クラスに基づく言語モデルのための単語クラスタリング

森 信介[†] 西村雅史[††] 伊東伸泰[††]

本論文では，クラスに基づく n-gram モデルのための単語クラスタリングの方法について述べる．クラスタリングの基準は，n-gram モデルの推定に用いるコーパスとは別に用意したコーパスのエントロピーであり，探索方法は貪欲アルゴリズムに基づいている．このため，局所的にではあるが最適な単語のクラス分類がクラス数をあらかじめ決めることなく得られる．この方法を日本語コーパスに適用して，59,956 個の単語をクラスに分類した結果，5,974 個のクラスが得られた．このクラス分類からクラスに基づく n-gram モデルを構成し，テストセットパープレキシティを計算すると 146.4 であり，同じ学習コーパスから得られる単語に基づく n-gram モデルによる値（153.1）や人間が与えた品詞に基づく n-gram モデルによる値（392.4）よりも低い値となった．

Word Clustering for Class-based Language Models

SHINSUKE MORI,[†] MASAFUMI NISHIMURA[††] and NOBUYASU ITOH[††]

In this paper we describe a word clustering method for class-based n-gram model. The measurement for clustering is the entropy on a corpus different from the corpus for n-gram model estimation and the search method is based on the greedy algorithm. For this reason this method gives us an optimum word classification without giving the number of classes. We applied this method to a Japanese corpus and classified 59,956 words. As the result we got 5,974 classes. The test set perplexity of the class-based n-gram model estimated from a training corpus with this classification was 146.4, which is less that of the word-based n-gram model (153.1) and that of the part-of-speech-based n-gram model (392.4) estimated from the same training corpus.

1. はじめに

統計的手法を用いた音声認識などでは，パラメータの推定や実装が容易とされているマルコフモデルを用いることが一般的である．このモデルでは，状態を単語に対応させて，コーパスから状態遷移確率を推定する（単語に基づく n-gram モデル）[1]．しかし，単語を状態に対応させると状態数は語彙数の n 乗と等しくなり，現在一般的に入手可能な量のコーパスでは，状態遷移確率を高い精度で推定できない．その結果，訓練データ以外のデータに対しては，単語列を正確に復元することができない．この問題に対処するため，クラスと呼ばれる単語のグループを 1 つの状態に対応させる方法が提案されている[2]．この手法によればモデルの訓練データが十分でない場合にも，性質の似た単語をグループ化することにより状態数が減少し，未観測の単語列に対するモデルの信頼性が向上すると考えられる．さらに副次的な効果として，状態遷移確率の記憶量が減少する．このようなモデルを，単語に基づく n-gram モデルに対して，クラスに基づく n-gram モデルと呼ぶ．

クラスに基づく n-gram モデルにおいて問題となるのは，単語列を復元するという課題に対して最適なクラス分類を求めること（以下，クラスタリングと呼ぶ）である．英語を対象とした研究として文献 2) や 3) などがある．これらの研究では，クラス数をあらかじめ決めておき，それぞれの単語は唯一のクラスに属すると仮定したうえで，遷移確率の推定に用いるコーパスのエントロピーを最小にするという条件でクラスタリングしている．しかし，この基準による最適解は個々の単語をクラスとするクラス分類である．この理由は，この基準ではどのような単語の組合せに対しても，それらを同一視することで情報を損失することはあっても獲得することはないことである．そこで，クラス数をあらかじめ決めておき，そのクラス数になるまで情

† 京都大学工学研究科電子通信工学専攻
Department of Electronics and Communication, Kyoto University
†† 日本アイ・ビー・エム株式会社東京基礎研究所
Tokyo Research Labolatory, IBM Japan, Ltd.

報の損失が最小である併合を繰り返すことでクラスタリングを行っている．この方法の最大の問題点は，確率的言語モデルの最終的な評価基準であるクロスエントロピー（テストセットパープレキシティ）とは異なる基準を用いているため，最終的に得られるクラスに基づく言語モデルが単語に基づく言語モデルよりも良くなることが必ずしも期待できないことである．実際，文献 2) で報告されている実験結果では，クラスに基づく言語モデルのテストセットパープレキシティが単語に基づく言語モデルのそれよりも大きい値となっている．

本論文では，クラス推定のためのコーパスを単語 n-gram モデルの推定用のコーパスとは別に用意し，異なるクラス分類の評価基準をこの別に用意されたコーパスのパープレキシティとすることを提案する．この方法の利点は，確率的言語モデルの最終的な評価基準を模倣していることである．つまり，最終的な評価の対象であるテストセットを参照することはできないので，その代わりにテストセットを学習データの一部を使って模倣し，それに対して計算したテストセットパープレキシティを評価基準として単語のクラス分類を求めるので，最終的に得られるクラスに基づく言語モデルが単語に基づく言語モデルよりも良くなることが期待できるという点である．具体的には，削除補間[4]のように学習コーパスの一部分で頻度の計数を行い，残りの部分に対してクラス関数の変更の効果を評価する．この方法より，テストコーパスに対して最適となるであろうクラス関数に，きわめて近いクラス関数を得ることが期待できる．このような利点に加えて，このような評価基準を採用した場合，従来の基準からは導き出せなかった停止条件を評価基準から自然に導き出せるという利点もある．これは，我々の基準ではクラス分類の変化によってパープレキシティが増加することもあれば減少することもあることによる．たとえば，我々の行った実験では，すべての単語が別々のクラスに属している状態で，単語 n-gram モデルという観点で類似している「や/助詞」と「の/助詞」とを併合するとパープレキシティは減少したが，類似していない「や/助詞」と「，/記号」とを併合するとパープレキシティは増加した．我々の採用した探索方法はボトムアップ型であるが，このようにクラスを併合していく過程でどのような併合もパープレキシティを減少させない状態に到達すると，この状態を解としてクラスタリングを終了させる．このようにきわめて自然に停止条件を設定できる．

クラス推定のためのコーパスを単語 n-gram モデルの推定用のコーパスとは別に用意するというアイデアは，Kneser ら[5]によってすでに提案されている．しかし，この論文で報告されている実験に用いられている方法は，初期値となる言語モデルに特別な仮定をして計算量を減らしており，結果として元となるアイデアの近似となっている．この文献では，以下の 4 つをドイツ語と英語に対して適用した結果を報告している．

(1) 単語を単位とした bi-gram
(2) 人間が与えた品詞を単位とした bi-gram
(3) 提案手法のクラスタリングの結果得られたクラスを単位とした bi-gram
(4) 従来手法☆のクラスタリングの結果得られたクラスを単位とした bi-gram

ドイツ語に対する実験結果では，人間が与えた品詞を用いた場合が最も良い結果を与えている．もし，この文献で提案されている手法が本当に有効であるなら，人間が与えた品詞 n-gram を初期状態として，提案したクラスタリングを再度行えば，さらにパープレキシティの低いモデルが得られると思われるが，文献にはそのような考察あるいは実験については触れられていない．また，英語に対する実験結果では提案手法の結果と従来手法の結果はほとんど同じとなっている．これは元となるアイデアが必ずしも有効には働いていない結果であると考えることもできる．さらに，この文献で報告されている実験結果はこのように不安定なので，この手法を日本語などに適用した場合どのような結果が得られるかが予測できないという問題もある．

これに対して，本論文では元となるアイデアを直接実現し，日本語に対する実験により得られた結果を報告する．また，対象実験として行った，単語を単位とした場合と人間の付与した品詞を単位とした場合のモデルによるテストセットパープレキシティの計算結果も報告する．実験には EDR コーパス[6]を用いた．59,956 個の単語を対象としてクラスタリングした結果，5,974 個のクラスが得られた．このクラスに基づく bi-gram モデルのテストセットパープレキシティの値は 146.4 であった．これは，単語に基づく bi-gram モデルによる値（153.1）や人間の付与した品詞に基づく bi-gram モデルによる値（392.4）よりも低い．これら結果は，我々の提案する方法によって，単語に基づく言語モデルから，記憶容量とパープレキシティの両方において優れている言語モデルが得られることを意味する．

☆ n-gram モデルの推定用のコーパスがクラス推定のためのコーパスと同じである方法

2. クラスに基づく n-gram モデル

この章では，確率的モデルによる音声認識に用いられる言語モデルとして，まず n-gram モデルについて簡単に説明し，次にその一般化であるクラスに基づく n-gram モデルを説明する．

2.1 n-gram モデル

確率モデルによる音声認識とは，ある音響特徴量の時系列 $\mathcal{S}$ が与えられたときに，それがある単語列 $\mathcal{W}$ として認識される確率 $p(\mathcal{W}|\mathcal{S})$ を最大にする単語列 $\hat{\mathcal{W}}$ を求めることである．条件付き確率 $p(\mathcal{W}|\mathcal{S})$ はベイズ則を用いることで以下のように書き換えられる．

$$p(\mathcal{W}|\mathcal{S}) = \frac{p(\mathcal{S}|\mathcal{W})p(\mathcal{W})}{p(\mathcal{S})}$$

この式の分母は $\mathcal{W}$ によらないので，求める単語列は以下の式で与えられる．

$$\begin{aligned}\hat{\mathcal{W}} &= \mathop{\mathbf{argmax}}_{\mathcal{W}} \frac{p(\mathcal{S}|\mathcal{W})p(\mathcal{W})}{p(\mathcal{S})} \\ &= \mathop{\mathbf{argmax}}_{\mathcal{W}} p(\mathcal{S}|\mathcal{W})p(\mathcal{W})\end{aligned}$$

この式において，$p(\mathcal{S}|\mathcal{W})$ は音響モデルと呼ばれ，$p(\mathcal{W})$ は言語モデルと呼ばれる．$\mathcal{W} = w_1 w_2 \cdots w_l$ とすると，この単語列の出現確率は次の式で与えられる．

$$\begin{aligned}p(\mathcal{W}) &= p(w_1 w_2 \cdots w_l) \\ &= \prod_{i=1}^{l} p(w_i|w_1 w_2 \cdots w_{i-1})\end{aligned}$$

この式は，ある時点 i での単語 w_i の出現確率は最初の時点から時点 $i-1$ までのすべての単語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 個の単語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．

$$\begin{aligned}&p(w_i|w_1 w_2 \cdots w_{i-1}) \\ &\quad \approx p(w_i|w_{i-k} w_{i-k+1} \cdots w_{i-1})\end{aligned}$$

一般に，確率 $p(w_i|w_{i-k} w_{i-k+1} \cdots w_{i-1})$ 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 を単語列のコーパスにおける頻度として，以下の式で与えられる．

$$\begin{aligned}&p(w_i|w_{i-k} w_{i-k+1} \cdots w_{i-1}) \\ &\quad = \frac{N(w_{i-k} w_{i-k+1} \cdots w_i)}{N(w_{i-k} w_{i-k+1} \cdots w_{i-1})} \\ &\quad = \frac{N(w_{i-k} w_{i-k+1} \cdots w_i)}{\sum_w N(w_{i-k} w_{i-k+1} \cdots w_{i-1} w)}\end{aligned}$$

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

以上のことから，n-gram モデルを用いる場合，統計的手法に基づく音声認識における言語モデルの課題は，以下のように定式化される．

(1) コーパスに対して n-gram の頻度を求め，マルコフ情報源のパラメータを推定する．

(2) 入力単語列 $\mathcal{W}$ に対して確率 $p(\mathcal{W})$ を計算する．

実際の音声認識では，音響モデルと言語モデルの確率の積が最大となる単語列が解である．このような単語列の探索には，ワン・パス・アルゴリズム[7]やスタック・デコーディング[8]が用いられる．

以上に述べた確率的言語モデルは十分一般的であり，音声認識以外の確率を用いる認識系の言語モデルとして用いることができる．

2.2 クラスに基づくモデルへの一般化

単語に基づく n-gram モデルでは，i 番目の単語の予測に直前の長さ $n-1$ の単語列を用いる．このとき，i 番目の単語の確率分布を，長さ $n-1$ の単語列のすべての組合せに対して個々に推定しておき，認識のときに用いる．しかし，これらの直前の単語列のいくつかは，次の単語を予測するという目的においては区別する必要がないという場合がある．このような場合には，直前の事象を一定の長さのすべての単語列に分類することは，不必要に直前の事象を区別していることになる．その結果，限られたコーパスにおける出現回数を減少させ，推定される確率値の信頼性の低下を招く．Bahl ら[9]は，直前の事象を情報理論的に見て有効な限りにおいて分類することを提案している．しかし，この問題に対しての最適な事象の分類の計算には膨大な計算時間が必要であり，実際には貪欲アルゴリズムを用いて準最適解を求めることで妥協せざるをえない．評価としてこの文献で報告されているのは，20 番目以前の単語から 21 番目の単語を予測するというきわめて限られた問題の実験結果だけである．そこで，解空間を限定したり，探索の順序を制御して最適解により近い準最適解をより速く計算することが問題となる．解空間の限定の例として，可変長マルコフモデル[10]では，単語列 $w_{i-k+1} w_{i-k+2} \cdots w_{i-1}$ が区別されている場合にのみ単語 w_{i-k} を区別するか否かを計算対象とする．以下で説明するクラスに基づく n-gram モデル[2]では，あらかじめ単語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見なして分類する．このモデルでは，次の単語を直接予測するのではなく，次のクラスを予測したうえで次の単語を予測する．

$$p(\mathcal{W}) = \prod_{i=1}^{n} p(w_i|w_1 w_2 \cdots w_{i-1})$$

$$p(w_i|w_1 w_2 \cdots w_{i-1}) = p(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1})p(w_i|c_i)$$

単語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 $p(w_i|c_i)$ の値は，コーパスから最尤推定することで得られる．

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

$$p(w_i|c_i) = \frac{N(w_i, c_i)}{N(c_i)}$$

この式において，単語からクラスへの写像が全単射であれば，単語に基づく n-gram モデルと等価になること分かる．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．

2.3 低頻度事象への対処

前節で述べたように，n-gram モデルのパラメータ推定には，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[11]．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いう．

- 単語に基づく n-gram モデル

$$p'(w_i|w_{i-k}w_{i-k+1}\cdots w_{i-1}) = \sum_{j=0}^{k} \lambda_j^w p(w_i|w_{i-j}w_{i-j+1}\cdots w_{i-1}) \quad (1)$$

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

- クラスに基づく n-gram モデル

$$p'(c_i|c_{i-k}c_{i-k+1}\cdots c_{i-1}) = \sum_{j=0}^{k} \lambda_j^c p(c_i|c_{i-j}c_{i-j+1}\cdots c_{i-1}) \quad (2)$$

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

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（k 通り）にわたって行い，その平均値をとるという方法である．

同様の考えに基づいて，複数のクラスに基づく n-gram モデルの間での補間も提案されている[12]．これは，クラスに基づく n-gram モデルが h 個あるとし，それぞれのクラスを $c^1, c^2, \ldots, c^h$ とすると，以下の式で表される．

$$p'(w_i|w_{i-k}w_{i-k+1}\cdots w_{i-1}) = \sum_{j=1}^{h} \mu_j p(c_i^j|c_{i-k}^j c_{i-k+1}^j \cdots c_{i-1}^j)p(w_i|c_i^j)$$

$$\text{ただし } 0 \le \mu_j \le 1, \sum_{j=1}^{h} \mu_j = 1$$

係数 μ_j 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．

3. クラスタリング

この章では，まず，単語とクラスの対応関係を定義する．次に，クラスタリングの基準となる目的関数について述べる．最後に，クラス関数の探索方法について述べる．

3.1 単語とクラスの対応関係

前章で述べたように，クラスは単語の集合である．本論文ではさらに，単語は唯一のクラスに属することを仮定している．このとき，クラスの集合は単語の集合の直和分割となっているので，単語とクラスの対応関係 F は，W，C をそれぞれ単語の集合とクラスの集合とすると，関数 $f: W \mapsto C$ 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☆．

$$W = \bigcup_{w \in W} f(w)$$

$\forall w \in W$ に対し $w \in f(w)$

$$f(w_1) \neq f(w_2) \Rightarrow f(w_1) \cap f(w_2) = \phi$$

単語とクラスの対応関係に対して，以下の関数を定義する．

- 単語の移動を表す関数

$$move: F \times W \times C \mapsto F$$

$move(f, w, c)$ は，単語とクラスの関係 f に対して単語 w をクラス c に移動した結果得られる単

☆ f の値は単語の集合である（例：$f(w_1) = \{w_1, w_2, w_3\}$）．

語とクラスの関係を返す関数であり，以下のように定義される☆．

```
define move(f, w, c)
  f(w) := f(w) − {w}
  c := c ∪ {w}
  return f
```

3.2 目的関数

すでに述べたように，単語クラスタリングの目的はより良い言語モデルを構成することである．我々は，言語モデルの良さの尺度としてパープレキシティ PP を用いることとした．これは，各単語が等確率に選ばれると仮定した場合の後続可能単語数の幾何平均を表しており，確率的言語モデル M と文の列 $\mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_n$ の関数であり，以下の式で定義される．

$$H(M, \mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_n) = \frac{\sum_{i=1}^{n} -\log p_M(\mathcal{W}_i)}{\sum_{i=1}^{n} |\mathcal{W}_i|}$$

ただし $|\mathcal{W}|$ は $\mathcal{W}$ の単語数

$$PP(M, \mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_n) = 2^{H(M, \mathcal{W}_1, \mathcal{W}_2, \ldots, \mathcal{W}_n)}$$

次の章で述べる実験では，補間係数の推定のみならず，クラスタリングにも削除補間と同じ技術を用いている．つまり，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，残りの部分で上記の評価関数の値を計算するということをすべての組合せにわたって行い，その平均値を全体の評価関数の値とするという方法である．よって，クラス関数の評価関数は以下の式で与えられる平均テストセットパープレキシティである．

$$\overline{PP} = \left\{ \prod_{i=1}^{k} PP(M_i, \mathcal{C}_i) \right\}^{\frac{1}{k}} \tag{3}$$

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 n-gram モデル（補間係数の推定も含む）であり，$\mathcal{C}_i$ は i 番目の部分コーパス（文の列）を表す．

本論文で問題としているのは，確率的言語モデルとしてクラスに基づく n-gram モデルを用いた場合の単語のクラスタリングである．この場合，コーパス（文の列）は一定であり，確率的言語モデル M は単語とクラスの関係 F にのみ依存する．したがって，平均テストセットパープレキシティは，単語とクラスの関係の関数と見なすことができる．パープレキシティの値域は正の実数であるから，平均テストセットパープレキシティの値域も正の実数であり，これにより単語とクラスの関係に全順序関係を与えることができる．定義から明らかなように，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．以上のことから，クラスタリングの目的は，式 (3) で定義される平均テストセットパープレキシティを最小化する単語とクラスの関係を求めることであるといえる．

Brown ら[2]や Ney ら[3]も，クラスタリングの基準としてパープレキシティを用いているが，計算の対象とするコーパスは確率値の推定に用いるコーパスと同じである．我々は，これらの先行研究と異なり，クラスタリングのためのコーパスを確率値の推定用のコーパスとは別に用意することとした．この利点は，単語とクラスの関係を変えた場合に，パープレキシティが増加することもあれば減少することもあるので，閾値を設ける代わりに減少する場合のみクラスの変更を施すことができるということである．

3.3 アルゴリズム

クラスタリングの解空間はあらゆる可能な単語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，これらすべてに対してパープレキシティを計算し，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．パープレキシティ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持っているので，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あきらめ，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とおりである．なお，$\overline{PP}$ は式 (3) で与えられる平均テストセットパープレキシティである．

```
W を頻度の降順に並べ w_1, w_2, ..., w_n とする
foreach i (1, 2, ···, n)
  c_i := {w_i}
  f(w_i) := c_i
foreach i (2, 3, ···, n)
  c := argmin_{c∈{c_1, c_2, ..., c_{i-1}}} PP(move(f, w_i, c))
  if (PP(move(f, w_i, c)) < PP(f)) then
    f := move(f, w_i, c)
```

計算量は，2 番目の **foreach** での繰返しの回数は単語数 $|W|$ に比例し，**argmin** での繰返しの回数はクラ

☆ 正確には，同じクラスに属するすべての単語に対して f の値を改めなければならない．

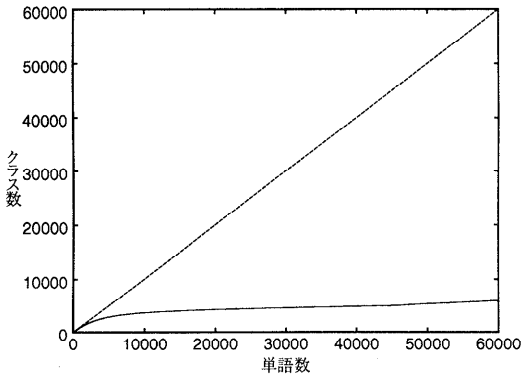


図1 単語数とクラス数の関係
Fig. 1 The relation between the number of words and that of classes.

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，すべ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，すべての単語が同一のクラスとなる場合に最小 ($|C|=1$) となる．したがって，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1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．単語数とクラス数の関係については考察を行っておらず，次章で述べる実験の結果を**図1**に掲げるにとどめる．計算時間は単語数とクラス数の関係を表す曲線と横軸に囲まれた部分の面積に比例することになるが，このグラフを見ると実際にはかなり線形に近いことが分かる．

頻度の高い単語から移動を試みることとしているのは，頻度の高い単語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，さらに単語の移動を試みること[3]やクラスの併合を試みること[2]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テストセットパープレキシティ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．

4. 実験結果

我々は，前章で説明したクラスタリングアルゴリズムを評価するために，同じ学習コーパスから推定された単語に基づく bi-gram モデルとクラスに基づく bi-gram モデルを，テストセットパープレキシティで比較した．この章では，この実験の条件と結果を提示し，考察を行う．

表1 コーパス
Table 1 Corpus.

コーパス	文数	形態素数
学習コーパス	187,022	4,595,786
テストコーパス	20,780	509,261

4.1 実験の条件

実験には EDR コーパス[6]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テストセットパープレキシティを求めるということを9通り行って得られる平均パープレキシティを評価基準とする．それぞれのコーパスに含まれる文と形態素の数は**表1**のとおりである．登録語は，2個以上の学習コーパスに現れる 59,956 個の単語とした．単語に基づく 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）と文区切りに対応する状態を持つ．同様に，クラスに基づくモデルは，登録語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．

単語に基づく bi-gram モデルとクラスに基づく 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を計算した．また，あらかじめ与えられた品詞をクラス分類とした場合のモデル（以下，品詞 bi-gram モデル）を構成し，同じテストコーパスに対してパープレキシティを計算した．それぞれの言語モデルの構成の手順は以下のとおりである．

- 単語 bi-gram モデル
  - (1) 削除補間により式 (1) の補間係数を推定
  - (2) すべての学習コーパスを対象に単語 bi-gram と単語 uni-gram を計数
- 品詞 bi-gram モデル
  - (1) 削除補間により式 (2) の補間係数を推定
  - (2) すべての学習コーパスを対象に品詞 bi-gram と品詞 uni-gram を計数
- クラス bi-gram モデル
  - (1) 削除補間により式 (1) の補間係数を推定

(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(2)の補間係数を推定
(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

パープレキシティの計算は，テストコーパスの未知語（登録語以外）を品詞ごとに異なる特別な記号に置き換えて行った．各品詞の未知語の文字列の出現確率は，品詞ごとに用意した未知語モデルによって与えられる．この未知語モデルは，学習コーパスから推定された文字 bi-gram モデルである．この部分は，どちらのモデルの場合も同じなので，パープレキシティの比を考える限り本質的な影響はない．

4.2 結果と考察

表 2 は各モデルのテストコーパスのパープレキシティである．クラスに基づく bi-gram モデルは単語に基づく bi-gram モデルや品詞に基づく bi-gram よりも低いパープレキシティとなっている．このことから，クラスに基づく bi-gram モデルが単語に基づく bi-gram モデルや品詞に基づく bi-gram よりも，予測力という点で良い言語モデルであると結論できる．Brown らの先行研究[2]や Ney らの先行研究[3]では，得られたクラス n-gram モデルの状態数は当然減少しているが，テストセットパープレキシティが上昇し予測力という点で良い言語モデルとなっていない．

なお，文献 5) は我々と同じアイデアに基づいており，予測力という点では，我々の結果と同様に，従来の方法を上回る結果が期待されるが，文献において報告されている実験結果では，提案手法により得られたクラスを単位としたモデルのテストセットパープレキシティは，ドイツ語を対象とした場合，人間が与えた品詞を単位としたモデルよりも劣っており，英語を対象とした場合，従来手法とほとんど同じであり，必ずしも予測されるような結果にはなっていない．しかしながら，我々の提案する基準では，予測力という点でより良い言語モデルとなることが，少なくとも日本語に対して，実験的に傍証されたといえる．

得られたクラスに基づくモデルの状態数は，単語に基づくモデルの状態数の約 10.0%である．このことは，記憶容量という点でも，クラスに基づく bi-gram モデルが単語に基づく bi-gram モデルよりも優れていることを示す．実験に用いた bi-gram モデルの状態遷移表の大きさは，配列による単純な実装を仮定すると，状態数の 2 乗に比例するので記憶容量は 0.998%に縮小する．また，非零要素の数を計数した結果，この数は単語に基づく bi-gram モデルでは 724,870 であり，クラスに基づく bi-gram モデルでは 245,283 であった．よって，ハッシュやリンクリストなどを用いて非零要素だけを記憶する場合の記憶容量の縮小率は 33.8%程度と推定される．現在実用となっている音声認識などでは，単語 tri-gram を用いるのが一般的である．この場合，記憶容量の差はさらに拡大するだろう．なお，品詞に基づくモデルは，記憶容量という点では非常に良いが，パープレキシティが高過ぎて実用的であるとはいえない．

表 2 実験結果
Table 2 An experimental result.

言語モデル	状態数	クラス数	パープレキシティ
単語 bi-gram	59,972	59,956	153.1
品詞 bi-gram	31	15	392.4
クラス bi-gram	5,990	5,974	146.4

クラス数に 15 品詞の未知語に対応する記号と文区切り記号の合計 16 を加算すると状態数になる．

付録 A.1 は得られたクラスタの例である．多くのクラスタが，**クラスタ 1** のように我々の言語直観に照らし合わせて，納得できるクラスタであった．このクラスタの「キロ/名詞」で示されるように，品詞の異なる単語が同一のクラスと見なされている場合も観測された．一方，**クラスタ 2** のように我々の言語直観に合致しないクラスタもあった．これは，我々が行った単語クラスタリングは，クラス bi-gram モデルの改善という観点からのクラスタリングであることと，得られた単語の分類が準最適解であることを考えると特に不自然であるとはいえないであろう．

5. おわりに

本論文では，クラスに基づく n-gram モデルを仮定して，単語の最適なクラス分類を求めるアルゴリズムについて述べた．このアルゴリズムは，クラス推定のためのコーパスを単語 n-gram モデルの推定用のコーパスとは別に用意するというアイデアに基づいている．このアルゴリズムを実装し，EDR コーパス[6]を用いて，59,956 個の単語を対象として bi-gram モデルを構成し，単語クラスタリングした結果，5,974 個のクラスが得られた．このクラスに基づくテストコーパスのパープレキシティを計算した結果，クラスに基づく n-gram モデルが単語に基づく n-gram モデルよりも，少なくとも日本語においては，予測力と記憶容量の両方の点で良い言語モデルとなることが分かった．

今後の課題として，Pereira らが動詞と目的語の関係に着目した名詞のクラスタリング[13]に対して適用しているような，各単語が複数のクラスに属することを

考慮に入れることや，クラスに基づくモデルを可変長マルコフモデル[10]と融合することなどがあげられる．

参考文献

1) Shannon, C.E.: Prediction and Entropy of Printed English, *Bell System Technical Journal*, Vol.30, pp.50–64 (1951).
2) Brown, P.F., Pietra, V.J.D., de Souza, 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).
3) 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).
4) Jelinek, F. and Mercer, R.L.: Interpolated estimation of Markov source parameters from sparse data, *Proc. Workshop on Pattern Recognition in Practice*, pp.381–397 (1980).
5) Kneser, R. and Ney, H.: Improved Clustering Techniques for Class-Based Statistical Language Modelling, *Eurospeech*, pp.21–23 (1993).
6) 日本電子化辞書研究所：EDR電子化辞書仕様説明書 (1993).
7) Ney, H.: The Use of One-Stage Dynamic Programming Algorithm for Connected Word Recognition, *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.32, No.2, pp.263–271 (1984).
8) 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).
9) Bahl, L., Brown, P., de Souza, P. and Mercer, R.: A Tree-Based Statistical Language Model for Natural Language Speech Recognition, *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol.37, No.7, pp.1000–1008 (1989).
10) Ron, D., Singer, Y. and Tishby, N.: The power of amnesia: Learning probabilistic automata with variable memory length, *Machine Learning Special Issue on COLT94* (1996).
11) 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).
12) McMahon, J.G. and Smith, F.J.: Improving Statistical Language Model Performance with Automatically Generated Word Hierarchies, *Computational Linguistics*, Vol.22, No.2, pp.217–247 (1996).
13) Pereira, F., Tishby, N. and Lee, L.: Distributional Clustering of English Words, *Proc. 31st Annual Meeting of the Association for Computational Linguistics*, pp.183–190 (1993).

付　録

A.1 得られたクラスタの例

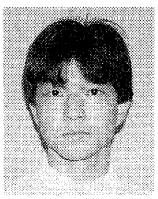
クラスタ 1

[人/接尾語 件/接尾語 カ所/接尾語 平方メートル/接尾語 点/接尾語 隻/接尾語 頭/接尾語 匹/接尾語 戸/接尾語 基/接尾語 世帯/名詞 校/接尾語 ヘクタール/接尾語 羽/接尾語 棟/接尾語 元/接尾語 票/接尾語 世帯/接尾語 キロ/名詞 リットル/接尾語 席/接尾語 桁/接尾語 巻/接尾語 編/接尾語 km/接尾語 曲/接尾語 mm/接尾語 マルク/接尾語 K/接尾語 品目/接尾語 床/接尾語 ミクロン/接尾語 ヵ所/接尾語 つがい/名詞 ズロチ/接尾語 首/接尾語 筆/接尾語 NZドル/接尾語 KHz/接尾語]

クラスタ 2

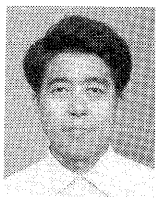
[の/助詞 や/助詞 および/接続詞 及び/接続詞 ないし/接続詞 ならびに/接続詞 もしくは/接続詞 イコール/接続詞 たる/助動詞 らしい/接尾語 カタロニア/名詞 や/接尾語 質/接尾語 はじめ/接尾語 テラピア/名詞 中心/接尾語]

(平成 9 年 3 月 26 日受付)
(平成 9 年 9 月 10 日採録)



森　信介（学生会員）
昭和 45 年生まれ．平成 7 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．計算言語学の研究に従事．言語処理学会会員．

西村　雅史（正会員）
昭和 33 年生まれ．昭和 56 年大阪大学基礎工学部生物工学科卒業．昭和 58 年同大学大学院物理系博士前期課程修了．同年日本アイ・ビー・エム（株）入社．以来，東京基礎研究所において，音声認識などの音声言語情報処理の研究に従事．電子情報通信学会，日本音響学会各会員．



伊東　伸泰（正会員）

昭和57年大阪大学基礎工学部生物工学科卒業．昭和59年同大学大学院修士課程修了．同年日本アイ・ビー・エム（株）入社．東京基礎研究所に勤務．文字認識および音声認識についての言語処理研究に従事．