音声言語情報処理 6－1
（1995．5．25）

確率モデルによる後続単語予測と大語彙日本語連続音声認識

周 旻　　中川 聖一
{xiu, nakagawa}@say1gw.tutics.tut.ac.jp

豊橋技術科学大学 情報工学系
〒441 愛知県 豊橋市 天伯町 字雲雀ヶ丘1－1

概 要 多種類の確率モデルによる日本語のモデル化を検討し、それらの言語モデルの性能をATRの日本語対話データベース (ADD) を用いて評価した。確率正規文法、確率文脈自由文法、bigram、trigram、bigram-HMM及び人間の作成した文節文法を言語モデルとして、品詞と単語レベルでテキスト文の複雑さ (パープレキシティ/エントロピー) と後続品詞・単語の予測率及びそれらのモデルパラメータ数の比較、考察を行なった。また、語彙数が4699のタスクについて、bigram言語モデルで連続音声の認識評価実験を話者適応化モデルの条件で行なった。

和文キーワード： 確率モデル、隠れマルコフモデル、文脈自由文法、bigram、trigram、エントロピー、パープレキシティ、単語予測

Japanese word prediction and large vocabulary speech recognition based on stochastic language model

Min ZHOU　　Seiichi NAKAGAWA

Toyohashi University of Technology, Department of Information & Computer Sciences
Tempaku-cho, 441 Toyohashi, Aichi

Abstract A study on comparing different types of stochastic language models based on the ATR dialog database is given. The comparison of a part of speech/word-unit perplexity and prediction ability for the following words based on hidden Markov model, stochastic context-free grammar, bigram, trigram, and bigram-HMM is evaluated. By using the bigram model, we performed the continuous speech recognition of a 4699 word vocabulary size on the speaker adaptation models.

Key Words: stochastic model, HMM, context-free grammar, bigram, trigram, entropy, perplexity, word prediction

1 はじめに

自然言語処理や音声認識の研究分野では言語モデルの研究が重要な役割を荷なっている。テキスト解析では、タギング (tagging) が重要であるが、音声認識では後続単語の予測が重要である。タギングに関しては、最近統計的手法（例えば 2 重隠れマルコフモデル）を用いることによって 95% 以上の精度が得られている [1] [2] [3]。音声認識にとって重要な後続単語の予測に関しては、従来、文節文法や文脈自由文法を用いるものが代表的であった [4] [5]。しかし、多量のテキスト入力文や自然な発話文に対してはこのような文法の構築は難しく、これらに確率を導入したり、単語対文法、bigram, trigram を用いることが欧米では盛んに研究されている [6] [7]。現実的には、確率文脈自由文法と trigram (bigram) の併用が有望だと考えられている [8]。

言語モデルが音声認識システムの探索空間を縮小させることに伴って、音声認識の性能を改善できる。連続音声認識における言語処理は文法、意味、文脈などのさまざまな高次知識を用いることができるため、音響的な特徴のみによる認識結果を補正し、認識精度を向上させることができる。単語予測の正確さは言語モデルのエントロピー/パープレキシティでも評価できる。一般的に言うと、認識性能と言語の複雑さには密接な関係があり、部分文に後続できる単語数に依存する。従って、タスクが決められた時に、複雑さが低い言語モデルを構築するのは認識システムの性能にとって重要である。

Black は人間が作成した文法と機械的に作成した文法を比較検討している [12]。中川は文献 [11] で音声認識用の言語モデルとして機械で自動獲得した確率文法、人間の作成した文法から学習した確率文法、人間が作成した文法、機械が自動獲得した文法の順で良いだろうと推測している。

本報告では、多数の確率言語モデルを検討し、それらに基づく後続品詞/単語の予測確率の計算法を検討する。また、日本語の ATR の対話データベース (ADD) を用いて品詞と単語レベルでそれらのモデルの有効性を比較して、言語モデルのエントロピーと後続品詞/単語の予測率を考察した。最後に、連続認識アルゴリズムで大語彙の朗読発声の文認識を行なった。

2 確率言語モデル

代表的な言語モデルとして、確率正規文法 (HMM)、確率文脈自由文法 (SCFG)、bigram や trigram (N-Gram) などがある。

N-Gram モデル N-Gram 確率モデルは柔軟性が有って、様々なテキスト集合文に対して構成しやすい。しかし、言語の文脈情報があまり表現できず、長い距離にわたる文のシンタックスの情報をほとんど含んでない。

隠れマルコフモデル 正規文法 (オートマトン) は文脈自由文法の特殊な場合である。確率正規文法は left-to-right 型 HMM と本質的に等価で、Forward-Backward アルゴリズムによる最尤推定法や、Viterbi ベストパスの認識法が提案されてから、音声認識以外にも言語モデルとして良く用いられている。

文脈自由文法モデル 一方、文脈自由文法は自然言語の統合規則として良く用いられてきた。また文脈自由文法の効率の良い解析法が良く知られており、少し拡張することにより、より精度の良い自然言語モデルを構成することができると言われてきた。更に、規則に確率を付けることによって得られる確率文脈自由文法によって文の確率を算出することができるため、曖昧さを含む文に対しても最適な構文解析木が決められる利点がある。

品詞モデル 単語系列$\boldsymbol{w}_1^L = w_1 w_2 \cdots w_L$が与えられる時に、言語モデル$\boldsymbol{G}$ による部分文の生成確率 $P(\boldsymbol{w}_1^n)$ は次式のように表される。

$$P(\boldsymbol{w}_1^n) = \prod_{t=1}^{n} P(w_t|\boldsymbol{w}_1^{t-1}, \boldsymbol{G}) \tag{1}$$

しかし、実際にこの事後確率 $P(w_t|\boldsymbol{w}_1^{t-1})$ を直接求めるのは困難であるため、我々は品詞レベルの言語モデルに基づいて、その単語列の確率及びテキストのエントロピーと後続単語の予測率を求める。

単語列$\boldsymbol{w}_1^L$ に対応する品詞列を $\boldsymbol{c}_1^L = c_1 c_2 \cdots c_L$、品詞 c_t の条件付き予測確率を $P(c_t|\boldsymbol{c}_1^{t-1})$ とすれば、単語の予測確率を

(第一次近似式)

$$P(w_t|\boldsymbol{w}_1^{t-1}) = \sum_{\{\boldsymbol{C}\}} P(\boldsymbol{c}_1^{t-1}|\boldsymbol{w}_1^{t-1}) * P(w_t|\boldsymbol{c}_1^{t-1}) \tag{2}$$

あるいは

(第二次近似式)

$$P(w_t|\boldsymbol{w}_1^{t-1}) = \sum_{\{\boldsymbol{C}\}} P(\boldsymbol{c}_1^{t-1}|\boldsymbol{w}_1^{t-1}) * (\sum_{x=1}^{K} P(c_x|\boldsymbol{c}_1^{t-1}) * P(w_t|c_x)) \tag{3}$$

で近似する。ここで、K は品詞の数である。$\{\boldsymbol{C}\}$ は$\boldsymbol{w}_1^{t-1}$ の可能な品詞列の集合である。以後、$\boldsymbol{w}_1^{t-1}$ に対して最尤の$\boldsymbol{c}_1^{t-1}$のみを考慮する。即ち、$P(\boldsymbol{c}_1^{t-1}|\boldsymbol{w}_1^{t-1}) = 1$ と仮定する。

2.1 N-Gram モデル

シンボル系列$\boldsymbol{w}_1^n = w_1 w_2 \cdots w_n$が与えられる時に、N-Gram モデルによる確率 $P(\boldsymbol{w})$ は

$$P(\boldsymbol{w}) = \prod_{t=1}^{n} P(w_t|\boldsymbol{w}_{t-N+1}^{t-1})$$

N = 3 の時は trigram (N = 2 の時は bigram) と言う。そのモデルに対して文の生起確率は

$$P_{\text{trigram}}(\boldsymbol{w}) = \prod_{i=1}^{n} P(w_i|w_{i-2}, w_{i-1}) \tag{4}$$

$$P_{\text{bigram}}(\boldsymbol{w}) = \prod_{i=1}^{n} P(w_i|w_{i-1}) \tag{5}$$

ただし$P(w_1|w_{-1},w_0)=P(w_1)$、$P(w_2|w_0,w_1)=P(w_2|w_1)$。

しかし、大語彙音声認識システムの時に、真のtrigram確率$P(w_t|w_{t-2},w_{t-1})$を求めるのが大変なので、ここでは次の二通りで近似する。

(第一次近似式)

$$P(w_1^n)=\prod_{t=1}^{n}P(w_t|c_{t-2},c_{t-1}) \tag{6}$$

(第二次近似式)

$$P(w_1^n)=\prod_{t=1}^{n}\sum_{x=1}^{K}P(c_x|c_{t-2},c_{t-1})*P(w_t|c_x) \tag{7}$$

ただし、$P(c_x|c_{-1}c_0)=P(c_x)$、$P(c_x|c_0c_1)=P(c_x|c_1)$。なお、bigramモデルの場合もこれと類似な方法で求める。

ある決められたデータベースに対して、trigramモデルの確率の求め方は次式のように二つ組と三つ組の出現頻度$C(c_{i-2},c_{i-1})$と$C(c_{i-2},c_{i-1},c_i)$の比によって推定できる（最尤推定)。

$$P(c_i|c_{i-2},c_{i-1})=\frac{C(c_{i-2},c_{i-1},c_i)}{C(c_{i-2},c_{i-1})} \tag{8}$$

ところが、語彙数が多い時には、多量の訓練サンプルを用いても、多くの三組については、信頼できる統計量を求めるのに足りるほど出現することはなく、上式を用いても精度良く確率を推定することは難しい。これに対して、スムージングの手法が良く用いられる（削除補完法)。

$$P(c_i|c_{i-2},c_{i-1})=\alpha f(c_i|c_{i-2},c_{i-1})+\beta f(c_i|c_{i-1})+\gamma f(c_i) \tag{9}$$

ただし、$\alpha+\beta+\gamma=1$。

2.2 隠れマルコフモデル

音声認識に用いられるHMM Mは次の六つ組$M=(S, Y, A, B, \pi, F)$で定義される。ここでSは状態の有限集合。Yは出力シンボルの集合。$A(=\{a_{ij}\})$と$B(=\{b_{ij}(k)\})$はそれぞれ状態遷移確率を出力確率の集合。$\pi=\{\pi_i\}$は初期状態確率の集合。Fは最終状態の集合。

確率正規文法は音声認識に用いられるleft-to-right型隠れマルコフモデル(LRHMM)と等価である。HMMの最尤推定法(Forward-Backwardアルゴリズム)は訓練データからHMMのパラメータを自動学習することができる。

隠れマルコフモデルの場合に、前向き確率$\alpha(i,t)$(y_1^tが出力され、時刻tに状態iに達する確率)を用いて、部分文w_1^tの確率は次のように求められる。[1]

(第二次近似式)

$$\begin{aligned}P_{\text{HMM}_1}(w_1^t) &= \sum_{i=1}^{S}\alpha_i(w_1^t)\\ &= \sum_{i=1}^{S}\sum_{x=1}^{K}\alpha_i(w_1^{t-1}c_x)*P(w_t|c_x)\end{aligned} \tag{10}$$

$$\alpha_i(w_1^{t-1}c_x)=\sum_{j=1}^{S}\alpha_j(w_1^{t-1})a_{ji}b_{ji}(c_x)$$

ここでSはHMMの状態数、$\alpha_i(w_1^{t-1}c_x)$はHMMで列$w_1\cdots w_{t-1}c_x$が生成され、且つ時刻tに状態iに達する確率である。その$\alpha_i(w_1^{t-1}c_x)$はForwardアルゴリズムによって計算される。

2.3 Bigram駆動型HMM

BigramとHMMの統合であるbigram駆動型HMMはシンボルのbigram確率とHMMを合成することによって生成され、出力シンボルの確率は前時刻の条件つき確率で表現される[14]。品詞カテゴリの集合を$V=\{v_1,v_2,\cdots,v_K\}$とすれば、bigram-HMMの出力シンボル確率は次式の二通りで定義される。

$$b_{ij}(c_t|c_{t-1}) = \frac{p(c_t|c_{t-1})b_{ij}(c_t)}{\sum_{m=1}^{K}p(v_m|c_{t-1})b_{ij}(v_m)} \tag{11}$$

$$b_{ij}(c_t|c_{t-1}) = \frac{p(c_{t-1}c_t|ij)}{p(c_{t-1}|ij)} \tag{12}$$

ここで$b_{ij}(c_t)$は従来のHMMにおいて状態iから状態jの遷移でシンボルc_tが出力する確率で、$p(c_t|c_{t-1})$は全品詞データのグロバールbigram確率である[14]。式(11)は全データに対するbigramをHMMに結合する時の出力確率である。式(12)はViterbiベストパスによって状態ごとにセグメンテーションしたデータから求めたbigramをHMMに結合する方法である。つまり、テキストの局所的単語列の集合をHMMの状態数にクラスタリングし、それぞれのクラスのbigramを求めていることに対応する。

Bigram-HMMによる部分品詞列の生成確率は

(第二次近似式)

$$\begin{aligned}P_{\text{bi-HMM}_1}(c_1^t) &= \sum_{j=1}^{S}\alpha_j(c_1^t)\\ &= \sum_{j=1}^{S}\sum_{i=1}^{S}\alpha_i(c_1^{t-1})a_{ij}b_{ij}(c_t|c_{t-1})\end{aligned} \tag{13}$$

また、部分文の生成確率は

$$P_{\text{bi-HMM}_1}(w_1^t)=\sum_{x=1}^{K}P_{\text{bi-HMM}}(w_1^{t-1}c_x)P(w_t|c_x) \tag{14}$$

で求められる((10)式参照)。

以上のモデル化法は第二次近似した式(3)に対応するものなので、パラメータ数は割と少なくなるけれども、単語レベルの予測能力があまり良くならなかった。もう一つのモデル化法は第一次近似した式(2)に対応

[1] 入力が単語系列なので、1時刻進むと1単語を出力して遷移する。

するものである。即ち、bigram 確率は $P(w_t|c_{t-1})$ を用いて、HMM と結合する方法である。単語カテゴリの集合を$\boldsymbol{W} = \{w_1, w_2, \cdots, w_V\}$ とすれば、 $b_{ij}(w_t|c_{t-1}) = \sum_{x=1}^{K} b_{ij}(c_x|c_{t-1})P(w_t|c_x)$ の代わりに次式を用いる。

(第一次近似式)

$$b_{ij}(w_t|c_{t-1}) = \frac{p(w_t|c_{t-1})b_{ij}(w_t)}{\sum_{m=1}^{V} p(w_m|c_{t-1})b_{ij}(w_m)} \qquad (15)$$

なお、品詞の予測では一番効果のあった式(12)に対応する bigram-HMM$_2$ は trigram$_2$ よりもパラメータが多くなるので試みなかった。

2.4 確率文脈自由文法[7]

確率文脈自由文法(SCFG)は四つ組$G = (V_N, V_T, P, S)$で定義される。V_NとV_T は各々非終端記号と終端記号の集合、 P は文脈自由規則の集合

$$A \xrightarrow{f_r} \gamma \qquad (\sum f_r = 1)$$

ここで A は非終端記号 ($A \in V_N$)、γ は終端記号か非終端記号の列 ($\gamma \in (V_N \cup V_T)^*$) である。 S は全ての文を生成するための開始記号である。 f_r は書き換え規則の確率、文を導出する時に非終端記号 A が終端記号か非終端記号の系列γ に書き換えられる確率である。

この SCFG が求められれば、曖昧な文を解析する時、一番確率が高い解析木を正しい結果として選ぶことができる。また、同じタスク (同じ書換え規則の集合) における SCFG は CFG よりエントロピーが小さく、従って、音声認識における探索空間が小さくなり、認識率も向上する。

2.4.1 Inside-Outside による SCFG の学習[15]

訓練用データセット $\mathbf{O} = \{o_1, o_2, \cdots, o_T\}$ を用いて Inside-Outside アルゴリズムで SCFG を学習することは inside 確率 ($e_i(s,t)$) と outside 確率 ($f_i(s,t)$) を求めることに基づいて行う。

$$e_i(s,t) = P(i \stackrel{*}{\Rightarrow} o_s \cdots o_t / G) \qquad (16)$$

$$f_i(s,t) = P(S \stackrel{*}{\Rightarrow} o_1 \cdots o_{s-1}, i, o_{t+1} \cdots o_T / G) \qquad (17)$$

以上の句構造確率を求めてから、SCFG の確率の再推定を次式によって行なう。

$$\hat{a}(i \to jk) = \frac{\sum_{s=1}^{T-1} \sum_{t=s+1}^{T} w(s,t,i,j,k)}{\sum_{s=1}^{T} \sum_{t=s}^{T} v(s,t,i)}$$

$$\hat{b}(i \to m) = \frac{\sum_{t \in \{o_t = m\}} v(t,t,i)}{\sum_{s=1}^{T} \sum_{t=s}^{T} v(s,t,i)}$$

そこで、i, j, kは非終端記号、m は終端記号。w と v は以下のように計算される。

$$w(s,t,i,j,k) = \sum_{r=s}^{t-1} a(i \to jk) e_j(s,r) e_k(r+1,t) f_i(s,t)$$

$$v(s,t,i) = e_i(s,t) f_i(s,t)$$

品詞データから学習された SCFG を用いて、品詞列 (c_1^t) の生成確率は

$$P_{\mathrm{SCFG}}(c_1^t) = \sum_{i=1}^{N} e_i(1,t) \qquad (18)$$

従って、SCFG による単語列 w_1^t の生成確率は

(第二次近似式)

$$P_{\mathrm{SCFG}}(w_1^t) = \prod_{i=0}^{t-1} \sum_{x=1}^{K} P(c_x|c_1^i) * P(w_{i+1}|c_x) \qquad (19)$$

で求められる。

2.4.2 CKY 解析法による SCFG の学習

この手法は前もって文脈自由文法が知られている場合に、テキスト文の解析情報を利用しながら文法規則の確率を学習する方法である。ただし、CKY 法で構文解析を行なうために CFG は等価な *Chomsky* 形に変形しなければいけない。学習アルゴリズムは文献[16]を参考されたい。

3 確率言語モデルによる後続予測実験

言語モデルのパフォーマンスを比較するために、品詞レベルと単語レベルで以上の種々の確率言語モデルによる後続品詞と後続単語の予測及びモデルパラメータ数とエントロピーの関係などの比較、評価実験を行なった。

3.1 日本語対話データベース

ATR で作成された日本語の旅行案内に関する問い合わせの対話データベース (ADD) を用いて評価実験を行なった。間投詞を取り除いて、実際の文章に出現する単語の数は 4514 種類、品詞の数は 24 種類である。学習データとテストデータはそれぞれ 10504 文と 1073 文を使った。品詞は表 1 に示される通りの 24 種類に分類された。なお、エルゴディック HMM による ATR テキストデータベースのモデル化の検討は村上らによってなされている[18]。

表 1: ATR の対話データに用いる品詞

係助詞	形容詞	普通名詞	サ変名詞
代名詞	数　詞	副　詞	連体詞
接続詞	感動詞	助動詞	副助詞
接続助詞	格助詞	終助詞	接尾語
接頭語	補助動詞	固有名詞	形容名詞
本動詞	準体助詞	並立助詞	記　号

3.2 エントロピーとパープレキシティ

言語モデル G において、文（単語列）$w_i = w_1^{T_i}$ の出現確率を $P(w_i)$ とすれば、文集合 $\{w_1, \cdots, w_N\}$ のエントロピーは次式で求められる [19]。

$$H(L) = -\sum_{i=1}^{N} P(w_i) \log_2 P(w_i) \qquad (20)$$

全テキスト文の連接を $W = w_1 w_2 \cdots w_T$ とすれば

$$H(L) = -\log_2 P(W)$$

一単語当たりのエントロピーは

$$H_0(L) = -\frac{\log_2 P(W)}{\sum_i T_i} \qquad (21)$$

また言語の複雑さ・パープレキシティは

$$P(L) = 2^{H_0(L)} \qquad (22)$$

と定義される。言語の複雑さは、全ての生成可能な文に対する平均値だが、実際の学習・認識実験は有限の文集合であるため、テスト文集合に対するパープレキシティを使う方が現実的である。これを特に、テストセットパープレキシティと呼び、(22) 式で求める。

3.3 品詞レベルの確率言語モデルの比較

Trigram と bigram-HMM の場合に品詞レベルの確率モデルのパラメータ数を比べると、後者のモデルパラメータ数がかなり少なく、エントロピーも小さい（表 2 参照）。SCFG のパラメータ数が少ないのは、例えば NP が主語であろうと目的語であろうと同一の確率の書き換え規則を用いるためで「結び (tied)」にした HMM に対応するためである。同程度の予測率が得られれば、パラメータ数が少ないモデルの方が良いモデルと言える。なお、確率文節文法の場合は規則数である。

SCFG は Inside-Outside アルゴリズムの方法で学習したモデルである。Bigram-HMM の bigram 確率は二つの方法で求めた。一つは全データ共通の bigram 、もう一つは HMM のベストパスによる入力文のセグメンテーションによるもので、各状態ごとのローカル bigram を求めて、HMM と結合する（表中 bi-HMM(seg) の方）。後者のモデルは trigram よりもかなり優れている (表 2、表 3 参照)。

表 2: 品詞モデルのパラメータ数とエントロピー

各種モデル	parameters	test	train
bigram	4576	2.67	2.57 (2.42)
trigram	13824	2.48	2.38 (2.28)
hmm (s=7)	1225	2.75	2.72 (2.54)
hmm (s=10)	2500	2.58	2.54 (2.42)
hmm (s=15)	5625	2.50	2.44 (2.33)
tied-hmm (s=7)	217	2.91	2.87 (2.67)
tied-hmm (s=10)	340	2.82	2.74 (2.61)
tied-hmm (s=15)	435	2.76	2.67 (2.55)
bi-hmm (s=7)	1801	2.48	2.42
bi-hmm (s=10)	3076	2.18	2.12
bi-hmm(seg 7)	5257	2.33	2.30
bi-hmm(seg 10)	8260	2.09	2.04
SCFG (s=7)	511	2.94	2.88
SCFG (s=10)	1240	2.73	2.67
SCFG (s=15)	3735	2.68	2.62
等確率文節文法	161	4.24	4.14
確率文節文法	1725	3.18	2.96

（括弧内の数字はモデルパラメータによって計算されたエントロピー）

ATR の日本語対話データに対して、我々は確率文節文法を用いて解析による学習実験を行なった。口語体文節文法は本研究室で作成した 11 状態オートマトンである [13]。トポロジを変えない条件で確率を学習させた確率文節文法と等確率文節文法を用いて、この文法の任意個の繰り返しで言語モデルを表現した。このオートマトンは ATR データの 91.6% しか解析できない (テストデータも同じぐらい)。従って、テストデータパープレキシテは無限大になるので、表 2では解析できる文集合のみで求めた。予測率は HMM よりも悪く、自然な対話文の文法の構築の難しさがわかった。

Inside-Outside アルゴリズムによる SCFG の学習の計算量が大きいので、表 2 と 表 3の SCFG による結果は文の長さが 20 以内のデータだけを使う時に得られたエントロピーと予測率である。

表 3: ATR 旅行案内対話データの後続品詞予測的中率 (%)

データベース	日本語の ATR 対話データ							
	テストデータ				学習データ			
各種モデル	一位	二位	五位	十位	一位	二位	五位	十位
bigram	43.4	61.2	83.7	95.8	44.2	62.3	84.3	95.6
trigram	46.7	64.9	85.8	96.2	47.3	66.0	86.6	96.6
HMM (s=15)	44.8	63.3	85.3	95.7	45.2	64.3	85.8	96.4
tied-HMM(s=15)	42.1	61.2	83.2	95.3	42.5	61.7	83.8	95.0
bi-HMM (s=10)	47.1	65.2	85.9	96.5	47.6	66.4	86.9	96.7
bi-HMM(seg 10)	48.1	65.8	86.2	96.7	48.3	67.2	87.3	96.8
文節文法*	5.48	10.9	27.1	52.9	5.71	11.2	27.6	53.7
確率文節文法	22.7	29.8	48.9	67.0	23.3	30.1	48.9	67.0

3.4 単語レベルの確率言語モデルの比較

音声認識における言語モデルの一つの重要な役割はある時点まで認識できた単語列から、次にどんな単語がどんな確率で生じるかを求めることである。従って、よい言語モデルとは次の単語が正しく予測できる確率が高いモデルのことである。本稿では、単語列 $w_1 w_2 \cdots w_i$ と後続可能な単語 w_x を結合した $w_1 w_2 \cdots w_i w_x$ の生起確率を言語モデルで求め、確率の高い順の w_x を予測順位とする。

パラメータ数の関係は品詞レベルの時とそれほど変わらない。Trigram は相変わらずパラメータ数が一番多く、bigram-HMM よりかなり多い（表 4 参照）。

後続単語の予測的中率 ADD について、我々はエルゴディック HMM_1、bigram、trigram、SCFG 及び bigram-HMM で言語のエントロピーと予測的中率の比較実験を行なった（表 4、5 参照）。

HMM の場合は式(10)によるもので、SCFG は inside-outside アルゴリズムの方法で学習したモデルである。

表4と表5の $trigram_1$($bigram_1$) と $trigram_2$($bigram_2$) はそれぞれ式(7)と(6)による結果で、bigram-HMM_0 と bigram-HMM_1 はそれぞれ式(11)と(12)の二通りによる結果である。

この中で注目されるのは式(15)を用いた bigram-HMM_2 の場合である。パラメータの数が $trigram_2$ モデルよりかなり少ないにもかかわらず、エントロピーと予測率は $trigram_2$ とほとんど変わらない。今回の ATR タスクについて、言語モデルとしては trigram より性能が良いと言える。

川端は HMM によって SCFG を動的に変換する方法を提案し、クローズデータに対して約 80 のパープレキシティ(約 6.3 bit)を得ているが[17]、我々とデータベースが少し異なり、直接比較できないが、我々のモデルの方がパープレキシティは小さい。

また同タスクについて、データのスパースさを調べてみた。学習データ中に出現回数が 2 回以上であった組合せは、三つ組 (品詞、品詞、品詞) は 14.4%、(品詞、品詞、単語) は 0.19%であった。二つ組 (品詞、品詞) は 65.6%、(品詞、単語) は 3.1%であった。即ち、有効な $trigram_2$ のパラメータ数は 4983 個、bigram-HMM_2 の有効パラメータ数は 3376 個である。bigram-HMM_2 は trigram(品詞$_{t-2}$、品詞$_{t-1}$、単語$_t$) を品詞$_{t-2}$ について最適なように自動的にクラスタリングしていると言える。

4 連続音声の認識実験

音声認識における言語モデルの役割はあくまでも、文認識性能を向上させるためである。その有効性はパープレキシティである程度評価できるが、具体的には実際の認識システムに組み込んで評価する必要がある。このために、我々は同タスクの大語彙認識システムで、日本語の連続音声認識実験を行なった。

表 4: 単語レベルの確率モデルのパラメータ数及びエントロピー

($p(w_j/c_i)$ のパラメータの最大数$\alpha = 108336$)

モデル	パラメータの数	エントロピー test	エントロピー train
$bigram_1$	576 + α	6.45	6.24
$trigram_1$	13824 + α	6.10	5.85
$bigram_2$	108336	5.87	5.56
$trigram_2$	2600064	5.62	5.01
tied-HMM_1 (s=07)	217 + α	8.22	8.05
tied-HMM_1 (s=10)	340 + α	8.07	7.94
tied-HMM_1 (s=15)	585 + α	8.05	7.86
HMM_1 (s=07)	1225 + α	7.81	7.62
HMM_1 (s=10)	2500 + α	7.72	7.58
HMM_1 (s=15)	5625 + α	7.66	7.55
bi-HMM_0(s=07)	1801 + α	7.32	7.23
bi-HMM_0(s=10)	3076 + α	7.24	7.19
bi-HMM_0(s=15)	6201 + α	7.29	7.18
bi-HMM_1(s=07)	5257 + α	7.25	7.17
bi-HMM_1(s=10)	8260 + α	7.24	7.14
bi-HMM_1(s=15)	14265 + α	7.23	7.13
tied-HMM_2(s=07)	31647	7.84	7.66
tied-HMM_2(s=10)	45250	7.76	7.62
tied-HMM_2(s=15)	67950	7.71	7.60
tied-bi-HMM_2(s=07)	139983	6.40	6.27
tied-bi-HMM_2(s=10)	153576	6.32	6.18
tied-bi-HMM_2(s=15)	176271	6.27	6.12
bi-HMM_2(s=07)	329571	5.83	5.52
bi-HMM_2(s=10)	559846	5.74	5.39
bi-HMM_2(s=15)	1124211	5.68	5.28
SCFG (s=07)	511 + α	8.11	8.02
SCFG (s=10)	1240 + α	8.06	7.89
SCFG (s=15)	3735 + α	8.02	7.83
等確率文節文法	161 + α	10.5	9.83
確率文節文法	1725 + α	8.37	8.23

音声データと音声分析 ATR 観光案内対話データのテストデータから抽出した 100 文章を男性話者三人が発声したものを評価用データとし、各話者 30 文で話者適応化を行なった。表 6に分析条件と認識条件を示す。

文認識アルゴリズム 従来の連続音声認識のアルゴリズムとして知られている Viterbi サーチ (one-pass DP) は各フレーム毎に各単語の境界と仮定して、言語モデルによる確率の対数値と音響累積尤度を足すことを繰り返すことによって、次の式を満たす最尤の単語列の候補を計算することができる。

$$P(w^*|y_1^T) = \arg\max_{\{w_1^N\}\{t_1^N\}} \{\sum_{n=1}^{N} \log(P_a(w_n|y_{t_{n-1}+1}y_{t_{n-1}+2}\cdots y_{t_n})) + \text{weight} * \sum_{n=1}^{N} \log(P(w_n|w_{n-1}))\} \quad (23)$$

ここで $P_a(w_l|y_{t_{n-1}+1}y_{t_{n-1}+2}\cdots y_{t_n})$ は観測パターン系列 $y_{t_{n-1}+1}y_{t_{n-1}+2}\cdots y_{t_n}$ に対して単語 w_l の音響的な出現確率、$P(w_l|w_{l-1})$ は単語 w_{l-1} の次に単語 w_l が接続する確率である。即ち、(HMM の音響尤度 + 単語 bigram の尤度 + Viterbi ベストスコア + ビームサーチ) を基本的な認識の枠組みとして、各時刻・各状態において累積尤度を計算する。

ビームサーチ 各時刻 (フレーム) について、全て予測される単語を全部連接し、保存していくと可能な単語列が爆発的に増える可能性がある。従って、各フレーム毎において累積尤度の低い単語列は以後の探索から除外する。そのため、フレーム毎に最尤なものからビーム幅で制限した単語列候補に対してのみ計算を続けることにより、計算量及びメモリ量が大幅に減らしている。

認識実験 ビーム幅や、単語 bigram の尤度の重みなどの設定等が不十分なため、村上らの報告している結果と比べてまだ十分な認識精度が得られていない (現在のところ単語認識率は約 60 %)。

表 5: ATR 旅行案内の対話データの後続単語予測的中率 (20 位内)(%)

	テストデータ					学習データ				
予測順位のランク	一位	二位	五位	十位	二十位	一位	二位	五位	十位	二十位
$bigram_1$	10.0	22.4	34.0	44.8	53.5	11.4	22.8	34.9	45.7	54.2
$trigram_1$	11.2	23.3	33.7	45.0	54.2	11.6	23.1	34.5	45.7	54.6
$bigram_2$	19.9	31.8	47.1	59.9	69.7	20.2	31.6	47.1	60.3	70.3
$trigram_2$	22.5	34.8	51.3	63.3	71.1	22.9	34.6	52.2	64.5	74.2
tied-HMM_1(s=07)	6.7	14.0	25.1	35.6	48.1	8.5	15.2	26.2	36.8	49.7
tied-HMM_1(s=10)	7.2	14.5	25.6	36.2	48.5	8.9	15.6	26.7	37.2	50.1
tied-HMM_1(s=15)	7.3	14.4	25.7	36.3	48.5	8.9	15.6	26.9	37.4	50.2
HMM_1 (s=07)	7.6	15.6	26.7	37.0	49.6	9.6	16.5	27.9	38.0	50.6
HMM_1 (s=10)	7.7	15.6	26.6	37.0	49.3	9.6	16.5	27.8	38.0	50.3
HMM_1 (s=15)	7.3	15.5	26.7	37.2	49.5	9.0	16.3	27.8	38.2	50.8
bi-HMM_0(s=07)	10.8	20.8	32.7	42.5	51.5	11.6	21.1	33.8	44.0	52.8
bi-HMM_0(s=10)	10.8	20.7	32.7	43.2	51.5	11.6	21.1	33.8	44.5	52.9
bi-HMM_0(s=15)	9.9	20.1	33.0	42.3	51.3	11.0	21.5	33.9	43.8	52.8
bi-HMM_1(s=07)	11.3	21.2	32.9	42.8	51.7	12.2	21.7	34.2	44.6	53.1
bi-HMM_1(s=10)	11.4	21.2	33.1	42.8	51.8	12.2	21.8	34.2	44.7	53.2
bi-HMM_1(s=15)	11.4	21.2	33.1	42.8	51.8	12.2	21.8	34.2	44.7	53.2
tied-HMM_2 (s=07)	8.3	17.7	29.5	39.3	51.1	11.3	19.6	32.2	42.0	52.6
tied-HMM_2 (s=10)	8.4	17.8	29.7	39.5	51.3	11.6	20.1	32.6	42.3	53.9
tied-HMM_2 (s=15)	8.4	17.7	29.6	39.7	51.3	11.7	20.2	32.7	42.5	54.1
tied-bi-HMM_2(s=07)	18.1	27.5	40.1	52.2	60.4	20.6	29.5	42.4	54.5	63.4
tied-bi-HMM_2(s=10)	18.5	27.8	40.3	52.6	60.9	20.7	29.9	42.7	54.7	63.7
tied-bi-HMM_2(s=15)	18.4	28.1	40.2	53.4	60.7	20.8	30.2	42.8	55.4	63.9
bi-HMM_2(s=07)	21.5	32.5	49.8	62.2	70.0	22.1	33.5	50.9	63.3	73.2
bi-HMM_2(s=10)	22.1	34.3	50.8	62.7	70.4	22.4	33.7	51.2	63.7	73.5
bi-HMM_2(s=15)	22.6	34.6	51.2	63.0	70.8	22.9	34.2	51.8	64.1	74.0
SCFG (n=07)	6.9	14.2	25.3	35.8	48.3	8.8	15.5	26.5	37.1	50.0
SCFG (n=10)	7.2	14.5	25.7	36.2	48.5	8.9	15.6	26.7	37.2	50.1
SCFG (n=15)	7.4	14.5	25.8	36.5	48.6	8.9	15.7	26.9	37.4	50.4
等確率文節文法	2.2	3.8	5.6	8.3	11.3	2.3	3.9	5.8	8.5	11.4
確率文節文法	4.3	8.4	13.4	20.5	28.8	4.6	8.7	13.8	21.1	29.5

表 6: ATR 連続音声認識の実験条件

語彙	4699 単語
言語情報	P_{bigram} (単語 | 品詞)
Bigram の学習データ	10495 文 (117027 単語)
実験文数	100 文× 3 人
発声様式	朗読発話
発声内容	旅行案内に関する問い合わせ
音節モデル	5 状態 4 ループ HMM
音節カテゴリ数	113 音節
話者適応化文数	30 文 / 話者
平均文長	9 単語 / 文
サンプリング周波数	12kHz
窓関数	21.33ms ハミング窓
フレーム周期	8ms
分析	14 次元の LPC 分析
特徴パラメータ	10 次 LPC メルケプストラム + 10 次回帰係数 + パワー

5 結び

本報告で我々は種々の確率モデルで日本語の旅行案内対話データを用いてエントロピーと後続単語の予測率を求める比較実験を行なった。Bigram、trigram、HMM、SCFG、bigram-HMM の中で、trigram モデルのパラメータ数が一番多い。Bigram-HMM_2 は trigram とほぼ同等の予測能力を持っているが、パラメータの数が trigram よりはるかに少ない。比較したモデルの中では一番良い言語モデルであった。

しかし大語彙連続音声認識システムの構築に、音声認識に対する性能改善がここで述べた予測モデルの結果と完全に一致するかどうかは実際の音声認識評価実験で判断するしかない。現在、ATR 観光案内対話データベースと富士山観光案内タスクに関する我々のデータベースに対する認識実験により、以上の種々の確率モデルの評価をやっているところである。

参考文献

[1] J.Kupiec : "Robust part-of-speech tagging using a hidden Markov model". Computer Speech and Language, Vol. 6. pp. 225-242 (1992)

[2] A.Kempe : "Probabilistic tagging with feature structures". Proc. Coling, pp.161-165 (1994)

[3] A.-C.Lin, T.-H.Chiang and K.-Y. Su "Automatic model refinement – with an application to tagging". Proc. Coling, pp, 148-153 (1994)

[4] 中川、伊藤 : "音節標準パターンと逆時間向き係り受け解析法を用いた日本語文音声の認識"。信学論、 Vol. 70-D, No.2, pp. 2469-2478 (1987)

[5] 中川、大黒、橋本 : "構文解析駆動型日本語連続音声認識システム– SPOJUS-SYNO"。信学論、Vol.72-DII,No.8 pp.1726-1280 (1989).

[6] 伊藤、中川 : "確率オートマトンと品詞の 3 字組出現確率を用いた文節音声認識"。音響学会講演論文集, 3-5-18 (1987)

[7] 中川聖一 : "確率モデルによる音声認識"。電子情報通信学会 (1988) .

[8] J.H.Wright, G.J.F.Jones and E.N.Wrigley, "Hybrid grammar-bigram speech recognition system with first-order dependent model", Proc. ICASSP pp.I-169-172. (1992).

[9] L.R. Bahl, P.F. Brown, P.V.Souza & R.L.Mercer, "Tree-based stastical language model for natural language speech recognition", IEEE Trans. ASSP-37, 1001-1008 (1989)

[10] S.Nakagawa and I.Murase. "Comparison of language models by context-free grammar, bigram and Quasi/Simplified-trigram". IEICE Trans. Inf & Syst. Vol.E74, No. 7, pp. 1897-1906 (1991)

[11] 中川聖一 : "確率・統計的手法による音声認識"。音響学会誌、50 巻 2 号、 pp. 126-132, (1994).

[12] E.Black. "Parsing English by Computer: The state of the art". Proc. International Symposium on Spoken Dialogue. pp 77-81 (1993-10)

[13] 周旻、中川聖一 : "日本語及び英語の言語モデルに関する検討"。「自然言語処理における学習」シンポジウム、電子情報通信学会、pp. 57-64, (1994.11).

[14] 高橋敏、松岡達雄、鹿野清宏 : "VQ コードの Bigram で制約した音韻 HMM による音声認識"。信学論、Vol. J76-D-II No.7 pp.1346-1353 (1993-7).

[15] Lari.K & Young.S.J "The estimation of stochastic context-free grammar using the inside-outside algorithm." Computer Speech and Language, Vol.4, 35-56 (1990)

[16] T.Fujisaki, F.Jelinek, J.Cocke, E.Black & T.Nishino. "A Probabilistic Parsing Method for Sentence Disambiguation." Proc. International Parsing Workshop Pittsburgh. pp.85-94 (1989)

[17] 川端 豪 : "音声理解システム JUNO における構文制御"。音響講演論文集、1-Q-5 (1994.10)

[18] 村上仁一、山本寛樹、嵯峨山茂樹 : "Erogdic HMM による確率つきネットワーク文法の獲得の可能性について"。人工知能学会研究会資料 SIG-SLUD pp. 17-24 (1992-4)

[19] 中川聖一 : "情報理論の基礎と応用"。近代科学社 (1992)

[20] 村上仁一、松永昭一 : "単語の trigram を利用した文音声認識と自由発話認識への拡張"。電子情報通信学会、信学技報 SP93-127 pp. 71-78 (1994-1)