Vol.2022-NL-254 No.9
Vol.2022-SLP-144 No.9
2022/11/29

二項事後分布に基づく N-gram 言語モデルにおける継承係数の文脈依存性の検討

込山航大[1]　川端豪[1]

概要：本研究は，音声認識に用いる N-gram 言語モデルの一つである BPD バックオフ法において，継承係数の文脈依存性について検討する．評価用テキストに含まれる各単語の bigram 確率のヒストグラムを観察し，横軸「-対数確率」に着目し，全ての文脈に同じ継承係数を設定するのと特定の文脈に異なる継承係数を設定するのとでは，どのようにパープレキシティに影響を与えるのかについての検討を行う．

キーワード：言語モデル，パープレキシティ, N-gram，バックオフ

1. はじめに

音声認識技術の現状は，残念ながら 100%の認識率が得られてはいない．この性能追及には終わりはない．国語辞典の語彙が更新されていくように，日々新しい単語が増えていく．音声認識の性能は認識対象となる単語の種類数に影響されるので，増加するすべての語彙を常に認識しようとするのは無理がある．

数万語の語彙を扱う大語彙連続音声認識システムの設計において，実効的な語彙数を絞り込むことは重要な設計要素になると考えられる．

近年，大語彙連続音声認識は実用域に入り，その要因としては機械が読める言語データの整備が進み，また計算機の進歩により大規模なデータの取り扱いが可能となった．

発話全体が文章であるような連続音声認識を行うに際し，そこまでの認識結果を用いて，次に来る単語を予測することが必要になる（発話仮説）．音声認識においては，認識の対象となる単語の数が大きいほど探索空間が大きくなり，精度が悪くなり，処理時間も大きくなる．言い換えれば，音声認識のある段階で次に発声される単語のバリエーションが少ない方が認識性能は良くなる．

音声認識の過程で，次に発声される単語の種類を確率的に平均した値をパープレキシティと呼ぶ．言語モデルの性能はこのパープレキシティをできるだけ小さく抑えることで評価される．

音声認識の言語モデルとして有望な手法として N-gram モデルが知られている [1]．これは N 個の単語の連鎖を単位として統計的に次の単語を予測する手法であり，モデルのパラメータを大量の言語データから学習することが必要になる．しかし，評価用テキストに未知語が含まれると，パープレキシティは無限大になるという問題があるため，適切な平滑化を行なうことが重要である．

平滑化の手法の一つとしてバックオフ・スムージングがある．これは，例えば N-gram 確率の推定が不確かな時には，(N-1)-gram 確率を用いて平滑化を行う手法である．各種の平滑化手法が提案されているが，本報告では，二項事後分布の継承に基づく BPD バックオフ法に注目し，継承係数の文脈依存性について検討する．継承係数とは，N-gram 確率と(N-1)-gram 確率のバランスを指定する値で，従来法では一つの N に対し，共通の値を設定していたが [2]，理論的には異なる文脈については各々異なる値を設定しても構わない．

本報告は，N-gram 確率のヒストグラムの観察から，文脈をいくつかのクラスタに分類し，各々に異なる継承係数を設定し，パープレキシティがより減少しないか探索する．

2. 二項事後分布に基づく N-gram 言語モデルにおける継承係数の文脈依存性

2.1 パープレキシティ

言語モデルを評価するにあたって言語モデルが言語をどれだけ正確に近似しているかということが問題となる．本報告では評価方法として，情報理論に基づいた評価尺度であるパープレキシティに基づいて言語モデルの評価を行う．

パープレキシティについて説明していく．まず言語は単語列 $w_1^n = w_1 \cdots w_n$ を生成する情報源であると考えることができる．言語 L における単語列の生成確率を $P(w_1^n)$とすれば言語 L のエントロピーは式(1)となる [3].

$$H(L) = -\sum_{w_1^n} \frac{1}{n} P(w_1^n) \log P(w_1^n) \qquad (1)$$

H(L)は言語から生成される単語を特定されるために必要な情報量(ビット)であり，各単語の後には平均して２の H(L)乗の単語が後続可能であることを示している．よって，パープレキシティは式(2)となる．

1 関西学院大学 理工学部
Kwansei Gakuin University. School of Science and Technology

Vol.2022-NL-254 No.9
Vol.2022-SLP-144 No.9
2022/11/29

$$PP = 2^{H(L)} \quad (2)$$

言語のパープレキシティが大きいと単語を特定するのが難しくなり言語として複雑であることを意味している．逆にパープレキシティが低いと単語の候補の数が少ないことを表しており，言語モデルの役割である単語の絞込みがうまく機能していることになる．

パープレキシティのメリットとしては実際に音声認識をしなくてもパープレキシティという値だけで評価が行える点とパープレキシティの単位を単語にとることで，人間にとって意味が理解しやすい値となっていることである．

本報告では，このパープレキシティを用いて言語モデルを評価する．

2.2 言語コーパス

日本語をはじめとする様々な言語を分析する基礎資料として，書き言葉や話し言葉の資料を体系的に収集し，研究用の情報を付与した「ことば」情報のデータベースを言語コーパスと呼ぶ．

日本語話し言葉コーパス(CSJ：Corpus of Spontaneous Japanese) [4]から学習用テキストおよび評価用テキストを抽出する．

CSJ は，日本語の自発音声を大量に集め，情報を付加したデータベースであり，自発音声の音声データに加え，それを書き起こしたテキストを含んでいる．本報告では，この単語に分割されたテキストを用いて，言語モデルの検討を行う．表 1 に学習用テキストおよび評価用テキストの分量を示す．

2.3 N-gram 確率の平滑化

ある時点で生起する事象の確率が，その直前の N 個の時点で生起した事象だけの影響を受けるとき，これを N 重マルコフ過程と呼ぶ．さらに，N-gram モデルというのは，単語の生起を（N－１）重マルコフ過程で近似したモデルのことである．すなわち，N-gram モデルでは，ある時点での単語の生起は直前の（N－１）単語にのみ依存すると考えられる．

表 1 言語モデルの検討に用いる学習用および評価用テキストの構成

	学習用テキスト	評価用テキスト
文数	798,602	76,506
単語数	1,597,204	-
語彙数	63,503	-

ユニグラム (unigram) は，単語が直前の語に影響されずに独立に生起するということであり，これは単語の生起確率と等しい．また，すべての単語が等確率で生起すると考えたモデルのことをゼログラム（zerogram）と呼ぶ．

ここで, バイグラム (bigram) による単語列 w_1^n の生成確率を式(3)に表す．

$$P(w_1^n) = \prod_{i=1}^{n} P\left(w_i | w_{i-1}\right) \quad (3)$$

N-gram 確率は，学習データ中に出現する単語の N 個組と（N-1）個組の相対頻度から次の式(4)のように表すことができる．

$$P(w_n | w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n})}{C(w_{n-N+1}^{n-1})} \quad (4)$$

N の値が大きいほど，学習テキストデータから信頼性の高い N-gram の値を推定するのが難しくなるため，通常は bigram や trigram を用いることが多い．また，評価用テキストに未知語が含まれると，分子がゼロになり，パープレキシティは無限大になる．

2.3.1 BPD バックオフ法

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．BPD バックオフ法では，この bigram 確率を次式のように計算する．

$$\widetilde{P}(w_t | w_{t-1}) = \frac{c(w_{t-1}, w_t) + \gamma_i (c(w_t) + 1)}{c(w_{t-1}) + \gamma_i (N + M)} \quad (5)$$

従来法では γ_i 値は一つの共通の値を与えていた [2]．この場合のベース性能を示しておく．表２に，共通の γ_i の値を変化させ，パープレキシティが最小になるようにした場合の共通 γ_i の値とパープレキシティの値を示す．

表２　共通の γ_i の値を変化させ，パープレキシティが最小になるようにした場合の γ_i の値とパープレキシティの値

	unigram	bigram
Optimum Gamma	2.1	2.0 e-5
Perplexity	72.4	25.7

Vol.2022-NL-254 No.9
Vol.2022-SLP-144 No.9
2022/11/29

i = 0 (unigram) の時のγの最適値はγ_0 = 2.1 であり，そのときのパープレキシティの値は 72.4.

i = 1 (bigram) の時のγの最適値はγ_1 = 2.0 e-5 であり，そのときのパープレキシティの値は 25.7.

本報告では，継承係数γ_iの値を全体で共通にするのではなく，文脈ごとに異なる値を設定することを検討する．(5)式において，左辺値を全ての w_t について加算したものは 1 にならなければならない．すなわち，ある文脈 (直前の単語が w_{t-1}) において(6)式が成立しなければならない．

$$\sum_{w_t} \widetilde{P}(w_t|w_{t-1}) = 1 \tag{6}$$

(6)式に(5)式の値を代入すると，(7)式のように計算が行なえる．

$$\begin{aligned}&\sum_{w_t} \frac{c(w_{t-1},w_t)+\gamma_i(c(w_t)+1)}{c(w_{t-1})+\gamma_i(N+M)}\\ &=\frac{\sum_{w_t}(c(w_{t-1},w_t)+\gamma_i(c(w_t)+1))}{c(w_{t-1})+\gamma_i(N+M)}\\ &=\frac{\sum_{w_t}c(w_{t-1},w_t)+\gamma_i\sum_{w_t}((c(w_t)+1))}{c(w_{t-1})+\gamma_i(N+M)}\end{aligned} \tag{7}$$

ここで，

$$\sum_{w_t} c(w_{t-1},w_t) = c(w_{t-1}) \tag{8}$$

$$\sum_{w_t} c(w_t) = N \tag{9}$$

$$\sum_{w_t} 1 = M \tag{10}$$

(8)式，(9)式，(10)式の関係を用いれば，(7)式は恒等的に 1 であることが証明できる．この計算は，γ_1の値に依存しないので，同じ文脈 (w_{t-1}) に同じ値を設定すれば，任意の値に変更してもよいということがわかる．

このように継承係数γの値は，文脈ごとに違う値を設定することができる．次節以降に実験的検討を行う．

2.3.2 継承係数の文脈依存性の検討

図 1 に，評価テキストに含まれる各単語の bigram 確率に関するヒストグラムを示す．横軸「-対数確率」が 0～5，5～15，15 以上の 3 つの範囲に分布の集中が観察された．各クラスタが確率的に異なる性質を持っていると考えられる．

はじめに，「0＜-対数確率 ＜5」のクラスタに含まれる単語が属する文脈に注目し，その文脈に対して設定する継承係数γ'_1を変化させ，パープレキシティの変化を観察した．この結果を図 2 に示す．この条件下では，特定の文脈に異なる継承係数を与えてもパープレキシティの削減は観察されなかった．

次に，「5＜-対数確率 ＜15」のクラスタに含まれる単語が属する文脈に注目し，その文脈に対して設定する継承係数γ''_1を変化させ，パープレキシティの変化を観察した．この結果を図 3 に示す．この条件下では，特定の文脈に異なる継承係数を与えてもパープレキシティの削減は観察されなかった．

図 3 を観察すると，γ''_1＝2.5 e-5 付近にパープレキシティの最小値がありそうである。より細かくγ''_1を設定して実験を行った結果を図 4 に示す．

このように，5.0＜-log(p)＜15 となる文脈に標準値(2.0 e-5)とは異なるγ'''_1を与えることで，多少ではあるがパープレキシティが減少することが分かった．

最後に，「15＜-対数確率」のクラスタに含まれる単語が属する文脈に注目し，その文脈に対して設定する継承係数γ'''_1を変化させ，パープレキシティの変化を観察した．この結果を図 5 に示す．図の横軸は，継承係数γ'''_1であり，縦軸にパープレキシティを示す．γ'_1 = 5.0 e-5 の時にパープレキシティが最小になり，特定の文脈に異なる継承係数を与えることでパープレキシティが減少することを確認した．

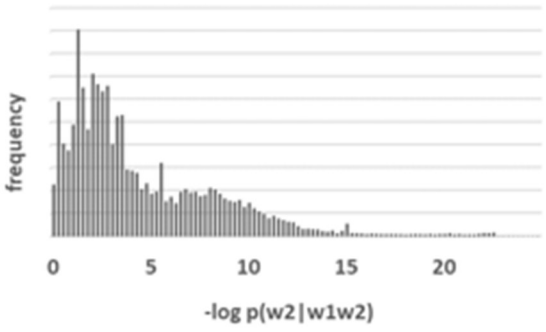

図 1　評価用テキスト中の単語に対する
対数確率のヒストグラム

Vol.2022-NL-254 No.9
Vol.2022-SLP-144 No.9
2022/11/29

3. まとめ

音声認識に用いる N-gram 言語モデルの一つである BPD バックオフ法において,継承係数の文脈依存性を検討する．評価テキストに含まれる各単語の bigram 確率のヒストグラムを観察し，横軸「-対数確率」が 0～5，5～15，15 以上の 3 つの範囲に分布が集中していることを発見した．全ての文脈に同じ継承係数を設定するのではなく，特定の文脈に異なる継承係数γ’1 を与えることでパープレキシティが減少することを確認した．

参考文献

[1] 北研二．確率的言語モデル．東京大学出版社, 1999,
[2] 川端豪，二項事後分布に基づく N-gram 記号連鎖確率の推定．日本音響学会誌, 2005, vol. 65, no. 8, p. 441.
[3] 吉田正太郎．二項事後分布の継承と W-B 平滑化に基づく音声認識のための言語モデル．信学技法 SP2012-118, vol. 112, no. 450, p. 1-2.
[4] 篠崎隆宏，古井貞熙．日本語話し言葉コーパスを用いた講演音声認識．情報処理学会論文誌, 2002, vol. 43, no. 7, p. 2098-2107.

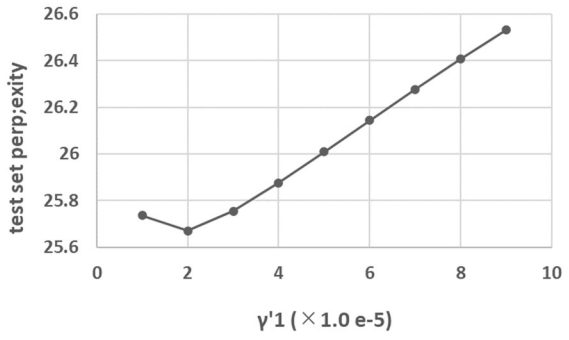

図 2　0 < -log(p) < 5 となる文脈に対し継承係数γ’1 を与えた場合の perplexity の変化

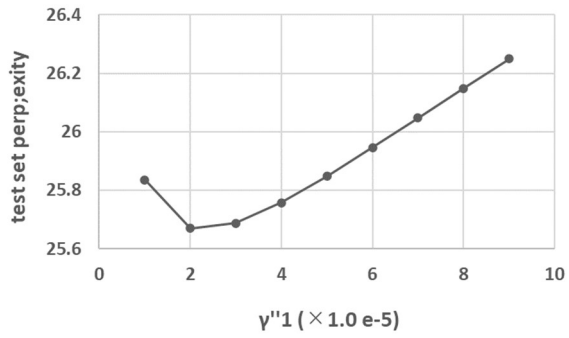

図 3　5 < -log(p) < 15 となる文脈に対し継承係数γ’’1 を与えた場合の perplexity の変化

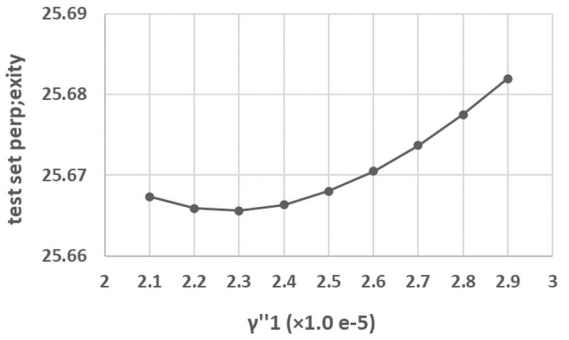

図 4　図 3 の結果をより細かいγ’’1 を設定した場合の perplexity の変化

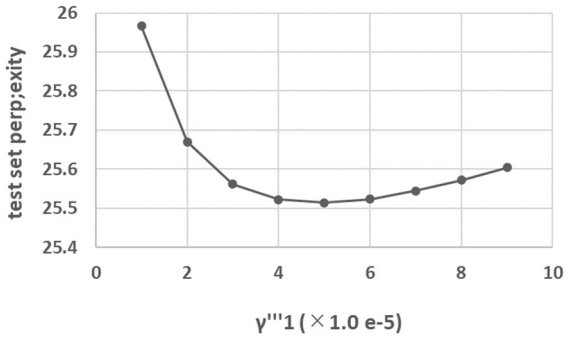

図 5　15 < -log(p)となる文脈に対し継承係数γ’’1 を与えた場合の perplexity の変化