最大エントロピー法を用いてnグラム確率をバイグラム確率で補完する方法

Interpolation of n-gram probability by 2-gram probability using maximum entropy heuristics

江原 暉将
Terumasa EHARA
NHK 放送技術研究所
NHK Science and Technical Research Laboratories
eharate@strl.nhk.or.jp

1 はじめに

文字や単語の n グラム確率は統計的言語処理で強力な道具立てである。しかし、n が大きくなると異なりデータ数が増大し、確率を推定することが困難になるとともに、確率の値を格納する領域も大きくなるという問題があった。本文では、最大エントロピー法をヒューリスティックスとして利用して、バイグラム確率から $n \geq 3$ なる n グラム確率を求める方法について述べる。

2 最大エントロピー法の適用

ある言語 $\mathcal{L}$ の文字あるいは単語の集合を $\mathcal{W}$ とする。n 個の確率変数 $W_1, \cdots, W_n$ を $\mathcal{L}$ の中で、$\mathcal{W}$ の要素が n 個連続して出現する事象を表す確率変数とする。このとき、$w_i \in \mathcal{W}, i = 1, \cdots, n$ に対して、確率 $p(W_1 = w_1, \cdots, W_n = w_n)$ が考えられる。これは n グラム確率と呼ばれ $p(w_1, \cdots, w_n)$ とも書かれる。n グラム確率によって確率変数 $W_1, \cdots, W_n$ の同時分布が得られる。また、$i = 1, \cdots, n-1; j = i+1, \cdots, n$ に対して、$p(W_i = w_i, W_j = w_j)$ が考えられる。これは、バイグラム確率と呼ばれ、$p_{ij}(w_i, w_j)$ とも書かれる。バイグラム確率によって W_i, W_j に関する2次の周辺分布が得られる。$j-i+1$ のことをバイグラムのギャップ長と呼ぶ。

通常のバイグラムは w_i と w_j が隣接している場合、つまりギャップ長が0の場合のみをさすが、ここでは、必ずしも隣接していなくても良い。そこで、隣接している場合には、その旨を明記して、「隣接バイグラム」と書く。

本文でバイグラムと呼ぶものは、文字認識などで用いられてきた positional digram [6] や最近提案されている D-bigram [7] と同一の概念である。

周辺分布は同時分布から周辺和をとることによって求められる。そこで、バイグラム確率は n グラム確率から

$$p(w_i, w_j) = \sum_{w_1 \in \mathcal{W}} \cdots \sum_{w_{i-1} \in \mathcal{W}} \sum_{w_{i+1} \in \mathcal{W}} \cdots \sum_{w_{j-1} \in \mathcal{W}} \sum_{w_{j+1} \in \mathcal{W}} \cdots \sum_{w_n \in \mathcal{W}} p(w_1, \cdots, w_i, \cdots, w_j, \cdots, w_n) \quad (1)$$

と計算できる。

では、逆にバイグラム確率から n グラム確率を求めるには、どうすればよいであろうか。これは、周辺分布から同時分布を求めることであり、一般的には、解が一意的に求まるということはない。そこで、最大エントロピー法を適用する。つまり、n グラム確率の推定値 $p_h(w_1, \cdots, w_n)$ を以下の評価関数 $Q(p)$ が最大となる p とするものである。つまり

$$p_h = \arg\max_p Q(p) \quad (2)$$

$$Q(p) = -\sum_{w_1 \in \mathcal{W}} \cdots \sum_{w_n \in \mathcal{W}} p(w_1, \cdots, w_n) \log p(w_1, \cdots, w_n) \quad (3)$$

とする。ここで、p は制約式 (1) を満足しなければならない。このような制約付き最大値問題は Lagrange の未定係数法を用いて解くことができ [4]、以下のように解が求まる。

$$p_h(w_1,\cdots,w_n)=\prod_{i=1}^{n-1}\prod_{j=i+1}^{n}h_{ij}(w_i,w_j) \quad (4)$$

ここで、h_{ij}は未定の関数である。h_{ij}は制約式 (1) から比例反復法 [3, pp.235-238] を用いて以下のように求めることができる。$h_{ij}(w_i,w_j)$ の $r+1$ 回目の反復値は r回目の反復値から

$$h_{ij}^{(r+1)}(w_i,w_j)=\frac{p_{ij}(w_i,w_j)}{p_{h,ij}^{(r)}(w_i,w_j)}h_{ij}^{(r)}(w_i,w_j) \quad (5)$$

となる。ここで、$p_{h,ij}^{(r)}(w_i,w_j)$ は $h_{ij}^{(r)}(w_i,w_j)$ を式 (4) に代入して得られたものから式 (1) に従って周辺和をとることで計算される。

式 (4) から最大エントロピー法の解は対数線形分布となることが分かる。つまり $p_h(w_1,\cdots,w_n)$ の対数が

$$\log p_h(w_1,\cdots,w_n)=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\log h_{ij}(w_i,w_j) \quad (6)$$

と、バイグラムのみから計算される量である $\log h_{ij}(w_i,w_j)$ の線形和として書ける。

3 実験

3.1 実験の概要

最大エントロピー法の有効性を確認するために、漢字かな交じり文に対して、文字トライグラム確率の推定実験を行なった。$n=3$ の場合を実験したことになる。確率を推定するための学習データ ($\mathcal{D}_L$) として、NHK の放送ニュース原稿から 1992 年 7 月と 8 月分の約 89 万文字を用いた。$\mathcal{D}_L$から次の 4 種の手法でトライグラム確率を推定した。各手法の内容は 3.2節で説明する。

- 手法 1　最大エントロピー法
- 手法 2　$\mathcal{D}_L$のトライグラムによる推定
- 手法 3　$\mathcal{D}_L$の隣接バイグラムによる推定
- 手法 4　手法 2 と手法 3 の混合

推定されたトライグラム確率の値を試験データ ($\mathcal{D}_T$) を用いて評価した。$\mathcal{D}_T$は放送ニュース原稿の 1992 年 7 月から 1993 年 7 月分を用いた。試験データは学習データを含んでいるが、10 倍以上の量がある (約 1,100 万文字)。

評価方法は以下のようにした。$\mathcal{D}_T$からランダムにトライグラム 531 データを抽出し、D_Tとした。D_Tの各トライグラム (w_1,w_2,w_3) に対し、$\mathcal{D}_T$ から最尤推定されたトライグラム確率を真の値とみなし $p(w_1,w_2,w_3)$ とした。これと、手法 i ($i=1,2,3,4$) で推定されたトライグラム確率 $p_{h,i}(w_1,w_2,w_3)$ を比較した。ここで、n_{w_1,w_2,w_3}を $\mathcal{D}_T$での (w_1,w_2,w_3) の度数、Nを $\mathcal{D}_T$の総延べ度数 (11,204,540) とすると、

$$p(w_1,w_2,w_3)=\frac{n_{w_1,w_2,w_3}}{N} \quad (7)$$

で与えられる。

評価は非被覆度と近接度の 2 種で行なった。非被覆度は 531 個のテストデータのうち、$p_{h,i}(w_1,w_2,w_3)$ Nの値が 0.5 より小さなデータの占める割合であり、学習データによって被覆されていない試験データの割合を表す。

一方、近接度は

$$P=\sum_{(w_1,w_2,w_3)\in D_T}\frac{(p(w_1,w_2,w_3)-p_{h,i}(w_1,w_2,w_3))^2}{p(w_1,w_2,w_3)} \quad (8)$$

で計算される分布の間の近さを測る数値である。非被覆度、近接度ともに 0 であることが望ましい。

3.2 推定手法

前節で述べた 4 種の推定手法について述べる。手法 1 は 2節で述べた最大エントロピー法である。$\mathcal{D}_L$ でのバイグラム (w_1,w_2)、(w_1,w_3) の度数 m_{w_1,w_2}、m_{w_1,w_3}から周辺分布 $p_{12}(w_1,w_2)$ と $p_{13}(w_1,w_3)$ を

$$p_{12}(w_1,w_2)=\frac{m_{w_1,w_2}}{M} \quad (9)$$

$$p_{13}(w_1,w_3)=\frac{m_{w_1,w_3}}{M} \quad (10)$$

と求める。ここで、Mは$\mathcal{D}_L$の総延べ度数 (896,645) である。そして、式 (5) に従って比例反復法を実行する。初期値は

$$h_{ij}^{(0)}(w_i,w_j)=p_{ij}(w_i,w_j)^{\frac{1}{2}} \quad (11)$$

とした。

手法 2 は $\mathcal{D}_L$でのトライグラム (w_1, w_2, w_3) の生起度数を m_{w_1,w_2,w_3}とするとき、

$$p_{h,2}(w_1, w_2, w_3) = \frac{m_{w_1,w_2,w_3}}{M} \tag{12}$$

である。

手法 3 は 2 つの隣接バイグラム (w_1, w_2) と (w_2, w_3) およびユニグラム w_2の生起度数 m_{w_1,w_2}、m_{w_2,w_3}、m_{w_2}から

$$p_{h,3}(w_1, w_2, w_3) = \frac{m_{w_1,w_2}\, m_{w_2,w_3}}{m_{w_2}\, M} \tag{13}$$

と求められる。

最後に手法 4 は $p_{h,2}$と $p_{h,3}$から

$$\begin{aligned} &p_{h,4}(w_1, w_2, w_3) = \\ &\quad \mu\, p_{h,2}(w_1, w_2, w_3) + \\ &\quad (1-\mu)\, p_{h,3}(w_1, w_2, w_3) \end{aligned} \tag{14}$$

と smoothing をするものである。今回の実験では $\mu = 0.95$ で固定した。

3.3 最大エントロピー法の実行

最大エントロピー法の実行経過について述べる。比例反復法は 9 回の反復でほぼ収束した。収束後の $p_{ij}(w_i, w_j)$ と $h_{ij}(w_i, w_j)$ の関係を図 1に示す。両者にはあまり相関がない。一例として、w_1 ="▼"、w_2 ="窒"、w_3 ="素"の場合を観察すると

$$\begin{aligned} p_{12}(w_1, w_2) &= 1.115 \times 10^{-6} \\ p_{23}(w_2, w_3) &= 5.57 \times 10^{-6} \\ p_{13}(w_1, w_3) &= 1.115 \times 10^{-6} \end{aligned} \tag{15}$$

$$\begin{aligned} h_{12}(w_1, w_2) &= 1.507 \\ h_{23}(w_2, w_3) &= 2.57 \times 10^{-2} \\ h_{13}(w_1, w_3) &= 2.52 \times 10^{-5} \end{aligned} \tag{16}$$

となり、かなり食い違いがあるが

$$p(w_1, w_2, w_3) = 1.115 \times 10^{-6} \tag{17}$$

$$p_h(w_1, w_2, w_3) = 9.75 \times 10^{-5} \tag{18}$$

とトライグラム確率の推定値は真値にほぼ一致している。

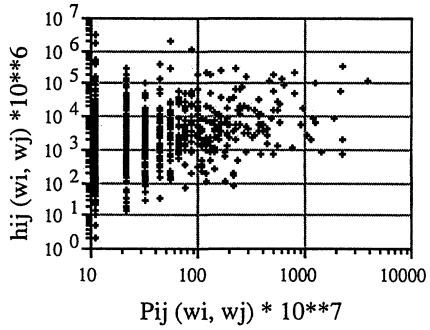


図 1: $p_{ij}(w_i, w_j)$ と $h_{ij}(w_i, w_j)$ の関係

3.4 実験結果

D_Tに含まれる各トライグラムに対する p の値と各 $p_{h,i}$の値の関係を図 2に示す。ただし、ここでは、確率の値そのものではなく、確率値に Nを乗じた値である $c = p\,N$や $c_{h,i} = p_{h,i}\,N$を用いた。そこで、c の値の最小値は 1 となる。図中、$c_{h,i}$の値が 0.01 以下になる場合は 0.01 とした。傾き 1 の直線上にデータが並べは、最も推定精度が高い。図から以下のことが言える。

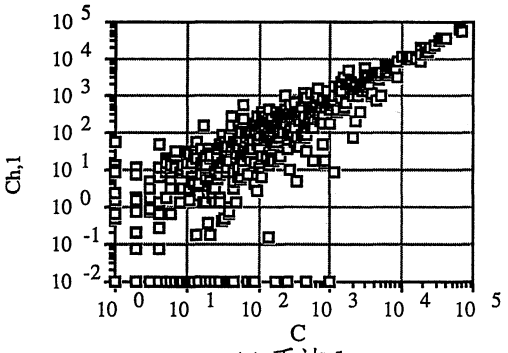
(1) 手法 2 では、$c_{h,2}$の値が 12.5 と 0.01 の間の値をとらない。これは、試験データが学習データの 12.5 倍の量があるためであり、試験データに含まれる頻度の小さいトライグラムに対する c と $c_{h,2}$ の値は大きく食い違っている。

(2) 手法 3 では、逆に、試験データに含まれる頻度の大きいデータに対して、c と $c_{h,3}$の値に開きが大きい。これは、隣接バイグラムのみを用いたのでは精度の高いトライグラム確率の推定ができないことを示している。特に c より c_hの方が小さい値をとる場合が多い。

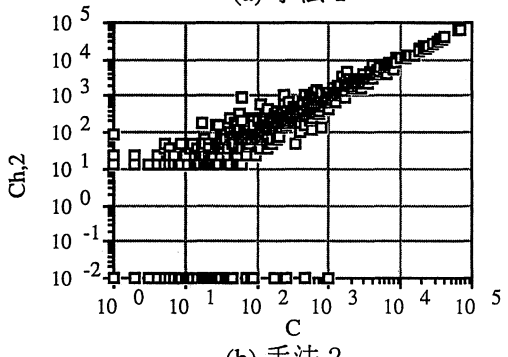
(3) 手法 1 と 4 は、手法 2 と 3 の中間的な推定値を持っているが、手法 1 は手法 4 よりも傾き 1 の直線上にデータが集中しており精度が高い。

3.5 評価結果

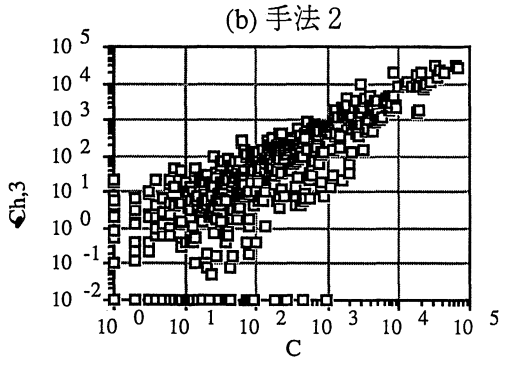
3.1節で述べた評価法によって、非被覆度と近接度を調べた。結果を図 3に示す。手法 2 と手法 4 は近接度は良いが非被覆度が悪い。一方、手法 3 はその逆である。手法 1 は最も原点に近く、最良の推定法であることがわかる。



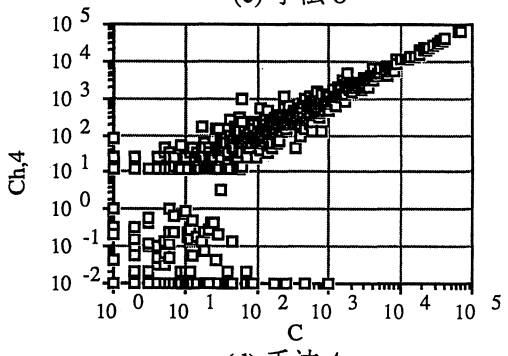
(a) 手法 1



(b) 手法 2



(c) 手法 3



(d) 手法 4

図 2: トライグラム確率の真値と推定値

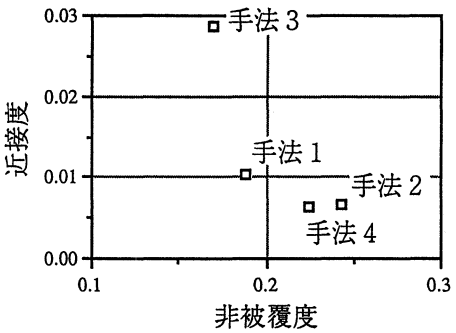


図 3: 非被覆度と近接度

4 おわりに

最大エントロピー法を用いて n グラム確率を推定する方法を提案し、実験によって有効性を確認した。本手法は、周辺分布から同時分布を求めるという確率論の一般的な方法を応用したものであり、n グラム確率の推定だけでなく、確率モデルによる様々な言語処理にも適用できる。[1] では、ゼロ主語の補完に、[5] では構文解析に、[2] ではタグつけに、同様な手法が利用されている。

参考文献

[1] 江原暉将, 金淵培. 確率モデルによるゼロ主語の補完. 自然言語処理, 掲載予定.

[2] Alexander Franz. An exploration of stochastic part-of-speech tagging. In *Proc. of NLPRS'95*, pp. 217–222, 1995.

[3] 廣津千尋. 離散データ解析. 教育出版, 1982.

[4] Gerald Minerbo. Ment: A maximum entropy algorithm for reconstructing a source from projection data. *Computer Graphics and Image Processing*, Vol. 10, pp. 48–68, 1979.

[5] Adwait Ratnaparkhi, Salim Roukos, and Todd R. Ward. A maximum entropy model for parsing. In *Proc. of ICSLP94*, pp. S16–7.1–S16–7.4, 1994.

[6] Edward M. Riseman and Alen R. Hanson. A contextual postprocessing system for error correction using binary n-grams. *IEEE Trans. Computers*, Vol. c-23, No. 5, 480–493 1974.

[7] 堤純也, 中西正和. 自然言語処理における統計情報量 D-bigram の提案. 情全大, 第 3 巻, pp. 7–8, 1995.