確率モデルに基づく日本語の後続単語予測評価

周 晏　　中川 聖一
豊橋技術科学大学 情報工学系

1 はじめに

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、音声認識の研究分野で重要な役割を荷なっている。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。

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 [1] [2]。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 [3] [4]。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 [5]。

単語予測の正確さは言語モデルの複雑さ（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ことは認識システムの性能を向上させるのに重要である。勿論、入力文に対するカバーレイジが十分大きいことが必要条件である。

本報告では、ATR の日本語対話データ (ADD) を用いて、隠れマルコフモデル (HMM)、確率文脈自由文法 (SCFG)、bigram、trigram、bigram-HMM、及び人間の作成した文節文法による言語のモデル化を検討し、テキスト文集合の複雑さ (パープレキシティ) 及び後続単語の予測を考察した。なお、品詞の予測に関しては既に報告した [6]。

2 確率言語モデル

大語彙連続音声認識システムでは確率言語モデルを用いることによって、タスクに対して、発声可能な入力文の確率が計算でき、探索空間が縮められ、誤認識率を大幅に減少させることができる。その中で、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も構成しやすい。

一方、確率文脈自由文法は自然言語のシンタクティック構造とその構造の確率の計算を統合し、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。その特殊な場合が確率正規文法である。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利用でき、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。

N-Gram が二つ組の場合 (bigram) に、HMM との統合モデルが音響モデルとして有効であることが示されており [7]、この手法を若干拡張した言語モデルを日本語対話データについて検討し、他の確率モデル (bigram, trigram, HMM, SCFG) との比較をする。また、人間が作成した日本語の文節文法、及びこれに確率を付与した文法も言語モデルとして評価する。

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。

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

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。

品詞列 $\boldsymbol{c}_1^t$ の条件付き予測確率を $P(c_t|\boldsymbol{c}_1^{t-1})$ とすれば、単語の予測確率は

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

あるいは

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

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。

2.1 N-Gram モデル

N-Gram モデルはある時刻の単語 w_t が直前の $N-1$ 個の単語 $\boldsymbol{w}_{i-N+1}^{i-1}$ にしか依存しないようなモデルである。従って、N-Gram モデルの事後確率は次の式で求める。

$$P(w_n|\boldsymbol{w}_1^{n-1}) \approx P(w_n|\boldsymbol{w}_{n-N+1}^{n-1})$$

また、N = 3 の時は trigram と言う。その時の生起確率は次式で与えられる。

$$P(\boldsymbol{w}_1^n) \approx \prod_{t=1}^{n} P(w_t|w_{t-2}, w_{t-1})$$

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。

$$P(\boldsymbol{w}_1^n) = \prod_{t=1}^{n} \sum_{x=1}^{K} P(c_x|c_{t-2}, c_{t-1}) * P(w_t|c_x) \quad (3)$$

$$P(w_1^n) = \prod_{t=1}^{n} P(w_t|c_{t-2}, c_{t-1}) \quad (4)$$

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。

2.2 隠れマルコフモデル

隠れマルコフモデルの場合に、部分文w_1^tの確率は次のように求められる。

$$P_{\mathrm{HMM}_1}(w_1^t) = \sum_{i=1}^{S} \alpha_i(w_1^t) \quad (5)$$

$$\alpha_i(w_1^t) \approx \sum_{x=1}^{K} \alpha_i(w_1^{t-1}c_x)) * P(w_t|c_x)$$

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

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。

これは式(3)に対応するもので、式(4)に対応する HMM としては、次式で定義される。

$$P_{\mathrm{HMM}_2}(w_1^t) = \sum_{j=1}^{S}(\sum_{i=1}^{S} \alpha_i(c_1^{t-1}) a_{ij} b_{ij}(w_t)) \quad (6)$$

$$\alpha_i(c_1^t) = \sum_{j=1}^{S} \alpha_j(c_1^{t-1}) a_{ji} b_{ji}(c_t)$$

2.3 確率文脈自由文法

確率文脈自由文法 (SCFG) は四つ組$\boldsymbol{G} = (\boldsymbol{V}_N, \boldsymbol{V}_T, \boldsymbol{P}, \boldsymbol{S})$で定義される。$\boldsymbol{V}_N$と$\boldsymbol{V}_T$ は各々非終端記号と終端記号の集合、 $\boldsymbol{P}$ 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 ($A \in \boldsymbol{V}_N$)、 γ は終端記号か非終端記号の列 ($\gamma \in (\boldsymbol{V}_N \cup \boldsymbol{V}_T)^*$)、 f_r は書き換え規則の確率、$\boldsymbol{S}$ は全ての文を生成するための開始記号である。

SCFG が *Chomsky* 標準形で表される時、 inside アルゴリズムによるシンボル系列 c_1^n の生成確率は以下のように再帰的に求める [8]。

$$\begin{aligned} e_i(s,t) &= P(i \Rightarrow c_s \cdots c_t / G) \\ &= \sum_{j,k=1}^{N} \sum_{r=s}^{t-1} a_{ijk} e_i(s,r) e_k(r+1,t) \end{aligned}$$

この inside 確率によって、品詞列の確率 $P(c_1^t)$ が求められる。

$$P_{\mathrm{SCFG}}(c_1^t) = \sum_{i=1}^{N} e_i(1,t) \quad (7)$$

従って、SCFG による単語列 w_1^t の生成確率は

$$P_{\mathrm{SCFG}}(w_1^t) = \prod_{i=0}^{t-1} \sum_{x=1}^{K} P(c_x|c_1^i) * P(w_{i+1}|c_x) \quad (8)$$

で求められる。

2.4 Bigram 結合 HMM

従来の HMM はシンボルの出力確率が単純な離散分布である。Bigram 駆動型 HMM の出力確率は前時刻の条件つき確率で表現される。品詞カテゴリの集合を $\boldsymbol{V} = \{v_1, v_2, \cdots, v_T\}$ とすれば、Bigram 駆動型 HMM の出力シンボル確率は次式の二通りで定義される。

$$b_{ij}(c_t|c_{t-1}) = \frac{p(c_t|c_{t-1}) b_{ij}(c_t)}{\sum_{m=1}^{T} p(v_m|c_{t-1}) b_{ij}(v_m)} \quad (9)$$

$$b_{ij}(c_t|c_{t-1}) = \frac{p(c_{t-1}c_t|ij)}{p(c_{t-1}|ij)} \quad (10)$$

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。式(9)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(10)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

$$\begin{aligned} P_{\text{bi-HMM}}(c_1^t) &= \sum_{j=1}^{S} \alpha_j(c_1^t) \quad (11) \\ &= \sum_{j=1}^{S} \sum_{i=1}^{S} \alpha_i(c_1^{t-1}) a_{ij} b_{ij}(c_t|c_{t-1}) \end{aligned}$$

また、部分文の生成確率は

$$P_{\text{bi-HMM}}(w_1^t) = \sum_{x=1}^{K} P_{\text{bi-HMM}}(w_1^{t-1}c_x) P(w_t|c_x) \quad (12)$$

で求められる ((5) 式参照)。

3 実験及び結果

言語モデルのパフォーマンスを比較するために、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。このデータベースは品詞ラベル付けのデータベースである。間投詞を取り除いて、単語数は 4514 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 文を用いた。品詞分類として、表 1 に示される 24 の品詞を用いている。

3.1 エントロピーとパープレキシティ

言語モデル$\boldsymbol{G}$ 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 [10]。

$$H(L) = -\sum_{i=1}^{N} P(w_i) log_2 P(w_i) \quad (13)$$

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

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

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

一単語当たりのエントロピーは

$$H_0(L) = -\frac{log_2 P(W)}{\sum_i T_i} \tag{14}$$

また言語の複雑さ・パープレキシティは

$$P(L) = 2^{H_0(L)} \tag{15}$$

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、(15) 式で求める。

3.2 確率モデルのパラメータ数の比較

Trigram と bigram-HMM と比べると、後者のモデルパラメータ数がかなり少ない (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)」 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。

3.3 エントロピーと後続単語の予測的中率

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。

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 2、3 参照)。非終端数 7、10、15 の SCFG は inside-outside アルゴリズムの方法でモデルを学習したものである。HMM の場合は式 (5) によるもので、状態数 7、10、15 のモデルで実験した。表 2と表 3の $trigram_1(bigram_1)$ と $trigram_2(bigram_2)$ はそれぞれ式 (3) と (4) による結果で、$bi\text{-}HMM_0$ と $bi\text{-}HMM_1$ はそれぞれ式 (9) と (10) の二通りによる結果である。

表 2: 確率モデルのパラメータ数及びエントロピー

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	6.45	6.24
$trigram_1$	13824	6.10	5.85
$bigram_2$	108336	5.87	5.56
$trigram_2$	2600064	5.62	5.01
tied-HMM_1 (s=07)	217	8.22	8.05
tied-HMM_1 (s=10)	340	8.07	7.94
tied-HMM_1 (s=15)	585	8.05	7.86
HMM_1 (s=07)	1225	7.81	7.62
HMM_1 (s=10)	2500	7.72	7.58
HMM_1 (s=15)	5625	7.66	7.55
bi-HMM_0(s=07)	1801	7.32	7.23
bi-HMM_0(s=10)	3076	7.24	7.19
bi-HMM_0(s=15)	6201	7.29	7.18
bi-HMM_1(s=07)	5257	7.25	7.17
bi-HMM_1(s=10)	8260	7.24	7.14
bi-HMM_1(s=15)	14265	7.23	7.13
SCFG (s=07)	511	8.11	8.02
SCFG (s=10)	1240	8.06	7.89
SCFG (s=15)	3735	8.02	7.83
等確率文節文法	161	10.5	9.83
確率文節文法	1725	8.37	8.23

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、我々とデータベースが少し異なり、直接比較できない [9]。我々はパープレキシティと後続単語の予測的中率両方を同時に求めている。表 2と 3の結果を見ると、学習データとテストデータのエントロピー及び予測率がほぼ等しいことから、品詞モデルのパラメータ推定に用いるデータ量が十分であることがわかった。また、$trigram_2(bigram_2)$ モデルのパラメータの数が他よりかなり多くて、エントロピーが一番小さく、予測率も一番良い。これを除ければ、パラメータの数が割と少ない bigram-HMM で良い予測率が得られている。

また、ATR の日本語対話データについて、人手で作成した文節文法を用いて確率文節文法を学習し、評価実験を行なった。日本語の口語体文節文法は本研究室で作成した 11 状態のオートマトンである [6]。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。このオートマトンのエントロピーは小さいが、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であった)。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、 3 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。

表 3: ATR 旅行案内の対話データの予測的中率 (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	五位内	十位内	二十位内	一位	二位	五位内	十位内	二十位内
$bigram_1$	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
SCFG (vn=07)	6.9	14.2	25.3	35.8	48.3	8.8	15.5	26.5	37.1	50.0
SCFG (vn=10)	7.2	14.5	25.7	36.2	48.5	8.9	15.6	26.7	37.2	50.1
SCFG (vn=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

4 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て比較実験を行なった。Bigram、trigram、HMM、SCFG、bigram-HMM の中で、$trigram_2$ モデルのパラメータ数が一番多いが、エントロピーが一番小さく、予測率も一番良かった。

後続品詞の予測に関しては、入力データを HMM による状態でセグメントし、状態ごとの bigram を構成してからの bigram-HMM 結合モデル (セグメントの bigram-HMM) はパラメータ数が trigram より少なく、後続品詞の予測率とエントロピーは trigram よりやや良く、比較したモデルの中では一番良い言語モデルであった [6]。

一方、後続単語の予測に関しては、trigram が一番良かった。$Trigram_1$ と HMM_1 は同等の予測率であるので、HMM_2(式 (6)) は $trigram_2$ と同等の性能が得られると予想される。今後の予定として、これらの言語モデルを用いて、実際の音声データの認識実験で評価するつもりである。

参考文献

[1] 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)

[2] 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).

[3] 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)

[4] 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) .

[5] 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).

[6] 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).

[7] 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).

[8] 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)

[9] 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)

[10] 中川聖一： "情報理論の基礎と応用"。近代科学社 (1992)