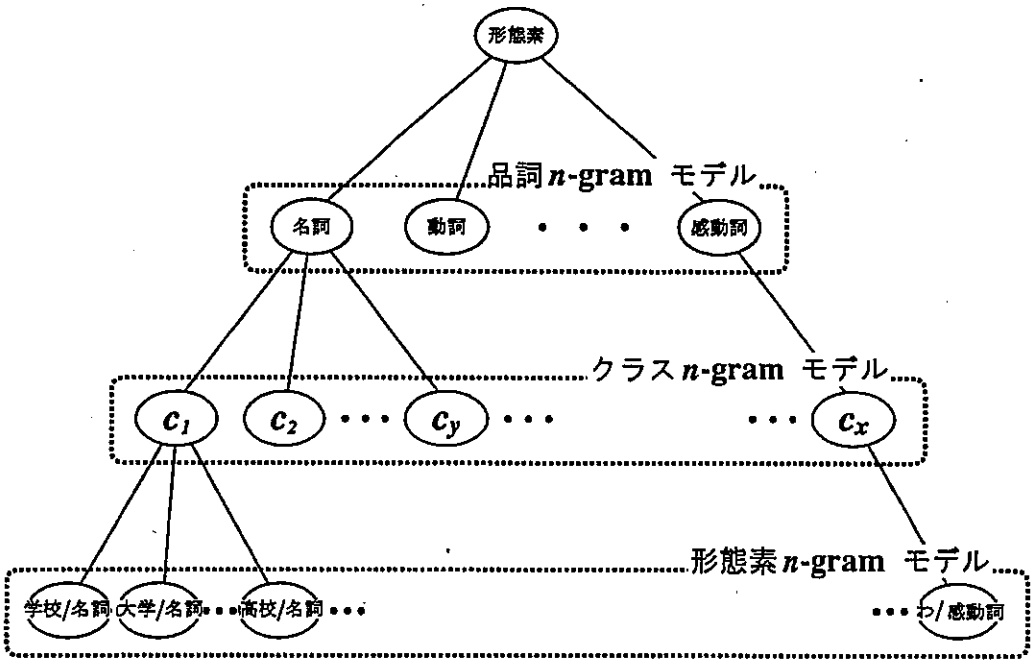

図 1　各 n-gram モデルの概念図

解探索のアルゴリズム中で用いるために，形態素とクラスの対応関係に対して，以下の関数を定義する．

- $move : F \times M \times C \mapsto F$

  $move(f, m, c)$ は，形態素とクラスの関係 f に対して形態素 m をクラス c に移動した結果得られる形態素とクラスの関係を返す．

クラスタリングの解空間はあらゆる可能な形態素とクラスの対応関係である．しかし，この数はある程度の大きさの語彙数に対しては非常に大きいため，これら全てに対して平均クロスエントロピーを計算し，これを最小化するクラス関係を選択するということは，計算量という観点から不可能である．平均クロスエントロピーの値はクラス関係の一部分の変更が全体に影響するという性質をもっているので，分割統治法や動的計画法を用いることもできない．以上のことから，我々は最適解を求めることを諦め，貪欲アルゴリズムを用いることにした．このアルゴリズムは以下の通りである (図 2参照)．なお，$\overline{H}$は式 (14) で与えられる平均クロスエントロピーでり，$t(m)$ や $t(c)$ は形態素 m やクラス c の品詞を表わす．同一の品詞である形態素に対してのみ併合を試みるので，結果としてどのクラスも同一の品詞の形態素のみを要素に持つことに注意しなければならない．

Mを頻度の降順に並べ m_1, m_2, ..., m_n とする
foreach i $(1, 2, \cdots, n)$
　$c_i := \{m_i\}$
　$f(m_i) := c_i$
foreach i $(2, 3, \cdots, n)$
　$c := \mathbf{argmin}_{c \in \{t(c)=t(m_i) \mid c_1,\, c_2,\, \ldots,\, c_{i-1}\}} \overline{H}(move(f, m_i, c))$
　if $(\overline{H}(move(f, m_i, c)) < \overline{H}(f))$ **then**
　　$f := move(f, m_i, c)$

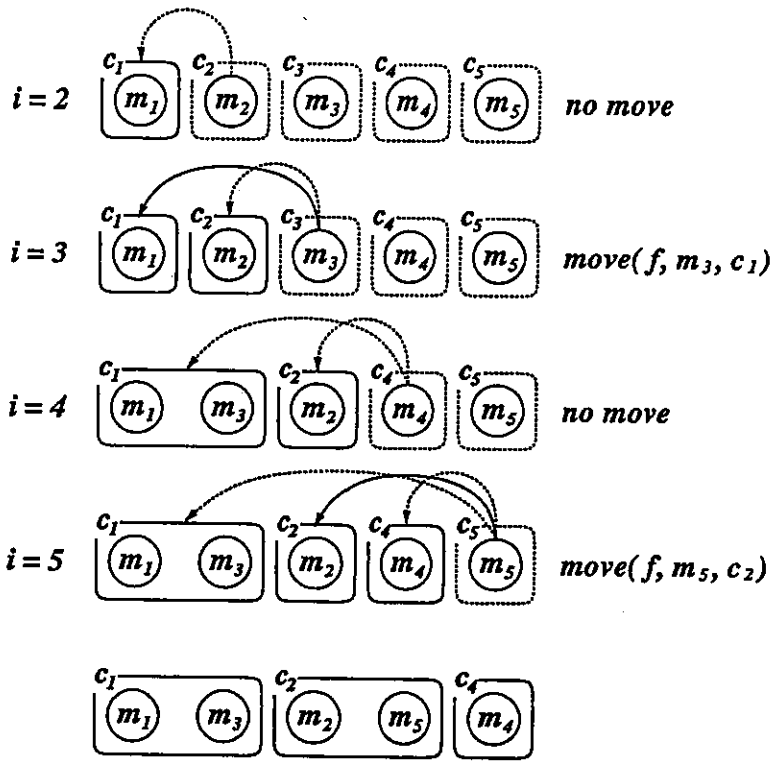

図 2　クラスタリングの概念図

計算量は，二番目の foreach での繰り返しの回数は単語数 $|W|$ に比例し，argmin での繰り返しの回数はクラス数 $|C|$ に比例するので，全体で $O(|W| \cdot |C|)$ である．クラス数 $|C|$ は，全ての単語が独立したクラスに分けられる場合に最大 $(|C| = |W|)$ となり，全ての単語が同一のクラスとなる場合に最小 $(|C| = 1)$ となる．従って，初期化における全体の計算量は，最良の場合が

$O(|W|)$ であり，最悪の場合が $O(|W|^2)$ である．ただし，単語の並べ替えや一番目の foreach の計算量は係数が非常に小さいと考えられるので，考慮に入れていない．次節で述べる実験の結果では，頻度の高い形態素を対象とする段階では多くの形態素がクラスに併合されずクラス数は形態素数に比例し最悪の場合に近い挙動を示したが，頻度の低い形態素を対象とする段階ではほとんどの形態素がクラスに併合されてクラス数はほとんど一定となり最良の場合に近い挙動を示した．頻度の低い形態素が多数を占めるので，計算量は実際にはかなり線形に近いと考えられる．

頻度の高い形態素から移動を試みることとしているのは，頻度の高い形態素の移動のほうがパープレキシティに与える影響が大きいと考えられるので，早い段階での移動が後の移動によって影響されにくく，収束がより速くなると考えたためである．

上述のアルゴリズムによって得られたクラス分類からさらに探索を進めてより良いクラス分類が得られるかを試みることができる．このアルゴリズムとして，さらに形態素の移動を試みること (Ney et al. 1994) やクラスの併合を試みること (Brown et al. 1992) が考えられる．我々は，これらのアルゴリズムを小さなコーパスに対する予備実験で適用してみたが，必要となる計算時間が膨大である割にはクロスエントロピーの改善が小さかった．よって，次章では，上述のアルゴリズムによる実験結果について述べる．

5　実験結果とその評価

前節で述べた方法の有効性を確かめるため，以下の点を明らかにするための実験を行った．

(1)　外部辞書による解析精度の向上

(2)　形態素クラスタリングによる解析精度の向上

以下では，まず形態素解析精度の評価基準について述べ，実験の条件を明確にし，上述の実験の結果を提示し評価する．また，文法の専門家による形態素解析器との解析精度の比較を行なった結果について述べる．なお以下では，「クラス n-gram モデル」などの言語モデルを表す表現を，文脈から明らかな場合には，その言語モデルに基づく形態素解析器を表すためにも用いる．

5.1　評価基準

我々が用いた評価基準は，(永田 1995) で用いられた再現率と適合率であり，次のように定義される．EDR コーパスに含まれる形態素数を N_{EDR}，解析結果に含まれる形態素数を N_{SYS}，分割と品詞の両方が一致した形態素数を N_{COR} とすると，再現率は N_{COR}/N_{EDR} と定義され，適合率は N_{COR}/N_{SYS} と定義される．例として，コーパスの内容と解析結果が以下のような場合を考える．

コーパス

外交 (名詞)　政策 (名詞)　で (助動詞)　は (助詞)　な (形容詞)　い (形容詞語尾)

解析結果

外交政策 (名詞)　で (助詞)　は (助詞)　な (形容詞)　い (形容詞語尾)

この場合，分割と品詞の両方が一致した形態素は「は (助詞)」と「な (形容詞)」と「い (形容詞語尾)」であるので，$N_{COR} = 3$ となる．また，コーパスには 6 つの形態素が含まれ，解析結果には 5 つの形態素が含まれているので，$N_{EDR} = 6$, $N_{SYS} = 5$ である．よって，再現率は $N_{COR}/N_{EDR} = 3/6$ となり，適合率は $N_{COR}/N_{SYS} = 3/5$ となる．

5.2　実験の条件

実験には EDR コーパス (日本電子化辞書研究所 1993) を用いた．まず，これを 10 個に分割し，この内の 9 個を学習コーパスとし，残りの 1 個をテストコーパスとした．前章で述べたように，クラス関数の推定では，この 9 個の学習コーパスのうちの 8 つから n-gram モデルを推定し，残りの 1 つのコーパスに対してクロスエントロピーを求めるということを 9 通り行なって得られる平均クロスエントロピーを評価規準とする．それぞれのコーパスに含まれる文と形態素と文字の数 (のべ) は表 1の通りである．既知形態素は，2 個以上の学習コーパスに現れる 59,956 個の形態素とした．形態素 bi-gram モデルは，これらに対応する状態の他に，各品詞の未知語に対応する状態 (15 個) と文区切り (文末と文頭) に対応する状態を持つ．同様に，クラ

スbi-gramモデルは，既知形態素をクラスタリングすることで得られるクラスに対応する状態と，各品詞の未知語に対応する状態と文区切りに対応する状態を持つ．

表1　EDRコーパスの大きさ

コーパス	文数	形態素数	文字数
学習コーパス	187,022	4,595,786	7,252,558
テストコーパス	20,780	509,261	802,576

形態素bi-gramモデルとクラスbi-gramモデルを比較するために，これらを同じ学習コーパスから構成し，同じテストコーパスに対してパープレキシティや形態素解析の精度を計算した．それぞれの言語モデルの構成の手順は以下の通りである．

- 形態素bi-gramモデル
  - (1)　削除補間により式(8)の補間の係数を推定
  - (2)　すべての学習コーパスを対象に形態素bi-gramと形態素uni-gramを計数
- クラスbi-gramモデル
  - (1)　削除補間により式(8)の補間の係数を推定
  - (2)　前章で述べた方法($k=9$)でクラス関数を推定
  - (3)　削除補間により式(13)の補間の係数を推定
  - (4)　すべての学習コーパスを対象にクラスbi-gramとクラスuni-gramを計数

未知語モデルは共通であり，各品詞(15個)に対して形態素bi-gramモデルと同様の手順で構成される．本実験では行なっていないが，文字に対するクラスタリングを行ない，これをクラスbi-gramモデルとすることも可能である．外部辞書の形態素集合は，EDR日本語単語辞書(日本 1993)の見出し語から既知形態素を除いた形態素集合と学習コーパスには出現するが既知形態素とならなかった形態素集合(分割された学習コーパスの1個にのみ現れた形態素)の和集合とした．

品詞毎の形態素数とクラスタリングの結果得られたクラスの数を表2に掲げた．平均要素数は，形態素数をクラス数で割った値である．この値は，内容語において高く，機能語において低いことが観測される．このことから，品詞n-gramモデルにおいては機能語を一般化し過ぎており，形態素n-gramモデルにおいては内容語を特殊化し過ぎているということが分かる．

なお，対象となる59,956の形態素をクラスタリングするのに要した時間は，SPARC Station 20 (150MHz)で約4日であった．

表 2　品詞毎の形態素数とクラス数

品詞	助詞	名詞	語尾	動詞	記号
形態素の数	108	44453	95	8352	80
クラスの数	59	3680	74	1260	30
平均要素数	1.83	12.08	1.28	6.63	2.67

助動詞	接尾語	数字	副詞	形容動詞	形容詞
110	789	1473	1411	2035	572
69	262	90	193	199	89
1.59	3.01	16.37	7.31	10.23	6.43

連体詞	接続詞	接頭語	感動詞	合計
128	148	170	32	59956
36	31	71	13	6156
3.56	4.77	2.39	2.46	9.74

$$\text{平均要素数} = \frac{\text{形態素の数}}{\text{クラスの数}}$$

5.3　外部辞書と形態素クラスタリングによる精度向上の評価

図 3は，形態素クラスタリングの結果を用いたクラス bi-gram モデルの，外部辞書を持つ場合と持たない場合の，クロスエントロピーと形態素解析の精度である．このグラフから次のようなことが分かる．まず，学習コーパスの大きさと解析精度の関係であるが，解析精度は，コーパスの大きさに対して単調に増加している．しかし，コーパスがある程度大きくなるとこの増加量は小さくなっている．このことは，さらなる精度向上を達成するためには，学習コーパスを増やすという単純な方法は，コーパスの作成コストを考えると，得策ではないということを意味する．次に，外部辞書を付加することによる解析精度の向上であるが，クロスエントロピーの減少から予測される通り，外部辞書を付加することにより解析精度が向上した．グラフから分かるように，学習コーパスの大きさが小さい方が，外部辞書を付加することによる効果が大きい．この理由は，学習コーパスが大きくなると，外部辞書の元となる辞書などに記述されている形態素の大部分が学習コーパスに含まれることになり，テストコーパスに含まれる未知形態素の割合が減少することであると考えられる．この議論から，確率的形態素解析器を用いて学習コーパスと異なる分野の文を解析する場合には，未知形態素となるであろうその分野特有の用語 (表記と品詞) を収集しておき，これを外部辞書として付加することでかなりの精度の向上が望めると考えられる．分野特有の用語の収集方法としては，その分野の専門用語辞書などを直接用いることや，その分野の大量の文例から n-gram 統計を用いて抽出し品詞を推定する

こと (森・長尾 1995) などが考えられる.

表 3は，外部辞書を備えない場合と備えた場合の，形態素 bi-gram モデルとクラス bi-gram モデルによるクロスエントロピーと形態素解析の精度である．また，先行研究との比較のため，外部辞書を備えていない場合の品詞 tri-gram モデルによるクロスエントロピーも表中に記載している．この結果から，外部辞書の有無に関わらず，我々が提案する方法によって得られる単語のクラス分類を用いることで，形態素解析の精度が再現率と適合率の双方で向上していることが分かる．これは，クロスエントロピーの減少から予測される通りの結果である．このように，確率モデルを用いた言語の解析では，クロスエントロピーが減少するようにモデルを改善することで，自然に形態素解析などの解析精度が向上することが見込まれる．ただし，このクロスエントロピーと解析精度の関係は，単調であることが解析的に導出できるような確固たる関係ではないことに注意しなければならない．クロスエントロピーと解析精度の関係が逆になっている例 (上述の関係の反例) として，表 3の中の「形態素 bi-gram+外部辞書」と「クラス bi-gram」のエントロピーと適合率が挙げられる．

表 3　各言語モデルによるクロスエントロピーと形態素解析の精度

モデル	クロスエントロピー	再現率	適合率
形態素 bi-gram	4.6053	93.23%	89.36%
形態素 bi-gram+外部辞書	4.5437	93.37%	89.75%
クラス bi-gram	4.5654	93.32%	89.78%
クラス bi-gram+外部辞書	4.5039	93.41%	90.12%
品詞 tri-gram	5.8643	–	–

文献 (永田 1995) では，品詞 tri-gram モデルを用いた形態素解析をについて述べている．この文献では，我々が今回用いた評価規準と全く同じ評価規準ではなく，単語分割のみや読みも含めた再現率と適合率を報告している．このような評価の一つとして 72,000 文で学習した品詞 tri-gram モデルの単語分割の精度として 90.6%の再現率と 91.7%の適合率を報告している．このモデルとの比較を可能にするために，約 47,000 の学習コーパスで学習した「クラス bi-gram+外部辞書」の単語分割の精度を計算した．この結果，再現率は 94.8%であり，適合率は 94.9%であり，学習コーパスが少し小さいにもかかわらず品詞 tri-gram モデルの結果を双方で上回っている．解析精度に関しては全ての条件が同じというわけではないので単純な比較は適切ではないが，この結果は，本手法の優位性を実験的に示すと考えられる．また，クロスエントロピー (表 3参照) の差は十分有意であると考えられるので，この点からも本手法の形態素解析の精度という点での優位性が十分予測される．しかし，より長い文脈から次の品詞を予測しているという品詞 tri-gram モデルの良い点も無視できない．この点を採り入れて，形態素 tri-gram モデ

ルに対して形態素クラスタリングを実行し，その結果を用いてクラス tri-gram モデルを構築すれば，クロスエントロピーがさらに下がり，形態素解析の精度も上がると考えられる．ただし，実用とするためには，遷移表や解探索のための表が大きくなることによる記憶域の増大と可能な組合せの増加による解探索に必要な時間が増加するという問題にも注意を払う必要がある．

5.4　文法の専門家による形態素解析器との比較

我々は，上述の実験に加えて，文法の専門家による形態素解析器と確率的形態素解析器を解析精度という点で比較するという実験を行なった．この際に最大の問題となるのは評価基準である．確率的形態素解析器の解析精度の比較は容易に行なえる．つまり，我々が上述した実験で行なったように，同一の学習コーパスと同一のテストコーパスを用いた解析結果の再現率と適合率を比較すればよい．これは英文における単語の品詞推定の精度の比較にも用いられる標準的な方法である (英語では単語区切りに曖昧性がないので再現率と適合率は同じ値になる)．しかし，文法の専門家による形態素解析器の解析精度の比較は一般に容易ではない．これは，それぞれの文法の専門家によって形態素の定義 (品詞体系や単語区切り) に違いがあり，正解となるべき形態素解析結果を共有できないことに起因する．その結果，形態素解析器の評価としては，あるいくつかの文の解析結果を文法の専門家も含めた形態素解析器の製作者が観察することで計算される値が用いられる．また，テストは最後に一回だけ行なわれるのではなく，テストの結果を見て形態素解析器を修正するということもあり，完全なオープンテストになっていないこともある．このようなテストの結果得られる精度は，客観性に欠けるので，おおよその目安としてのみ意味があり，複数の形態素解析器の比較に用いることはできない．この問題は，文法の専門家による形態素解析器と確率的形態素解析器の解析精度の比較を行なう際にも現れる．

上述の問題を解決する方法として，同じ文法基準 (品詞体系や単語区切) を持つ形態素解析済みコーパスと文法の専門家による形態素解析器を用いることが考えられる．これが，本研究で我々が選択した解決方法である．具体的には，京都大学で開発された文法の専門家による形態素解析器 JUMAN (松本, 黒橋, 山地, 妙木, 長尾 1997) とその解析結果を人手で修正したコーパス (黒橋・長尾 1997) を用いた．つまり，コーパスを学習コーパスとテストコーパスに分割し (表 4)，学習コーパスから構成した確率的形態素解析器 (外部辞書を備えたクラス bi-gram モデル) と JUMAN を用いてテストコーパスを解析した結果を，テストコーパスにあらかじめ付与されている正解と比較して，それぞれの再現率と適合率を計算した．なお，外部辞書の形態素集合は，学習コーパスには出現するが既知形態素とならなかった形態素集合である．表 5はこの結果である．この表から，テストコーパスにおいては，確率的形態素解析器の誤りが文法の専門家による形態素解析器の誤りに対して 25%程度少ないことが分かる．この実験で使用した解析済みコーパスが JUMAN の出力の訂正であることや，コーパスの訂正の過程で訂正結果を参考にして JUMAN を改良していることを考えると学習コーパスでの比較が適切かも知れな

い．この場合は，確率的形態素解析器の解析精度は表5に示されるように圧倒的に良い．未知語モデルを文字クラスタリングしたクラス n-gram モデルとすることや，外部辞書の源として JUMAN の辞書や別のコーパスを JUMAN で解析した結果から得られる学習コーパスに現れない高頻度の形態素を用いることで，確率的形態素解析器の精度はさらに向上すると考えられる．

表 4　京都大学テキストコーパスの大きさ

コーパス	文数	形態素数	文字数
学習コーパス	8,584	206,812	366,599
テストコーパス	921	22,484	39,826

本実験で比較の対象とした文法の専門家による形態素解析器は，初版の完成から 10 年弱の期間を経ており，この間に莫大な人的資源を投入し様々な改良が施されている．一方，我々の確率的形態素解析器がパラメータ推定に用いた学習コーパスは 8,584 文であり，これを作成する費用はそれほど高くはない．これは，確率的形態素解析器が，文法の専門家による形態素解析器に対して優位である点の一つである．現状での学習コーパスの大きさは $10^{5.56}$文字と比較的小規模であり，図 3の EDR コーパスにおける学習コーパスの大きさと解析精度の関係から，コーパスを増量し確率言語モデルを再学習するということを繰り返すことで，この品詞体系でのより高精度の形態素解析器が容易に実現できると予測される．これと並行して確率言語モデルの改善を行なうことも重要である．以下に，より良い確率的形態素解析器を実現するための指針をまとめる．

- 解析済みコーパスの保守と増量

  コーパスの修正

  人手による修正を受けた解析済みコーパスにも誤りはあり，さらなる修正が必要である．確率的形態素解析器の出力との比較は，これらの誤りを指摘する上で有効であろう．

  コーパスの増量

  すでに指摘したように，学習コーパスは多ければ多いほど良い．新たな文に正解を付加するときには，人手による修正を受けたコーパスを全て用いて，最も良い言語モデルを学習し，その結果得られる確率的形態素解析器による解析結果を修正することで，人手による修正のコストを最小限に抑える必要がある．

品詞体系の変更

形態素解析器の出力を用いた研究や開発の過程で，品詞体系の変更が要求されることがある．例えば，京都大学テキストコーパス (黒橋・長尾 1997) では，「みんな/名詞」と「みんな/副詞」を区別していない．このような区別が必要になれば，まず解析済みコーパスの一部をこの区別を加えて修正し，これと残りのコーパス

表 5　文法家による形態素解析器と確率的形態素解析器の精度比較

形態素解析解析器	学習コーパス		テストコーパス	
	再現率	適合率	再現率	適合率
JUMAN3.2	94.29%	93.67%	94.51%	94.02%
クラス bi-gram+外部辞書	98.42%	98.48%	95.84%	95.67%

で問題となる形態素が出現しないコーパスから形態素解析器を学習し，問題となる形態素が出現する文を曖昧な部分以外を固定して解析し直すことで，人手による修正のコストを最小限に抑えることができる．

- 確率的言語モデルの改良

  確率言語モデルの改善方法は，本論文で提案した形態素クラスタリング以外にも提案されてる．これらは，未知語モデルにも適用できる．

  可変記憶長マルコフモデル

  n-gram モデルでの単語予測は固定長の文脈を条件部にもつが，これを先行する単語に応じて変化させる (Schütze and Singer 1994) (春野・松本 1996)．

  キャッシュモデル

  直前のいくつかの単語の分布 (キャッシュ) を用いて n-gram モデルのパラメータを動的に変化させる (Kuhn and de Mori 1990)．

  複数のモデルの補間

  複数のクラス n-gram モデルを補間したモデルを用いる (McMahon and Smith 1996)．

  これらの改良をうまく組み合わせることで言語モデルの予測力が向上し，結果としてより高い精度の形態素解析器が実現できる．

- 解探索のアルゴリズムやデータ構造の改良

  これによる解析速度や記憶容量の改良は，解析精度の向上にはつながらないが，実用とする上で重要である．解探索のアルゴリズムやデータ構造は，モデルのクラスに依存する点に注意しなければならない．

これらの改善は独立に行なえるので，組織的な取り組みが可能になる．このように，高い精度を実現するための方法論が確立していることが確率的手法の最大の利点であろう．

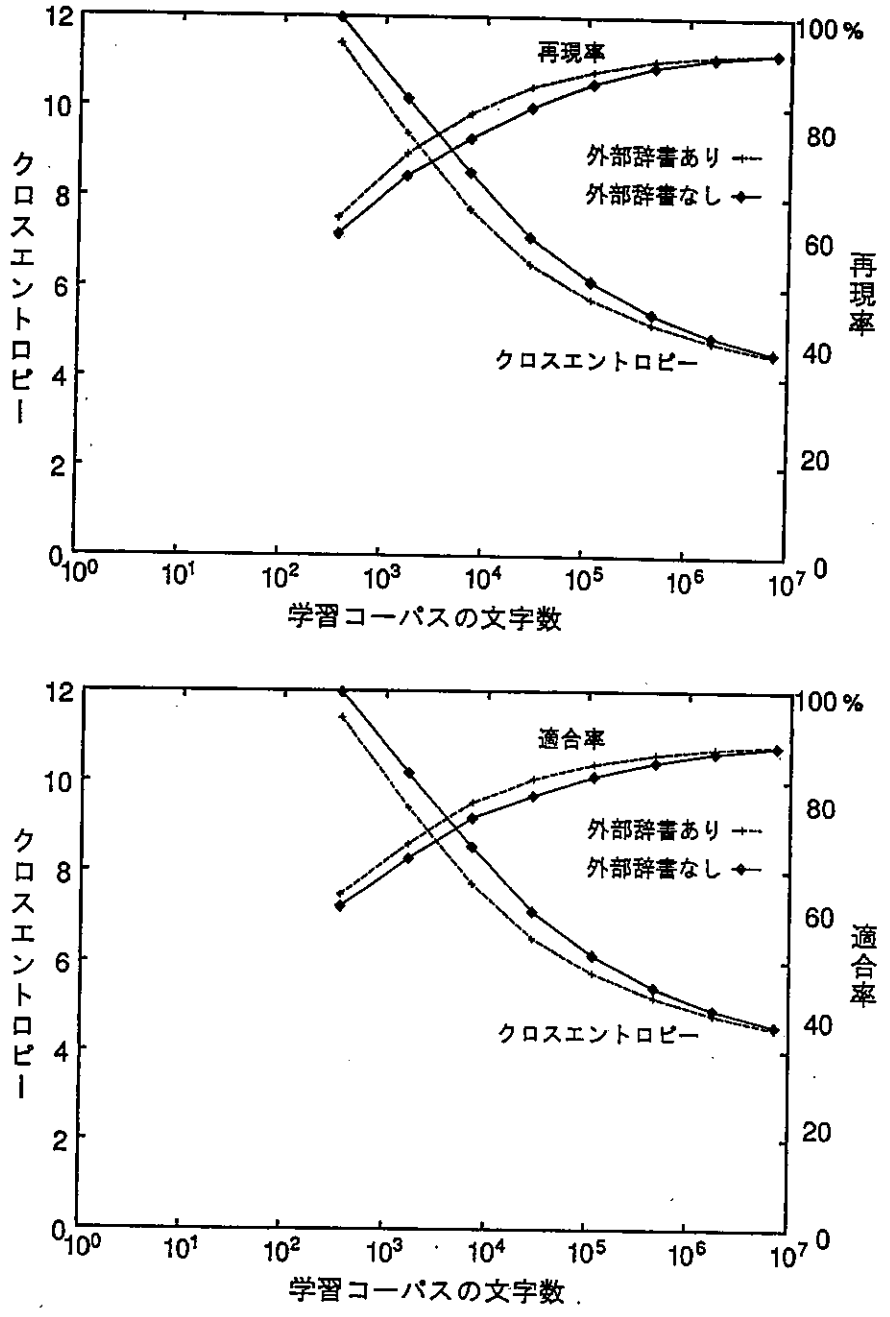

図 3　学習コーパスの大きさと形態素解析精度の関係

6　むすび

本論文では，形態素クラスタリングと外部辞書の付加による確率的形態素解析器の精度向上について述べた．形態素クラスタリングとしては，形態素 n-gram モデルをクロスエントロピーを基準としてクラス n-gram モデルに改良する方法を提案した．bi-gram モデルを実装し EDR コーパスを用いて実験を行なった結果，形態素解析の精度の向上が観測された．また，未知語モデルに外部辞書を付加する方法を提案した．同様の実験を行なった結果，形態素解析の精度の向上が観測された．これは，学習コーパスとは異なる性質を持つ分野の形態素解析器や解析済みコーパスを作成するのに特に有効であろう．両方の改良を行なったモデルによる形態素解析実験の結果の精度は，先行研究として報告されている品詞 tri-gram モデルの精度を上回った．これは，我々のモデルが形態素解析の精度という点で優れていることを示す結果である．これらの実験に加えて，人間の言語直感に基づく形態素解析器との精度比較の実験を行なった．この結果，確率的形態素解析器の誤りは文法家による形態素解析器の誤りに対して 25%程度少なかった．形態素解析における確率的な手法は，人間の言語直感に基づく形態素解析器と比較して，現時点で精度がより高いという長所に加えて，今後のさらなる改良にも組織的取り組みが可能であるという点で有利である．

謝辞

本研究を進めるに過程で，示唆に富んだコメントを頂いた日本アイ・ビー・エムの西村雅史氏と伊東伸泰氏に心から感謝する．また，本論文で報告している研究は文部省科学研究費補助金 (課題番号 00093069) の助成を受けている．ここに感謝の意を表する．

参考文献

Brill, E. (1992). "A Simple Rule-Based Part of Speech Tagger." In *Proceedings of the Third Conference on Applied Natural Language Processing*, pp. 152–154.

Brill, E. (1994). "Some Advances in Transformation-Based Part of Speech Tagging." In *Proceedings of the Twelveth National Conference on Artificial Intelligence*, pp. 722–727.

Brill, E. (1995). "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging." *Computational Linguistics*, **21** (4), 543–565.

Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). "Class-Based n-gram Models of Natural Language." *Computational Linguistics*, **18** (4), 467–479.

Charniak, E., Hendrickson, C., Jacobson, N., and Perkowitz, M. (1993). "Equations for Part-of-Speech Tagging." In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pp. 784–789.

Church, K. W. (1988). "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text." In *Proceedings of the Second Conference on Applied Natural Language Processing*, pp. 136–143.

Cutting, D., Kupiec, J., Pedersen, J., and Sibun, P. (1992). "A Practical Part-of-Speech Tagger." In *Proceedings of the Third Conference on Applied Natural Language Processing*, pp. 133–140.

de Marcken, C. G. (1990). "Parsing the LOB corpus." In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pp. 243–251.

Dermatas, E. and Kokkinakis, G. (1995). "Automatic Stochastic Tagging of Natural Language Texts." *Computational Linguistics*, **21** (2), 137–163.

DeRose, S. J. (1988). "Grammatical Category Disambiguation by Statistical Optimization." *Computational Linguistics*, **14** (1), 31–39.

Franz, A. (1997). *Automatic Ambiguity Resolution in Natural Language Processing.* Lecture Notes in Artificial Intelligence 1171. Springer.

Fu, K. S. (1974). *Syntactic Methods in Pattern Recognition*, Vol. 12 of *Mathematics in Science and Engineering.* Accademic Press.

Jelinek, F. and Mercer, R. L. (1980). "Interpolated estimation of Markov source parameters from sparse data." In *Proceeding of the Workshop on Pattern Recognition in Practice*, pp. 381–397.

Jelinek, F., Mercer, R. L., and Roukos, S. (1991). "Principles of Lexical Language Modeling

for Speech Recognition." In *Advances in Speech Signal Processing*, chap. 21, pp. 651–699. Dekker.

Kneser, R. and Ney, H. (1993). "Improved Clustering Techniques for Class-Based Statistical Language Modelling." In *Eurospeech*, pp. 21–23.

Kuhn, R. and de Mori, R. (1990). "A Cache-Based Natural Language Model for Speech Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12** (6), 570–583.

McMahon, J. G. and Smith, F. J. (1996). "Improving Statistical Language Model Performance with Automatically Generated Word Hierarchies." *Computational Linguistics*, **22** (2), 217–247.

Merialdo, B. (1994). "Tagging English Text with a Probabilistic Model." *Computational Linguistics*, **20** (2), 155–171.

Nagata, M. (1994). "A Stochastic Japanese Morphological Analyzer Using a Forward-DP Backward-A* N-Best Search Algorithm." In *Proceedings of the 15th International Conference on Computational Linguistics*, pp. 201–207.

Ney, H. (1984). "The Use of One-Stage Dynamic Programming Algorithm for Connected Word Recognition." *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **32** (2), 263–271.

Ney, H., Essen, U., and Kneser, R. (1994). "On Structuring Probabilistic Dependences in Stochastic Language Modeling." *Computer Speech and Language*, **8**, 1–38.

Schütze, H. and Singer, Y. (1994). "Part of Speech Tagging Using a Variable Memory Markov Model." In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 181–187.

Weischedel, R., Meteer, M., Schwartz, R., Ramshaw, L., and Palmucci, J. (1993). "Coping with Ambiguity and Unknown Words through Probabilistic Models." *Computational Linguistics*, **19** (2), 359–382.

黒橋禎夫・長尾眞 (1997). "京都大学テキストコーパス・プロジェクト." In *Proceedings of the Third Annual Meeting of the Association for Natural Language Processing*, pp. 115–118.

竹内孔一・松本裕治 (1995). "HMM による日本語形態素解析システムのパラメータ学習." 情報処理学会研究報告.

永田昌明 (1995). "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.

永田昌明 (1996). "単語頻度の期待値に基づく未知語の自動収集." 情報処理学会研究報告, 96-NL-116 巻.

日本電子化辞書研究所 (1993). EDR 電子化辞書仕様説明書.

春野雅彦・松本裕治 (1996). “文脈木を利用した形態素解析.” 情報処理学会研究報告.

丸山宏, 荻野紫穂, 渡辺日出雄 (1991). “確率的形態素解析.” 日本ソフトウェア科学会第 8 回大会論文集, pp. 177–180.

松本裕治, 黒橋禎夫, 山地治, 妙木裕, 長尾真 (1997). 日本語形態素解析システム JUMAN 使用説明書 version 3.2. 京都大学工学部長尾研究室.

森信介・長尾眞 (1995). “n グラム統計によるコーパスからの未知語抽出.” 情報処理学会研究報告.

略歴

森 信介: 1995 年京都大学大学院工学研究科電気工学第二専攻修士課程修了. 同年, 同大学大学院博士後期課程進学. 学術振興会特別研究員 (1997 –). 計算言語学の研究に従事. 情報処理学会会員.

長尾 眞: 1959 年京都大学工学部電子工学科卒業. 工学博士. 京都大学工学部助手, 助教授を経て, 1973 年より京都大学工学部教授. 国立民族学博物館教授を兼任 (1976 – 1994). 京都大学大型計算機センター長 (1986 – 1990), 日本認知科学会会長 (1989 – 1990), パターン認識国際学会副会長 (1982 – 1984), 日本機械翻訳協会初代会長 (1991 – 1993), 機械翻訳国際連盟初代会長 (1991 – 1993). 電子情報通信学会副会長 (1993 – 1995). 情報処理学会副会長 (1994 – 1996). 京都大学附属図書館長 (1995 – 1997). 京都大学大学院工学研究科長 (1997). 京都大学総長 (1997 –). パターン認識, 画像処理, 機械翻訳, 自然言語処理等の分野を並行して研究.

(1997 年 8 月 1 日 受付)
(1997 年 9 月 30 日 再受付)
(1997 年 10 月 24 日 採録)