音声言語情報処理　9－13
(1995. 12. 15)

二項事後分布に基づくN-gram言語モデルのBack-off平滑化

川端 豪　田本 真詞

NTT基礎研究所

〒243-01 神奈川県厚木市森の里若宮 3-1

0462-40-3584, kaw@siva.ntt.jp

あらまし　N-gram言語モデルは、自然音声言語を取り扱うための有力な手法の一つであるが信頼できるパラメータ推定のために、膨大な音声言語コーパスを必要とするという問題点があった。このため、疎(スパース)なデータからn-gram確率を推定する種々の手法が提案されている。Katzは学習コーパス中に出現しないn-gramの確率を(n-1)-gram確率から推定するback-off平滑の考え方を提案した。Katzの定式化はTuringの標本分布推定に基づくものであるが、状況によっては推定が不安定になることがある。本論文では、このback-off平滑法をKatzとは別の観点から理論的に再定式化することを試みる。二項事後確率分布の継承関係から導かれる新しいback-off平滑法は、より簡単な計算によって、希少標本からの安定な確率推定を実現する。

キーワード　自然言語処理、n-gram、バックオフ、JUNO

Back-off Method for N-gram Smoothing based on Binomial Posteriori Distribution

Takeshi KAWABATA and Masafumi TAMOTO

NTT Basic Research Laboratories

3-1 Morinosato-Wakamiya, Atsugi-shi 243-01 JAPAN

+81-462-40-3584, kaw@siva.ntt.jp

Abstract The n-gram language models are powerful for treating natural spoken languages, however need large amounts of spoken language corpus for estimating reliable model parameters. For estimating n-gram probabilities from sparse data, Katz's back-off smoothing method is promising. However, this approach is sometimes unstable because it uses singleton heuristics based on Turing's formula. This paper proposes a new back-off method based on Binomial Posteriori Distribution of n-gram probabilities, which achieves stable and more effective n-gram smoothing by sophisticated calculation formula with no heuristics.

key words　Natural language, n-gram, back-off, JUNO

1. まえがき

音声認識・理解系における言語処理の役割の一つは、発話仮説の生成に制限を加え、認識探索空間を絞り込むことである。この仮説生成に対する制限として、例えば伝統的な文法の枠組による文／非文判定などを用いることもできるが、現実に real world system の構築によく用いられているのは、n-gram と呼ばれる単純な単語連鎖の確率モデルである。このモデルでは、与えられた単語列の文らしさを、単語列が生成される先験確率として計算する。

[単語列 $w_1^I = w_1, w_2, \cdots, w_{i-1}, w_i, \cdots, w_I$ の確率]

$$P(w_1^I) = \prod_{i=1}^{I} P(w_i \mid w_1^{i-1}) \cong \prod_{i=1}^{I} P(w_i \mid w_{i-n+1}^{i-1}) \qquad (1)$$

最右辺の条件つき確率は、前もって大量の学習テキストから求めておく。問題は「大量の」といっても限界があることである。上記の条件つき確率の種類は語彙数を M とするとき M^n である。例えば 5000 単語の 3-gram (trigram) に対して 1.25×10^{11} であり、よほど巨大なテキストコーパスを用いても、精度の高い統計を行うことは難しい。実際、可能な n-単語連鎖がたまたま学習テキストに含まれず文全体の確率が 0 と推定されてしまう致命的な誤りが、頻繁に起こる。

この問題に対する一つの解答として Katz [1] により back-off 平滑法が提案されている。基本的な考え方は、n-単語連鎖の確率を (n-1)-単語連鎖の確率に基づいて平滑化するというものであり、単純な底上げ法 (flooring) よりも精度の高い確率推定が期待できる。この提案以来、方式の修整や改良、性能評価など多くの追随研究がなされている [2][3]。

本論文では、この back-off 平滑法を Katz とは別の観点から理論的に再定式化することを試みる。「二項事後確率分布 (BPD)」の継承関係から導かれる新しい back-off 平滑法は、より簡単な計算によって、稀少標本からの安定な確率推定を実現する。

2. Katz の back-off 平滑法

Katz は、Turing's estimates [4] として知られる個体数分布の関係式をもとに、学習テキストに出現しない n-単語連鎖の確率を (n-1)-単語連鎖の確率から求める back-off 平滑の考え方を提案した [1]。

標本の総数を N、標本集合中に r 回含まれる個体が n_r 種類あるとする。また r の最大値を $r_{\max}$ とする。このとき、明らかに次式が成立する。

$$N = \sum_{r=1}^{r_{\max}} r \cdot n_r \qquad (2)$$

このとき、標本集合中に r 回含まれる個体の先験確率を次式のように推定する (Turing's estimate)。

$$P_T = r^* / N \quad \text{where} \quad r^* = (r+1)\, n_{r+1} / n_r \qquad (3)$$

すなわち「出現回数の多い個体ほどその種類は少ない」と仮定する (i.e. $r \cdot n_r \cong const.$)。これは、個体 s の標本集合中での出現回数を $c(s)$ と表わすとき、次のように書ける。

$$\sum_{s:\, c(s)=r} P_T(s) \cong const. \quad (r=1, \ \cdots, \ r_{\max}) \qquad (4)$$

この関係を r に対する $r \cdot n_r$ の分布として模式的に書くと図 1 のようになる。Katz のアイデアの本質は、この関係を $r=0$ の場合に拡大適用したことである（図中の破線）。

$$\sum_{s:\, c(s)=0} P_T(s) \cong \sum_{s:\, c(s)=1} P_T(s) = \frac{n_1}{N} \qquad (5)$$

標本集合中に 1 個だけ含まれる個体を singleton という。singleton を数え上げ (n_1)、全標本数 N に対する割合を計算することによって、標本集合中に含まれない個体の確率の総和を推定する。

(5) 式により、標本集合中に含まれない個体に対して与えるべき確率の総和が定められる。あとは、そのような個体間で確率を分配すればよい。確率の分配は、一段階低いレベルの確率に比例させて行う (back off)。例えば m-gram の場合 (m-1)-gram 確率に比例させて分配する。最終的に次式が得られる。

$$P_s(w_m \mid w_1^{m-1}) = \qquad (6)$$

$$\begin{cases} \tilde{P}(w_m \mid w_1^{m-1}) & (c(w_1^{m-1}) > 0) \\ \alpha \cdot P_s(w_m \mid w_2^{m-1}) & (c(w_1^{m-1}) = 0, c(w_2^{m-1}) > 0) \\ P_s(w_m \mid w_2^{m-1}) & (c(w_1^{m-1}) = 0, c(w_2^{m-1}) = 0) \end{cases}$$

where

$$\tilde{P}(w_m \mid w_1^{m-1}) \equiv \frac{c(w_1^m)+1}{c(w_1^m)} \cdot \frac{n_{c(w_1^m)+1}}{n_{c(w_1^m)}} \qquad (7)$$

$$\alpha(w_1^{m-1}) \equiv \frac{1 - \sum_{w_m : c(w_1^m)>0} \tilde{P}(w_m \mid w_1^{m-1})}{1 - \sum_{w_m : c(w_1^m)>0} \tilde{P}(w_m \mid w_2^{m-1})} \qquad (8)$$

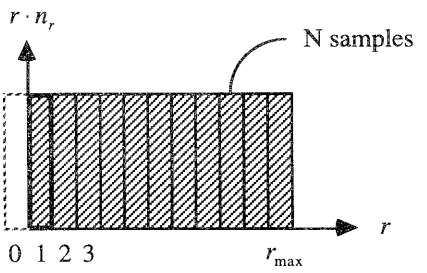

Fig.1 Turing's estimate

Katzの方法は2つの仮定（あるいは予想）に立脚している。第一の仮定は n-gram 確率が $(n-1)$-gram 確率に「似て」いるということであり、第二の仮定は前述の Turing's estimate である。ここで、現実のデータに基づいて、これらの仮定の妥当性を検討してみよう。

ある対話データ（実は後述の学習用テキスト）を音素列に変換し、音素 w_t の出現回数 $c(w_t)$、音素2つ組の出現回数 $c(w_{t-1},w_t)$、及び3つ組の出現回数 $c(w_{t-2},w_{t-1},w_t)$ を数え上げる。音素の bigram 確率、及び trigram 確率は、次式のように定義される。

Bigram probability:

$$P(w_t \mid w_{t-1}) = \frac{c(w_{t-1},w_t)}{c(w_{t-1})} \tag{9}$$

Trigram probability:

$$P(w_t \mid w_{t-2},w_{t-1}) = \frac{c(w_{t-2},w_{t-1},w_t)}{c(w_{t-2},w_{t-1})} \tag{10}$$

図2に w_t と w_{t-1} の一致で対応づけられた bigram 確率と trigram 確率の散布図を示す(両対数)。この図によって両者の間に強い相関があることが確認できる。実は、trigram 確率を w_{t-2} について、コンテキスト頻度 $(c(w_{t-2},w_{t-1}))$で重みづけて平均すると、

$$\frac{1}{\sum_{w_{t-2}} c(w_{t-2},w_{t-1})} \sum_{w_{t-2}} c(w_{t-2},w_{t-1}) \cdot P(w_t \mid w_{t-2},w_{t-1})$$

$$= \frac{1}{c(w_{t-1})} \sum_{w_{t-2}} c(w_{t-2},w_{t-1},w_t) = \frac{c(w_{t-1},w_t)}{c(w_{t-1})} \tag{11}$$

となり、厳密に bigram 確率と一致する。これらの議論から、 n-gram 確率に $(n-1)$-gram 確率を反映させる第一の仮定は妥当なものといえる。

次に、第二の仮定を検討するために、実際の対話データに含まれる音素2つ組連鎖について、出現回数 r と、出現回数が r である連鎖の種類数 n_r を数え、各 r に対する $r \cdot n_r$ の分布を作成してみた。この分布図を図3に示す。実際のデータに基づく分布は、r に関して必ずしも平坦ではなく、かなりのばらつきを伴うことがわかる。このばらつきがたまたま singleton の数え上げにおいて生じ、$n_1=0$ となると、(5)式によって標本集合中に含まれない音素連鎖への確率の割り当ても0になってしまう。すなわち Katz の方法では、可能な n-単語連鎖がたまたま学習テキストに含まれず文全体の確率が0と推定されてしまう致命的な誤りを、完全には防止することができない。

3. 二項事後分布に基づく back-off 平滑法

本論文では n-gram 確率に $(n-1)$-gram 確率を反映させるという back-off 平滑の考え方を生かしながら、稀少標本からの n-gram 確率推定問題を Katz とは別の観点に基づいて理論的に定式化する。

3.1 二項事後分布 [5]

無限の要素からなる母集団があり、有限個の抽出単位（前述の個体に相当）とそれらに対する確率によって規定されているとする。この母集団から無作

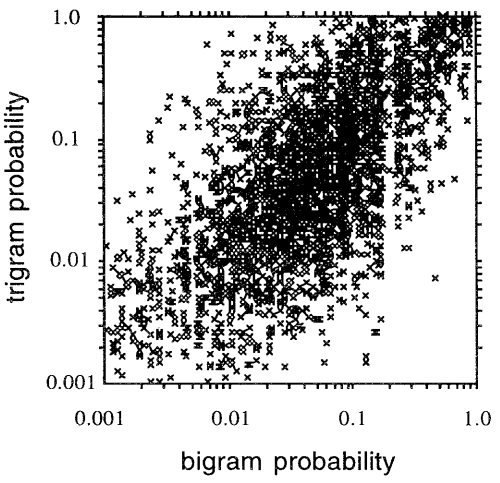

Fig. 2 Correlation between bigram and trigram probabilities

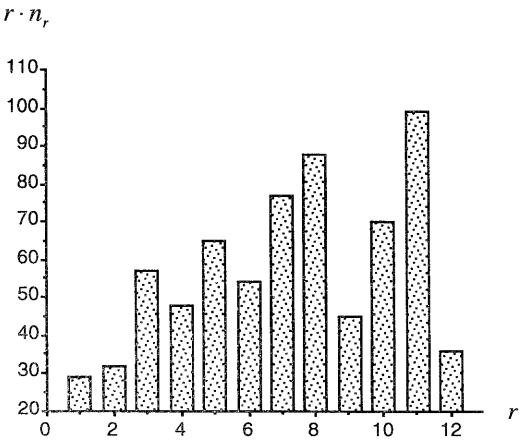

Fig. 3 An example of phoneme bigram statistics r vs. $r \cdot n_r$

為に n 個の標本抽出を行ったとき、確率 p の抽出単位が k 回抽出される確率は、次式の二項分布に従う。

$$P(k \mid n,p)={}_nC_k \cdot p^k(1-p)^{n-k} \tag{12}$$

逆に n, k が観測されたときの p の事後確率は Bayes の定理 [6] により、

$$P(p \mid n,k) = \frac{{}_nC_k \cdot p^k(1-p)^{n-k}P(p)}{\int_0^1 {}_nC_k \cdot q^k(1-q)^{n-k}P(q)dq} \tag{13}$$

となる。但し $P(p)$ は p の先験確率である。${}_nC_k$ を約分して、次式の二項事後分布 (BPD, Binomial Posteriori Distribution) を得る。

$$B_n^k(p) = \frac{p^k(1-p)^{n-k}P(p)}{\int_0^1 q^k(1-q)^{n-k}P(q)dq} \tag{14}$$

上式を用いることによって、k=0 すなわち標本単位が標本集合中に 1 度も出現しなかった場合でも、問題なく分布が推定できる。

3.2 エントロピー最小化に基づく確率推定

抽出単位 s に対する (真の) 確率を p_s、何らかの手段による推定値を $\tilde{p}_s$ とする。今、この推定値のエントロピーを次式のように定義する。

$$H \equiv -\sum_s p_s \log \tilde{p}_s \tag{15}$$

ちなみに 2^H はパープレキシティと呼ばれる量である。このエントロピーを、境界条件 ($0 \le \tilde{p}_s \le 1$, $\sum_s \tilde{p}_s = 1$) 下で最小化したい。

まず、関数 f を次式のように定義する。

$$f \equiv \sum_s p_s \log \tilde{p}_s + \lambda(1 - \sum_s \tilde{p}_s) \tag{16}$$

ここで、λ は Lagrange の未定乗数である。極値を求めるために f を各 $\tilde{p}_s$ で偏微分して 0 とおく。

$$\frac{\partial f}{\partial \tilde{p}_s} = p_s \frac{1}{\tilde{p}_s} - \lambda = 0 \quad \therefore \ \tilde{p}_s \propto p_s \tag{17}$$

実際には真の確率分布は未知であり (11) 式に基づいて標本集合から推定しなければならない。この場合のエントロピーは次式のように定義される。

$$H \equiv -\sum_s \int_0^1 p \cdot B_{n(s)}^{k(s)}(p) \cdot \log \tilde{p}_s \, dp \tag{18}$$

同じような手続きを経て、次の結論が導き出される。

$$\tilde{p}_s \propto \mu_s \quad (\mu_s \equiv \int_0^1 p \cdot B_{n(s)}^{k(s)}(p)dp) \tag{19}$$

3.3 二項事後分布に基づく back-off 平滑法

ここで、準備のため、分布

$$f(p) = p^{\alpha-1}(1-p)^{\beta-1} \tag{20}$$

の平均値と分散を求めておく。まず、平均値は、

$$\mu = \frac{\int_0^1 p \cdot f(p)dp}{\int_0^1 f(p)dp} = \frac{\int_0^1 p \cdot p^{\alpha-1}(1-p)^{\beta-1}dp}{\int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp} \tag{21}$$

部分積分法により、

$$\begin{aligned}
&\int_0^1 p \cdot p^{\alpha-1}(1-p)^{\beta-1}dp \\
&= \frac{1}{\alpha+\beta}\left(\left[p^\alpha(1-p)^\beta\right]_0^1 + \alpha\int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp\right) \\
&= \frac{\alpha}{\alpha+\beta}\int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp
\end{aligned} \tag{22}$$

よって、

$$\mu = \frac{\alpha}{\alpha+\beta} \tag{23}$$

と求められる。分散に関しても同様の手順で、

$$\begin{aligned}
\sigma^2 &= \int_0^1 (p-\mu)^2 f(p)dp \\
&= \int_0^1 \left\{p^2 - 2\mu p + \mu^2\right\} f(p)dp \\
&= \frac{\alpha+1}{\alpha+\beta+1} \cdot \frac{\alpha}{\alpha+\beta} - 2\left(\frac{\alpha}{\alpha+\beta}\right)^2 + \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\alpha+1}{\alpha+\beta+1} \cdot \mu - \mu^2
\end{aligned} \tag{24}$$

と求められる。

まず unigram について考えよう。この場合の抽出要素は各単語であり、p_s は単語の出現確率である。標本集合中の単語数を N、単語の種類を M、単語 w_t の出現回数を $c(w_t)$ とする。また先験確率分布として次式の分布を仮定する。

$$P(p) = (1-p)^{M-2} \tag{25}$$

この分布は、平均値が $1/M$ になるように、(20) 式において α=1, β=M-1 と設定したものである。結局 unigram 確率に関する BPD は次式で計算される。

$$\begin{aligned}
{}^{(uni)}B_N^{c(w_t)}(p) &= \frac{p^{c(w_t)}(1-p)^{N-c(w_t)}P(p)}{\int_0^1 q^{c(w_t)}(1-q)^{N-c(w_t)}P(q)dq} \\
&= \frac{p^{c(w_t)}(1-p)^{N-c(w_t)+M-2}}{\int_0^1 q^{c(w_t)}(1-q)^{N-c(w_t)+M-2}dq}
\end{aligned} \tag{26}$$

(19) 式より、unigram 確率の推定値を上記分布の平均値として計算する。

$$\tilde{P}(w_t) = \mu_t^{uni} = \frac{c(w_t)+1}{N+M} \tag{27}$$

次に bigram について考えよう。抽出単位は単語対であり、p_s は条件つき確率 $P(w_t \mid w_{t-1})$ である。標本集合中の 単語 w_{t-1} の出現回数を $c(w_{t-1})$、単語対 w_{t-1}, w_t の出現回数を $c(w_{t-1}, w_t)$ とする。また先験確率として次式の分布を仮定する。

$$P(p) = p^{\alpha-1}(1-p)^{\beta-1} \tag{28}$$

この先験確率分布を一段階低いレベルの確率分布に

基づいて設定するのが、本論文で提案する BPD back-off 平滑の考え方である。(28) 式の分布の平均値が (27) 式の値に一致するためには α, β の間に次の関係が成立することが必要である。

$$\frac{\alpha}{c(w_t)+1}=\frac{\alpha+\beta}{N+M}(=\gamma_1) \tag{29}$$

この γ_1 を用いて bigram の BPD を計算する。

$$\tilde{P}(w_t \mid w_{t-1})=\mu_t^{bi}=\frac{c(w_{t-1},w_t)+\alpha}{c(w_{t-1})+\alpha+\beta}$$
$$=\frac{c(w_{t-1},w_t)+\gamma_1(c(w_t)+1)}{c(w_{t-1})+\gamma_1(N+M)} \tag{30}$$

以降、この γ を継承係数 (inheritance factor) と呼ぶことにする。

Trigram の場合は、抽出単位は単語の 3 つ組であり、p_s として条件つき確率 $P(w_t \mid w_{t-2},w_{t-1})$ を求めることになる。先験確率分布として次式の分布を仮定する。

$$P(p)=p^{\alpha-1}(1-p)^{\beta-1}$$

where α,β:

$$\frac{\alpha}{c(w_{t-1},w_t)+\gamma_1(c(w_t)+1)}=$$
$$\frac{\alpha+\beta}{c(w_{t-1})+\gamma_1(N+M)}(=\gamma_2) \tag{31}$$

結果は次式の通りである。

$$\tilde{P}(w_t \mid w_{t-2},w_{t-1})=\mu_t^{tri}= \tag{32}$$
$$\frac{c(w_{t-2},w_{t-1},w_t)+\gamma_2(c(w_{t-1},w_t)+\gamma_1(c(w_t)+1))}{c(w_{t-2},w_{t-1})+\gamma_2(c(w_{t-1})+\gamma_1(N+M))}$$

以降、漸化的に *n*-gram に拡張していけばよい。

4. 評価実験と結論

提案する新しい back-off 平滑法を、大規模テキストデータベースにおけるパープレキシティの削減効果によって評価した。学習用テキストは、国際会議の参加登録を想定して収集したキーボード会話であり、句読点で区切られた約 3,000 の発話からなっている。異なり語彙数は約 6,000 、音素の種類は長母音や拗音を区別して 41 種類である。学習用テキストを単語列及び音素列に変換し、単語及び音素 *n*-gram (*n*=2..4) を計算する。この *n*-gram モデルを用いて、同じ条件下で収集した評価用テキスト約 20,000 発話に対する確率を計算し、単語及び音素パープレキシティを算出する。学習用及び評価用テキスト中の単語数及び音素数を表 1 に示す。

単純な確率の底上げ (flooring) 法、 Katz の方法と底上げ法の併用、及び提案する BPD back-off 平滑法の 3 つの手法を比較評価した。 Katz の方法に底上げ法を併用したのは、単独の適用では今回のデータに対して fatal error が出現し、パープレキシティ＝∞となってしまったためである。なお、底値 及び 継承係数 γ の値は、各々の場合で最良の結果が得られるように調整した。

音素パープレキシティの削減効果の比較を表 2 に示す。単純な底上げ法では、gram-数が大きくなって確率推定の信頼性が低くなってくると、それがパー

Table 1 The number of tokens in the corpus

corpus	utterances	words	phonemes
Texts for training	3,008	20,060	65,761
Texts for tests	20,000	158,601	522,304

Table 2 Reduction of Phoneme Perplexity

method	bigram	trigram	4-gram
flooring	9.1	8.5	8.8
Katz + flooring	9.1	7.2	6.6
BPD back-off	9.2	7.1	5.9

Table 3 Reduction of Word Perplexity

method	bigram	trigram	4-gram
flooring	173	518	1510
Katz + flooring	136	119	128
BPD back-off	129	102	102

プレキシティの増加に直結する。表2では4-gramにおいて、この劣化が始まっている。これに対しKatzの方法及びBPD back-off法では、パープレキシティは減少している。

単語パープレキシティの削減効果の比較を表3に示す。この厳しい実験条件においては、Katzの方法も耐えきれず、4-gramにおいてパープレキシティが増加してしまっている。BPD back-off法では、この場合にもパープレキシティの劣化は起きていない。

以上の結果より、提案するBPD back-off平滑法は、gram-数が大きくなって信頼性の高い確率推定が困難になるほど、他の方法に対する優位性が際だってくることが確認できた。

5. むすび

二項事後確率分布(BPD)の継承関係に基づく、新しい*n*-gram確率推定法を提案した。本手法は、記号の*n*個組が標本集合中で稀少であるとき(*n*-1)-gram確率を*n*-gram確率推定に反映させるという意味においてKatzの提案したback-off平滑法に似ている。しかし、その理論的背景は全く異なっており、また計算もずっと簡単である。

提案するback-off平滑法を、大規模テキストデータベースにおけるパープレキシティの削減効果によって評価した結果、gram-数が大きくなって信頼性の高い確率推定が困難になるほど、提案する手法の優位性が現われてくることを確認した。継承係数γの決定法については、稿を改めて論じたい。

謝辞

諸事にわたり研究をサポートして頂いているNTT基礎研究所 情報科学研究部 石井健一郎部長、並びに誉田雅彰グループリーダに感謝致します。いつも熱心に討論して頂く誉田研究グループの皆様に深謝致します。

参考文献

[1] Katz, S. M. : "Estimation of Probabilities form Sparse Data for the Language Model Component of a Speech Recognizer,"
IEEE Trans., ASSP-35, 3, pp.400-401 (Mar. 1987)

[2] Jelinek, F. : "Self-organized Language Modeling for Speech Recognition," Readings in Speech Recognition, pp.450-503,
Morgan Kaufman Pub. (1991)

[3] Kneser, R. and Ney, H. : "Improved Backing-off for M-gram Language Modeling," Proc. ICASSP-95, I, pp.181-184 (May 1995)

[4] Nadas, A. : "On Turing's Formula for Word Probabilities," IEEE Trans., ASSP-33, 6, pp.1414-1416 (Dec. 1985)

[5] 川端：「確率文法と話題マルコフモデルに基づく音声認識のための話題制御」, 信学論 D-II, J77-DII, 10, pp.1967-1972 (Oct. 1994)

[6] Duda, R. O. and Hart, P. E. : "Pattern Classification and Scene Analysis," A Wiley-Interscience Pub. (1973)