Vol.2012-MPS-90 No.8
2012/9/19

ベイズ決定理論にもとづく階層 N グラムを用いた最適予測法

末永 高志[1,2]　松嶋 敏泰[2]

概要：ユーザ入力をもとにシステムが予測し候補を提示する文書作成支援技術が普及している．この応用を想定し，自然文を構成する単語の生起モデルとして広く使われている，N グラムモデルにもとづいた単語の予測法を検討する．N グラムモデルは，ある単語が生起する確率が直前の $N-1$ 個の単語列に依存するモデルで学習データをもとに構築される．このデータは，単語列を構成する単語の組み合わせが膨大なため一般的には疎となる．モデル構築においては，高次のモデルを低次のモデルで補間するための，混合分布の仮定や，極少数にしか出現しない単語列の出現回数の割引といった方式が提案されてきた．これらの方式は，モデルのパラメータ推定が目的で単語の予測誤りに対する考慮がされていない．本稿では，単語の予測誤りに対して，ベイズ決定理論をもとに考察を行い，誤り率がベイズ基準のもとで最小となる階層 N グラム予測法を示す．更に，実データによる実験を行い，提案法が単語予測に有効であることを示す．

キーワード：ベイズ決定理論，ベイズ基準，N グラム，予測入力

An optimal prediction method using hierarchical N-gram based on Bayes decision theory

Suenaga Takashi[1,2]　Matsushima Toshiyasu[2]

Abstract: Predictive word is an input technology showing candidate words an system predict by user partial input. We treat predictive methods using an N-gram model familiar to the natural language processing field. The model is a contiguous sequence that a next word depends on a sequence in the form of a $N-1$–order Markov model produced by analyzing train data. The data is generally sparse because of enormous combinations of words in the sequences. Many researchers proposed methods to interpolate a lower order model into a higher order one, such as assuming mixture distribution or discounting extremely lower frequent sequence counts. These methods aim parameter estimation and are not considered about prediction errors. In this paper, we discuss prediction errors from a point of view about Bayesian decision theory and present an optimal prediction method with reference to the Bayes criterion for minimizing the errors. Experimental results using real documents show that our method performs good predictive words.

Keywords: Bayesian decision theory, Bayes criterion, N-gram, predictive word

1. はじめに

文字入力インタフェースの制限されたスマートフォン [1]，オフショア開発といった日本語非母語者向け [2] に，ユーザが文字入力した一部の系列データをもとに，単語候補を予測し提示する技術の検討が行われている．この中で，ユーザの入力済みの単語列に対して，N グラムモデルを用いた単語の予測では，この確率モデルをいかに構築するかが課題となる．

N グラムモデルの構築においては，N の数が大きい高次なモデルであるほど得られるデータは疎なため，統計的

[1] 株式会社 NTT データ 技術開発本部
NTT DATA CORPORATION
[2] 早稲田大学 基幹理工学部 応用数理学科
WASEDA UNIVERSITY, School of Fundamental Science and Engineering, Department of Applied Mathematics

な信頼性のあるモデルを構築することが困難になる．一般には，想定するモデルよりも低次なモデルで階層的に補完する平滑化の処理が行われる．これは，モデルの構築にあたり，高次のモデルの確率の推定に対して，低次のモデルの確率の推定結果をもとにした補間をいかに行うかが課題となる．

従来では，それぞれの次数のモデルに対し，重みをつけて足し合わせることが行われていた．この重みに対して，各次数を混合した分布から単語が生成すると仮定し，EM アルゴリズムにより算出した混合比を重みとしたり，1，2 回程度の極低頻度の単語列の出現回数をもとに割引係数を算出して重みの調整が行われていた [3]．この中でも，ニーザー・ネイ法 [4], [5] の有効性が経験的に知られているが，補間を行う形式や割引係数の算出方法について，理論的な説明や性能に対する保証は何もない．

ここで，N グラムの確率モデルの構築を整理すると，N の長さが未知である単語生成モデルに対する問題 [6] といえる．この問題に対してベイズ決定理論にもとづく考察を行い，単語の予測誤りの損失に対してベイズ基準 [7] を最適にする予測法を導出することで，従来研究と形式的に似た，モデルの事後確率で重みづけして足し合わせる方式が求まることを示す．

この方式に対して，日本語の文書データを用た入力支援技術への応用を想定した実験を行い，単語の予測精度を用いて，ニーザー・ネイ法と同程度かそれ以上となることを示す．さらに、学習データが少量の場合に，類似のデータにより事前分布を学習することにより，アルゴリズムを修正することなく単語の予測精度が向上することを示す．

2 節では N グラムモデルを対象にしたモデル構築の従来研究について述べる．3 節では入力支援を想定した単語予測に対し，ベイズ決定理論にもとづく最適予測法を提案する．4 節では 3 節で導出した予測法に対して，日本語文書を用いた実データによる実験を行う．5 節はまとめである．

2. 従来の N グラムモデル構築法

本稿で対象とする N グラムモデルによる単語予測では，系列が単語で構成された単語列ごとにその次に続く単語が確率的に生成すると仮定し，この確率モデルを用いて単語を予測する．

このモデルは，N がある程度の大きさであるほうが予測精度の高さが期待できる．一方で，モデルが高次になるに従いパラメータの数が指数的に増加し，一般に，パラメータの数と比較すると得られるデータは疎となる．このため，履歴となる単語系列に対して予測対象となる単語の組合せの数が少なく，統計的な信頼性のあるモデルを構築することが困難になる．

これに対して，N グラムモデルが階層モデル族であることから，低次のモデルを階層的に補完する処理が適用されてきた [3]．具体的には，N グラムモデルで想定する履歴となる単語列 $\boldsymbol{x}^{N-1} \in \boldsymbol{X}^{N-1}$ が与えられたときに，次に続く単語 $y \in \boldsymbol{Y}$ の確率を，N グラムモデル低次のモデル m とパラメータ $\boldsymbol{\theta}_m$，さらにモデルの重み $w(m)$ を設定して，

$$p(y|\boldsymbol{x}^{N-1}) = \sum_{m\in\{N\}} p(y|\boldsymbol{x}^{m-1}, \boldsymbol{\theta}_m, m)w(m) \qquad (1)$$

と算出する方式が検討されてきた．ただし，$\{N\}$ は次数が N 以下のモデルの集合，$\sum_{m\in\{N\}} w(m) = 1$ である．

モデルの重みは，次数の各モデルが混合した分布から単語が生起すると仮定して，EM アルゴリズムを用いて算出することが広く行われている [3], [5]．この方式は，学習データに対する尤度関数の最大化を考えている．この場合，$\boldsymbol{\theta}_m$ と $w(m)$ を同一データで算出すると最高次数のモデルが常に最尤となるため，$\boldsymbol{\theta}_m$ と $w(m)$ の算出のために異なるデータを用意する必要がある．一般には，学習データを分割することになり，疎なデータに対する対応として望ましくない．

また，極低頻度のデータの影響の制御においては，以下の，

$$\begin{aligned} & P_d(y|\boldsymbol{x}^m) \\ = & \frac{\max\{c(y|\boldsymbol{x}^m) - D, 0\}}{\sum_{y\in\boldsymbol{Y}} c(y|\boldsymbol{x}^m)} \\ & + \frac{D}{\sum_{y\in\boldsymbol{Y}} c(y|\boldsymbol{x}^m)} \big|y : c(y|\boldsymbol{x}^m) > 0\big| P_d(y|\boldsymbol{x}^{m-1}) \end{aligned} \qquad (2)$$

の形式により補間を行う方式が検討されている．ただし，D は $D \geq 0$ とする割引係数，$c(y|\boldsymbol{x}^m)$ は学習データに含まれる $\boldsymbol{x}^m$ の後に y が出現する系列の数，$|\cdot|$ は集合の要素数を表す．

D はいくつかの算出法が検討されているが，m_1 を学習データに 1 回出現した長さ m の単語列の頻度の和，m_2 を学習データに 2 回出現した長さ m の単語列の頻度の和としたときに，m ごとに極低頻度に出現した単語列の頻度を用いて，$D_m = m_1/(m_1 + 2m_2)$ のように算出する方法が提案されている [4], [5]．

これらは，検証データに対するクロス・エントロピーの観点での実験的な検証が中心である．単語の生成プロセスを想定したモデルのパラメータ推定方法の解析は行われているが，単語の予測に関する理論的な解析とはなっていない．

3. ベイズ決定理論にもとづく最適予測法

本節では，N グラムの確率モデルの構築にあたり，N の長さが未知である単語生成モデルに対する問題ととらえ，ベイズ決定理論にもとづいて，単語の予測誤りの損失に対してベイズ基準 [7] を最適にする予測法を導出する．

まず，ここで想定している N グラムモデルを整理すると，真の次数 $m^* \in \{N\}$ とそのパラメータ $\boldsymbol{\theta}_{m^*}$ が存在し，

履歴となる単語列 $\boldsymbol{x}^{N-1}$ が与えられたもとで，

$$p(y|\boldsymbol{x}^{N-1}) = p(y|\boldsymbol{x}^{N-1}, \boldsymbol{\theta}_{m^*}) \tag{3}$$

という確率で単語が生成されると仮定する．また，単語の予測のための決定関数を整理すると，履歴となる単語列とその次に続く単語の n 個の対である学習データ $\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n$，$y^n \in \boldsymbol{Y}^n$ と，単語系列 $\boldsymbol{x}_p^{N-1}$ が得られたもとで $\boldsymbol{x}_p^{N-1}$ の次に続く単語 $y_p \in \boldsymbol{Y}_p$ を予測することになる．これを定式化すると，

$$\hat{y} = D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n) \tag{4}$$

と定義できる．

以上の準備のもとベイズ決定理論にもとづく予測法を以下のように考察する．まず，(4) 式で示される決定関数を用いて，予測した結果の損失関数を定義する．ただし，学習データは確率的に与えられるため，学習データに対して (3) 式で示された真のモデルの分布で期待値を取った危険関数を定義する．この危険関数を最小にする予測法が最適な予測法といえるが，真のモデルの次数 m^* とそのパラメータ $\boldsymbol{\theta}_{m^*}$ は未知のため，これらに事前分布を仮定し，その事前分布に対して期待値をとったベイズ危険関数を最小化することを考える．このベイズ危険関数を最小化する基準をベイズ基準と呼ぶ．

本稿では，最初に簡単のためモデルの次数が既知の場合で議論し，次にモデルの次数が未知の場合の議論を行う．

3.1 真の次数が既知の場合

まず，予測した結果の正誤判定に対して距離

$$d(\hat{y}, y_p) = \begin{cases} 0 & (\hat{y} = y_p) \\ 1 & (\hat{y} \neq y_p) \end{cases} \tag{5}$$

を定義する．これは予測した結果が正しければ 0，誤っていれば 1 の距離をとることを意味する．

この距離に対して，$y_p \in \boldsymbol{Y}_p$ は確率変数であるため，真の分布 $\boldsymbol{\theta}$ で期待値をとった損失関数を定義すると[*1]，

$$\begin{aligned} &L(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p) \\ &= \sum_{y_p \in \boldsymbol{Y}_p} d(D, y_p) p(y_p|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) \end{aligned} \tag{6}$$

となる．

この損失関数を学習データについて期待値をとることで危険関数を定義すると，

$$\begin{aligned} &R(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p|\boldsymbol{\theta}, \boldsymbol{\mu}) \\ &= \sum_{\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n} \sum_{y^n \in \boldsymbol{Y}^n} L(D, \boldsymbol{Y}_p) \\ &\quad p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}) p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) \end{aligned} \tag{7}$$

[*1] 以下，決定関数 $D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n)$ と明らかな場合は D と省略する．

と記述できる．ただし，$\boldsymbol{\mu}$ は $\boldsymbol{x}^{N-1}$ のパラメータとする．

これに対し，パラメータの事前分布 $f(\boldsymbol{\mu})$，$f(\boldsymbol{\theta})$ を仮定し，危険関数を平均化したベイズ危険関数を導出すると，

$$\begin{aligned} &B_{risk}(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p) \\ &= \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\theta}} R(D, \boldsymbol{Y}_p|\boldsymbol{\theta}, \boldsymbol{\mu}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} f(\boldsymbol{\mu}) d\boldsymbol{\mu} \\ &= \sum_{\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n} \sum_{y^n \in \boldsymbol{Y}^n} \sum_{y_p \in \boldsymbol{Y}_p} \\ &\quad \int_{\boldsymbol{\mu}} \Big\{ 1 - \int_{\boldsymbol{\theta}} I_D(y_p) p(y_p|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) \\ &\quad f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \Big\} p(\{\boldsymbol{x}^{N-1}\}^n, y^n) \\ &\quad p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) f(\boldsymbol{\mu}) d\boldsymbol{\mu} \end{aligned} \tag{8}$$

となる．ただし，$I_D(y_p)$ は $D = y_p$ なら 1，$D \neq y_p$ なら 0 を返す関数である．

結局，ベイズ危険関数の最小値は，(8) 式に含まれる

$$\int_{\boldsymbol{\theta}} p(y_p|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \tag{9}$$

を最大化することで得られる．すなわち，

$$\begin{aligned} &\hat{y} = \arg\max_y \\ &\int_{\boldsymbol{\theta}} p(y|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \end{aligned} \tag{10}$$

となる $\hat{y}$ を予測値として出力することが，ベイズ基準のもとでの最適な予測法といえる．

ここで，(10) 式に含まれる予測分布と呼ばれる積分計算は，パラメータ $\boldsymbol{\theta}$ の事前分布 $f(\boldsymbol{\theta})$ にディリクレ分布を仮定することで，多項分布との自然共役の関係から，

$$\begin{aligned} &\int_{\boldsymbol{\theta}} p(y|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \\ &= \frac{c(y|\boldsymbol{x}^{N-1}) + \alpha(y|\boldsymbol{x}^{N-1})}{\sum_{y \in \boldsymbol{Y}} c(y|\boldsymbol{x}^{N-1}) + \sum_{y \in \boldsymbol{Y}} \alpha(y|\boldsymbol{x}^{N-1})} \end{aligned} \tag{11}$$

により容易に求められる [8]．ただし，$\alpha(y|\boldsymbol{x}^{N-1})$ は，$p(y|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta})$ に対応するディリクレ分布のパラメータを表す．

3.2 真の次数が未知の場合

次に，モデルの真の次数 m^* が未知のもとでの N グラムモデルに対する，ベイズ決定理論にもとづく最適な予測法を導出する．なお，距離関数は (5) 式を仮定する．

まず，N グラムモデルを構成するモデル m のパラメータを $\boldsymbol{\theta}_m$，単語の履歴 $\boldsymbol{x}_p^{N-1}$ に含まれる長さ $m-1$ の単語の履歴を $\boldsymbol{x}_p^{m-1}$ とし，各々のモデルで予測する場合の損失関数を定義すると，

$$\begin{aligned} &L_h(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p|m) \\ &= \sum_{y_p \in \boldsymbol{Y}_p} d(D, y_p) p(y_p|\boldsymbol{x}_p^{m-1}, \boldsymbol{\theta}_m, m) \end{aligned} \tag{12}$$

となる．

この損失関数に対する危険関数は，

$$R_h(D(\boldsymbol{x}_p^{N-1},\{\boldsymbol{x}^{N-1}\}^n,y^n),\boldsymbol{Y}_p|m,\boldsymbol{\theta}_m,\boldsymbol{\mu}) = \sum_{\{\boldsymbol{x}^{N-1}\}^n\in\{\boldsymbol{X}^{N-1}\}^n}\sum_{y^n\in\boldsymbol{Y}^n} L_h(D,\boldsymbol{Y}_p|m) \; p(y^n|\{\boldsymbol{x}^{N-1}\}^n,\boldsymbol{\theta}_m,m)p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) \quad (13)$$

と定義できる．

次に，モデル m の事前確率 $p(m)$ とそのパラメータの事前分布 $f(\boldsymbol{\theta}_m)$，$\boldsymbol{\mu}$ の事前分布 $f(\boldsymbol{\mu})$ を仮定すると，ベイズ危険関数は

$$\begin{aligned} & B_{h,risk}(D(\boldsymbol{x}_p^{N-1},\{\boldsymbol{x}^{N-1}\}^n,y^n),\boldsymbol{Y}_p) \\ &= \int_{\boldsymbol{\mu}}\sum_{m\in\{N\}} p(m)\int_{\boldsymbol{\theta}_m} \\ & \quad R_h(D,\boldsymbol{Y}_p|m,\boldsymbol{\theta}_m,\boldsymbol{\mu})f(\boldsymbol{\theta}_m)d\boldsymbol{\theta}_m f(\boldsymbol{\mu})d\boldsymbol{\mu} \\ &= \sum_{\{\boldsymbol{x}^{N-1}\}^n\in\{\boldsymbol{X}^{N-1}\}^n}\sum_{y^n\in\boldsymbol{Y}^n}\sum_{y_p\in\boldsymbol{Y}_p}\int_{\boldsymbol{\mu}}\sum_{m\in\{N\}} \\ & \quad p(m)\int_{\boldsymbol{\theta}_m} d(D,y_p)p(y_p|\boldsymbol{x}_p^{m-1},\boldsymbol{\theta}_m,m) \\ & \quad p(y^n|\{\boldsymbol{x}^{N-1}\}^n,\boldsymbol{\theta}_m,m)f(\boldsymbol{\theta}_m)d\boldsymbol{\theta}_m \\ & \quad p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu})f(\boldsymbol{\mu})d\boldsymbol{\mu} \end{aligned} \quad (14)$$

となる．ベイズ危険関数の最小値は (14) 式の，

$$\begin{aligned} & \sum_{m\in\{N\}} p(m)\int_{\boldsymbol{\theta}_m} d(D,y_p)p(y_p|\boldsymbol{x}_p^{m-1},\boldsymbol{\theta}_m,m) \\ & \quad p(y^n|\{\boldsymbol{x}^{N-1}\}^n,\boldsymbol{\theta}_m,m)f(\boldsymbol{\theta}_m)d\boldsymbol{\theta}_m \\ &= 1-\sum_{m\in\{N\}}\int_{\boldsymbol{\theta}_m} I_D(y_p)p(y_p|\boldsymbol{x}_p^{m-1},\boldsymbol{\theta}_m,m) \\ & \quad f(\boldsymbol{\theta}_m|y^n,\{\boldsymbol{x}^{N-1}\}^n,m)d\boldsymbol{\theta}_m p(m|\{\boldsymbol{x}^{N-1}\}^n,y^n) \end{aligned} \quad (15)$$

を最小化することで得られ，ベイズ基準のもとでの最適な予測法は

$$\begin{aligned} \hat{y} = \arg\max_y & \sum_{m\in\{N\}}\int_{\boldsymbol{\theta}_m} p(y|\boldsymbol{x}_p^{N-1},\boldsymbol{\theta}_m,m) \\ & f(\boldsymbol{\theta}_m|\{\boldsymbol{x}^{N-1}\}^n,y^n,m)d\boldsymbol{\theta}_m \\ & p(m|\{\boldsymbol{x}^{N-1}\}^n,y^n) \end{aligned} \quad (16)$$

となる $\hat{y}$ を出力することになる．これは，形式的には従来研究の (1) 式と類似しているが，低次のモデルごとに予測分布を求め，モデルの事後確率による重みで足し合わせたものが，単語の予測誤りの損失に対するベイズ基準を最適にするモデルといえる．

なお，予測分布は (11) 式と同様の形式で，モデルの事後確率はベイズの定理より求まる．

4. 実データによる単語予測実験

本節では，特定業務に対する入力支援を想定して，提案する予測法の効果を日本語の特許文書とシステム開発文書を対象に検証する．

検証では，既存の混合分布を仮定した方式，ニーザー・ネイ法，提案法のそれぞれで，学習データからモデルを構築し，このモデルをもとに検証データに対する単語予測の実験を行う．この実験では予測の正答率をもとに各方式の比較を行い，提案法が実用面においても有効であることを示す．

また，実応用で学習データの量が少ない場合においては，類似する他の業務で作成された文書を事前知識としてモデルの構築に活用することが行われる．提案法は，(16) 式の導出で仮定をおいたモデル m とパラメータ $\boldsymbol{\theta}_m$ の事前分布として，類似文書のデータにより学習した結果を導入することが可能であるという特徴を持つ．

そこで，学習データの量が少ない場合を想定し，事前分布の設定を，事前知識がない場合を無情報事前分布 [9]，事前知識がある場合を類似データにより学習した事前分布として，それぞれモデルを構築し検証データを用いた単語予測の正答率の比較を行う．これにより，類似データの学習結果を利用することで，学習データの量が少ない場合に予測の正答率が向上することを示す．

4.1 文書データの条件

対象とするデータは日本語の文書から名詞，助詞，動詞と連続する単語列を抽出し，さらに履歴のデータとしてこの単語列よりも前に出現する単語を品詞を区別することなく抽出することで作成した．また，助詞を含めそれより前の系列を履歴となる単語列，予測対象とする単語を動詞とした[*2]．

特許文書は，1,000 件の公開特許公報を無作為に選定し，上記の処理を行い 93,320 件の単語列を抽出した．システム開発文書は，いくつかのシステムの設計書やマニュアル等を含む 4,423 件の文書から 119,146 件の単語列を抽出した．さらに，これらの単語列のデータを学習データと検証データに文書種類ごとにそれぞれ同数に分割した．なお，学習データに含まれる予測対象となる動詞の単語は，特許文書が 5,409 種類，システム開発文書が 1,989 種類であった．

4.2 単語予測の正答率

単語予測の正答率の評価にあたっては，既存法として 2 節で説明した，混合分布を仮定したモデル，ニーザー・ネイ法を用いる．提案は (16) 式にもとづく予測法を用いる．それぞれの方式対し，学習データをもとにモデルを構築し，検証データによる単語予測の実験を行う．

[*2] 文書データに対する単語系列の分割および品詞の付与は，形態素解析ツール MeCab http://mecab.sourceforge.net/を利用した．

混合分布のモデル構築は，学習データを二つに分割し，一方を，単語の出現確率となるパラメータの算出，もう一方をモデルの重みの算出に用いた EM アルゴリズム [3] により実施した．また，データを入れ替えてパラメータと重みの算出を再度行い，モデルの重みは算出された二つの結果の平均とした．この後，学習データの全体で単語の出現確率となるパラメータをあらためて算出し，(1) 式に従い混合を行った．

ニーザー・ネイ法は，現在，高性能として広く知られている修正ニーザー・ネイ法 [5] を用いた．この方式は (2) 式に含まれる D の値の算出に対して，2 節で示した方法を拡張して，学習データに出現する長さ m の単語列に対して，1 から 3 回まで出現した単語列の頻度を考慮したものである．

また，提案法で用いるモデル m の事前確率は，データ圧縮の分野で広く使われている次数が高くなるに従い値を小さくする，2^{-m}，ただし $m = N$ の場合は 2^{N-1} とする方法で与えた．$\boldsymbol{\theta}_m$ の事前分布とするディリクレ分布のパラメータは，学習データに含まれる予測対象となる動詞の単語の種類数の逆数で与えることとした．

比較する N グラムモデルの N は 3 から 6 まで行い，各方式に対して，履歴となる単語系列が到達できる最高次のモデルで予測することとした．

利用シーンを考慮すると候補を複数提示しその中から適切なものを選択することも可能である．単語予測の評価においては，予測に使う単語の確率が最上位であった単語を一つ出力し，検証データの正解となる単語の一致した割合を表す正答率と，確率が大きいものから上位 5 件の単語を出力し，検証データの正解となる単語が含まれていれば正解とする正答率の二種類で行った．

検証データによる正答率の結果を表 1 に示す．表中の N は構築したモデルの単語列の最大長で，特許文書，システム開発文書のそれぞれに対する正答率を示している．最上位は，予測に使う単語を確率が最上位のものを一つ出力した場合，上位 5 位は，予測に使う単語を確率が大きいものから五つ出力した場合を表す．混合は EM アルゴリズムにより求められた混合分布を仮定したモデル，MKN は修正ニーザー・ネイ法，提案は提案法をそれぞれ指す．また，太文字は正答率の最良値である．

今回検討した範囲では，すべての方式で N を増加させることで正答率が向上している．文書ごとの傾向を確認すると，特許文書では，最上位の正答率と，上位 5 位を出力した場合の正答率の双方で，差は小さいものの提案法による予測が最良値となり，システム開発文書では，修正ニーザー・ネイ法が最良値となる場合と，提案法が最良値となる場合の双方の結果が見られた．ただし，正答率の差は 0.1% 程度でこちらも差は小さいといえる．

この結果から，実証的な検討から有効性が知られていた修正ニーザー・ネイ法とほぼ同等の単語予測の性能を持つといえ，理論的な最適性に加え，実証的な観点からも有効に機能するといえる．

4.3 事前知識の利用

本節では，学習データが少量の場合を想定し，提案法に対して類似したデータで学習した事前分布を導入することを効果を検証する．具体的には，学習データによるモデル構築時に用いる事前分布の設定を，無情報事前分布とした場合と類似したデータで学習した事前分布を導入した場合のそれぞれに対し，単語予測の実験を行い比較する．これは，それぞれ事前知識を利用しない場合と，利用する場合に相当する．

事前知識とする類似データは特許文書のデータとし，学習に用いる少量データはシステム開発文書の学習データから一部のデータを無作為に抽出し作成した．なお，業務継続におけるデータの増加を想定し，データ量を増やす場合は，抽出ずみのデータに対して追加することとした．

事前分布を無情報事前分布とする場合は，4.2 節の提案法と同様の設定とした．類似データの学習は特許文書はすべてのデータを用いて，4.2 節の提案法と同様の設定でモデルの事後確率と履歴となる単語列ごとに予測対象とする単語の出現頻度を算出し，それぞれを学習データに対する事前分布として利用した．

なお，単語の出現頻度はディリクレ分布のパラメータとして利用することになる．このパラメータは観測されたデータの個数に相当し，学習データに対する影響が大きすぎることがないよう調整が必要である．各々の履歴となる単語列に対して，出現する単語の頻度の和が $\rho > 0$ となるよう調整し，さらに，事前分布を無情報事前分布とする場合のパラメータに加算することとした．本実験では，履歴となる単語列が事前知識用のデータと学習データの双方に含まれる場合に，学習データの影響を優先することを狙い $\rho = 0.01$ とした．

$N = 3$ とした実験結果を表 2 に示す．表中の，学習データの件数は学習データとして用いたシステム開発文書のデータ件数を表す．事前知識なしが類似データを利用しない場合，事前知識ありが類似データを利用する場合である．この二つの結果に対して，予測に正答するか誤答するかが二項分布に従うと仮定して，母不良率の検定を行った．ここで，有意水準は 5% と 1% のそれぞれで行った．無為の場合は差がなく，有意の場合は差があることを意味する．

検定結果を確認すると，学習データの量が少ない場合は有意な差がみられるが，データ量が増えることで有意な差がなくなっている．また，正答率の差も学習データ量が増えるに従い小さくなる傾向がみられ，データが少量の場合は事前知識を活用し，データの増加に従い事前知識の影響が小さくなる性質をもつといえる．

表 1 各方式による単語予測結果の正答率（単位 %）

Table 1 Accuracy of word prediction using each method.

	特許文書						システム開発文書					
	最上位			上位 5 位を出力			最上位			上位 5 位を出力		
N	混合	MKN	提案	混合	MKN	提案	混合	MKN	提案	混合	MKN	提案
3	37.07	44.51	**44.52**	61.07	64.83	**65.00**	46.34	**52.17**	**52.17**	71.06	78.58	**78.73**
4	46.47	46.68	**46.70**	65.88	66.14	**66.16**	64.16	**64.43**	64.41	83.53	**83.93**	83.83
5	49.92	52.48	**52.49**	67.55	**68.99**	**68.99**	67.86	73.16	**73.19**	84.73	**86.26**	86.17
6	52.76	53.51	**53.57**	68.20	**69.14**	**69.14**	75.48	**75.84**	75.65	86.00	**86.57**	86.52

表 2 $N = 3$ での事前知識の利用有無での単語予測結果の比較（単位 %）

Table 2 Comparison of methods derived from pre and train data with just train one in $N = 3$.

	最上位				上位 5 位を出力			
学習データ件数	事前知識なし	事前知識あり	有意水準 5% の検定結果	有意水準 1% の検定結果	事前知識なし	事前知識あり	有意水準 5% の検定結果	有意水準 1% の検定結果
233	22.47	23.70	有為	有為	38.31	39.63	有為	有為
465	26.86	27.75	有為	有為	44.99	45.45	有為	無為
931	34.31	34.15	無為	無為	52.29	52.09	無為	無為
1862	37.31	37.14	無為	無為	57.53	57.33	無為	無為

5. おわりに

N グラムモデルをもとにした単語の予測に用いる確率モデルの構築の問題に対し，N の長さが未知である単語生成モデルに対する問題ととらえ，単語の予測誤りの損失に対してベイズ基準を最適にする予測法を導出した．本提案は，形式的には従来研究と類似しているが，重みの算出に対して，モデルの事後確率で重みづけして足し合わせる特徴をもつ．

実データによる実験により，現在，自然言語処理の分野で高性能と知られている修正ニーザー・ネイ法と比べてほぼ同等，もしくはやや上回ることを示した．また，学習データが少量しか得られない場合において，事前知識の利用が容易に行え予測の正答率を向上させることと，学習データ量が増えるに従い事前知識の影響が小さくなる性質をもつことを実験により示した．

実応用を想定すると，学習データが少量しか得られない場合でも，業務を継続することで対象データは増加することが想定される．単語予測モデルの個人適用や業務適用を想定した場合，予測法の性質として，データ量の増加に従い事前知識の影響が小さくなり，追加されたデータの影響が大きくなることが好ましいといえる．本提案法は，システム提供者が用意した事前知識が適用先との適合度合いが小さかったとしても，事前知識の影響度合いの調整なしにデータを追加するだけでモデル構築の可能性が示唆される．

今後の課題としては，単語予測の高精度化について，階層 N グラムモデルに限定しない様々なモデルの提案法への導入や，学習データが少量な場合の事前知識を導入する場合において，事前知識の影響を適切に調整する方式の検討があげられる．

参考文献

[1] 小町守, 木田泰夫: スマートフォンにおける日本語入力の現状と課題, 言語処理学会第 17 回年次大会, 言語処理学会, pp. 1095–1098 (2011).
[2] 末永高志, 松嶋敏泰: ベイズ決定理論にもとづく階層 N グラムを用いた最適予測法と日本語入力支援技術への応用, 言語処理学会第 18 回年次大会, 言語処理学会, pp. 6–9 (2012).
[3] 北研二: 言語と計算–4 確率的言語モデル, 東京大学出版会 (1999).
[4] Kneser, R. and Ney, H.: Improved backing-off for m-gram language modeling, *Proceedings of ICASSP*, Vol. 1, Association for Computational Linguistics Morristown, NJ, USA, pp. 181–184 (1995).
[5] Chen, S. F. and Goodman, J.: An Empirical Study of Smoothing Techniques for Language Modeling, *Proceedings of ACL*, pp. 310–318 (1996).
[6] 松嶋敏泰: 統計モデル選択の概要, オペレーションズ・リサーチ, Vol. 41, No. 7, pp. 369–374 (1996).
[7] 松嶋敏泰: 帰納・演繹推論と予測–決定理論による学習モデル–, 情報論敵学習理論ワークショップ予稿集, 情報理論とその応用学会 (1998).
[8] Bishop, C. M.: パターン認識と機械学習 上 – ベイズ理論による統計的予測, シュプリンガー・ジャパン, 東京 (2007).
[9] 繁桝算男: ベイズ統計入門, 東京大学出版会 (1985).