自然言語処理 113－5
音声言語情報処理 11－14
（1996. 5. 28）

最大エントロピー法を用いてバイグラム確率からnグラム確率を求める

江原 暉将
NHK 放送技術研究所
eharate@strl.nhk.or.jp

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が大きくなると推定すべきnグラム確率（パラメータ）の数が増加し、推定精度が低下するとともに、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、nグラム確率をn個の確率変数の同時分布と捉え、それらの確率変数に対する2次の周辺分布としてバイグラム確率を考える。その上で、周辺分布から同時分布を求める一般的手法である最大エントロピー法を適用して、バイグラム確率からnグラム確率を求める手法を提案する。本手法によって、パラメータの数を増加させることなく、精度良くnグラム確率を求めることができる可能性がある。次に、本手法を、文字トライグラムに適用して、有効性を実験する。その結果、（1）トライグラム自体を用いるよりも、未知データに対する被覆率が高くなること、（2）隣接バイグラムのみを用いる手法よりも、精度が高いこと、などが分かった。

Interpolation of n-gram probability by bi-gram probabilities using Maximum Entropy heuristics

Terumasa EHARA
NHK Science and Technical Research Laboratories
eharate@strl.nhk.or.jp

N-gram probability for characters or words is one of the most powerful tools in a statistical natural language processing. However, when n increases, the number of parameters, which should be estimated, also increases and estimation of them becomes inaccurate. And the area for storing these parameters becomes large. The author proposes the method to estimate n-gram probability from bi-gram probabilities using Maximum Entropy heuristics. This method provides n-gram probability from peripheral bi-gram probabilities accurately, not increasing the number of parameters. We show the experimental results to get character tri-gram from bi-grams. Here, character means Japanese Kanji character and/or Kana character. Their results show (1) our method has more wide coverage for unseen data than the method which uses tri-gram itself; (2) our method is more accurate than the method which uses usual contiguous bi-gram.

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が大きくなると推定するパラメータ数が増大し、推定精度が悪くなるとともに、パラメータの値を格納する領域も大きくなるという問題があった。例えば、文字nグラムの場合、文字の異なり数をMとすると、nグラムの異なり数は、M^nとなり、nの冪乗のオーダーである。

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。後述するように、nグラム確率を求めるためには、ギャップ長さが$n-1$以下のバイグラムが必要である。そこで、必要なパラメータの総数は、$(n-1)M^2$となり、nの線形オーダーとなる。

ここで、バイグラムと呼ぶものは、必ずしも隣接していなくても良く、間にギャップがあいているものも考えている。このようなバイグラムは、文字認識などで用いられてきた positional digram [8] や最近提案されている D-bigram [9] と同一の概念である。

以下、最大エントロピー法の適用方法につい

て、説明したあと、本手法を2種類の文字バイグラムに適用して、有効性を実験する。

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。$W_1,\cdots,W_n$ 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。

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

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

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。この問題に対する解法の1つに、最大エントロピー法を用いるものがある。これは、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

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

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

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 [6]、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 [5, 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

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

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。

最大エントロピー法の有効性を確認するために、2種類の実験を行なった。第1の実験は、漢字かな交じり文に対して、文字トライグラム確率を推定するものであり、第2の実験は、対話文の読みに関して、文字トライグラム確率を推定するものである。以下、これらの実験について述べる。

3 実験1：ニュース文の場合

3.1 実験の概要

本節では、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結果に

ついて述べる。$n=3$ 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、

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

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

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

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)、(w_2,w_3) の度数 m_{w_1,w_2}、m_{w_1,w_3}、m_{w_2,w_3}から周辺分布 $p_{12}(w_1,w_2)$、$p_{13}(w_1,w_3)$、$p_{23}(w_2,w_3)$ を

$$\begin{aligned}p_{12}(w_1,w_2)&=\frac{m_{w_1,w_2}}{M}\\ p_{13}(w_1,w_3)&=\frac{m_{w_1,w_3}}{M}\\ p_{23}(w_2,w_3)&=\frac{m_{w_2,w_3}}{M}\end{aligned} \tag{9}$$

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

$$h_{ij}^{(0)}(w_i,w_j)=p_{ij}(w_i,w_j)^{\frac{1}{2}} \tag{10}$$

とした。式（10）を選ぶ理由は、もし W_iと W_j が独立である、つまり $i=1,\cdots,n-1;j=i+1,\cdots,n$ に対して

$$p_{ij}(w_i,w_j)=p_i(w_i)\,p_j(w_j) \tag{11}$$

が成立すると仮定したとき

$$p_h^{(0)}(w_1,\cdots,w_n)=\prod_{i=1}^{n}p_i(w_i) \tag{12}$$

となるからである。

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、

$$p_{h,2}(w_1,w_2,w_3)=\frac{m_{w_1,w_2,w_3}}{M} \tag{13}$$

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

$$p_{h,3}(w_1,w_2,w_3)=\frac{m_{w_1,w_2}\,m_{w_2,w_3}}{m_{w_2}\,M} \tag{14}$$

と求められる。

最後に手法 4 は $p_{h,2}$と $p_{h,3}$から

$$\begin{aligned}&p_{h,4}(w_1,w_2,w_3)=\\ &\quad\mu\,p_{h,2}(w_1,w_2,w_3)+\\ &\quad(1-\mu)\,p_{h,3}(w_1,w_2,w_3)\end{aligned} \tag{15}$$

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。

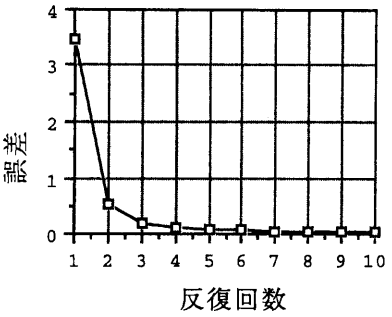


図 1: 反復回数と誤差の関係

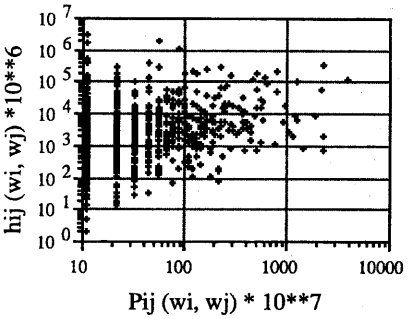


図 2: $p_{ij}(w_i, w_j)$ と $h_{ij}(w_i, w_j)$ の関係

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。$\mathcal{B}$をバイグラムの全体とし反復回数と誤差

$$e = \sum_{i=1}^{2} \sum_{j=i+1}^{3} \sum_{(w_i,w_j)\in\mathcal{B}} \frac{|p_{ij}(w_i,w_j) - p_{h,ij}^{(r)}(w_i,w_j)|}{p_{ij}(w_i,w_j)} \quad (16)$$

の関係を図 1に示す。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 2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

$$\begin{aligned} p_{12}(w_1,w_2) &= 1.115 \times 10^{-6} \\ p_{23}(w_2,w_3) &= 5.57 \times 10^{-6} \\ p_{13}(w_1,w_3) &= 1.115 \times 10^{-6} \quad (17) \\ h_{12}(w_1,w_2) &= 1.507 \\ h_{23}(w_2,w_3) &= 2.57 \times 10^{-2} \\ h_{13}(w_1,w_3) &= 2.52 \times 10^{-5} \quad (18) \end{aligned}$$

となり、かなり食い違いがあるが

$$\begin{aligned} p(w_1,w_2,w_3) &= 1.115 \times 10^{-6} \quad (19) \\ p_h(w_1,w_2,w_3) &= 9.75 \times 10^{-7} \quad (20) \end{aligned}$$

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。

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3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 $c = pN$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ば、最も推定精度が高い。図から以下のことが言える。

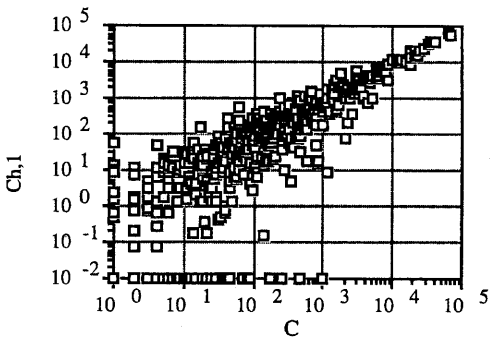
(1) 手法 2 では、$c_{h,2}$の値が 12.5 と 0.01 の間の値をとらない。これは、試験データが学習データの 12.5 倍の量があるためであり、試験データに含まれる頻度の小さいトライグラムに対する c と $c_{h,2}$の値は大きく食い違っている。

(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。

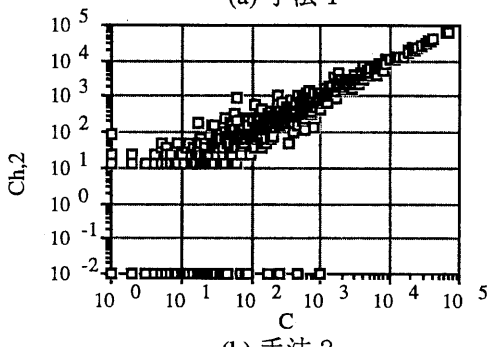
(3) 手法 1 と 4 は、手法 2 と 3 の中間的な推定値を持っているが、手法 1 は手法 4 よりも傾き 1 の直線上にデータが集中しており精度が高い。

3.5 評価結果

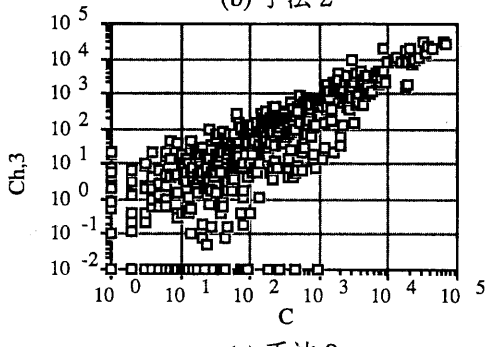
3.1節で述べた評価法によって、非被覆度と近接度を調べた。結果を図 4に示す。手法 2 と手法 4 は近接度は良いが非被覆度が悪い。一方、手法 3 はその逆である。手法 1 は最も原点に近く、最良の推定法であることがわかる。



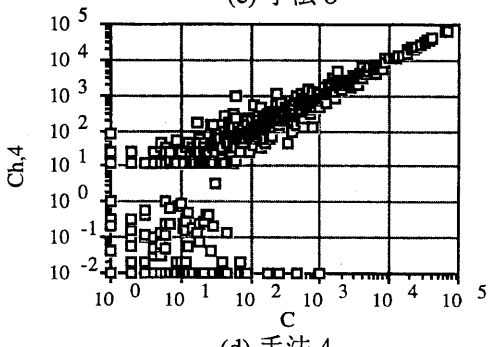
(a) 手法 1



(b) 手法 2



(c) 手法 3



(d) 手法 4

図 3: トライグラム確率の真値と推定値

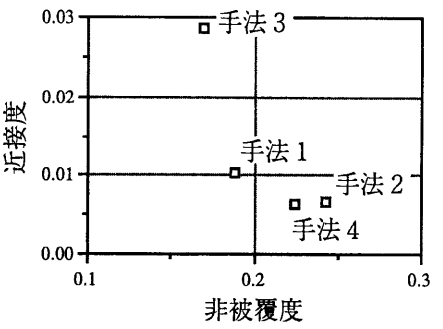


図 4: 非被覆度と近接度

4 実験 2：対話文の場合

4.1 実験の概要

本節では、ＡＴＲ対話データベース [2] から抽出した対話文 9,515 文を対象に、読みを取り出して、平かな、または、片かなで記述されたその読みのトライグラムに対して実験を行なった。今回は、データ量があまり多くないので、学習データそのものをテストデータとしても利用した。学習データは 28 万文字であり、$\mathcal{D}_{L2}$ と書く。本実験を行なう前に、バイグラム (w_i, w_j) のギャップ長 $j-i-1$ と w_i と w_j の独立性の強さを t-score を用いて評価した。

4.2 t-score による評価

バイグラム (w_i, w_j) の t-score は

$$t(w_i, w_j) = \frac{m_{w_i,wj} - \frac{m_{w_i} m_{w_j}}{M}}{\sqrt{m_{w_i,wj}}} \qquad (21)$$

で近似的に定義される。ただし分母が 0 の場合は 1 とした。$t(w_i, w_j)$ の値が 0 であれば、w_i と w_j の生起は独立である。

ギャップ長と t-score の絶対値が 4 以上のバイグラムの度数との関係を図 5に示す。ギャップ長 $j-i+1$ が長くなるに従って度数が減少し、7-gram 以上では、ほぼ独立であるといえる。

4.3 実験結果

3.4節に示した図 3と同様の図を実験 2 について、図 6に示す。手法 3 よりも、手法 1 の方が対角線に近くデータが分布しており、精度高くトライ グラム確率を再現していることが分かる。

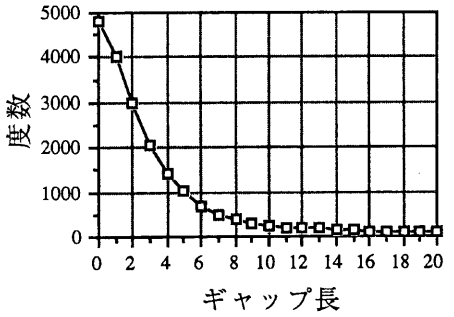
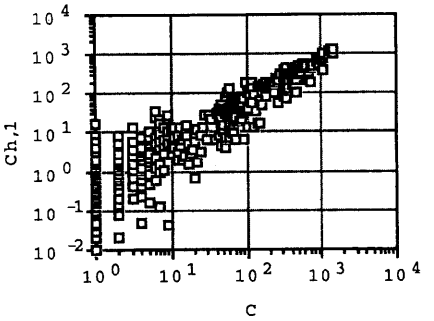
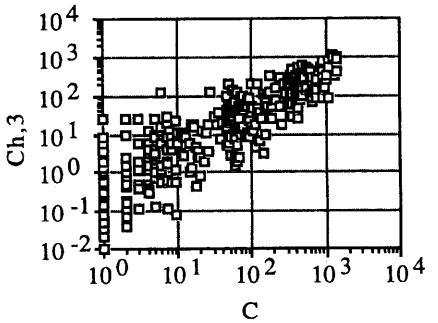


図 5: t-score の評価結果



(a) 手法 1



(b) 手法 3

図 6: トライグラム確率の真値と推定値

5 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適用できる。[3] では、ゼロ主語の補完に、[7] では構文解析に、[4] ではタグつけに、[1] では統計的な機械翻訳に、同様の手法が利用されている。

参考文献

[1] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, Vol. 22, No. 1, pp. 39–68, March 1996.

[2] 江原暉将, 小倉健太郎, 篠崎直子, 森元逞, 樽松明. 電話またはキーボードを介した対話に基づく対話データベースＡＤＤの構築. 情報処理学会論文誌, Vol. 33, No. 4, pp. 448–456, April 1992.

[3] 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.

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

[5] 廣津千尋. 離散データ解析. 教育出版, 1982.

[6] 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.

[7] 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.

[8] 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.

[9] 堤純也, ほか. d-bigram と trigram の相関に関する実験. 情報処理学会第 51 回全国大会, 第 3 巻, pp. 5–6, 1995.