情報処理学会論文誌　数理モデル化と応用　Vol.6　No.1　102–110 (Mar. 2013)

ベイズ決定理論に基づく階層 N グラムを用いた最適予測法

末永 高志[1,2,a)]　松嶋 敏泰[2]

受付日 2012年8月22日，再受付日 2012年10月11日，
採録日 2012年10月29日

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．この応用を想定した N グラムモデルを用いた単語の予測法を検討する．N グラムモデルは学習データをもとに構築されるが，単語列を構成する単語の組合せが膨大なため次数が増加するにつれ疎となる．モデル構築においては，高次のモデルの推定に対し，低次のモデルの推定結果をもとにいかに補間するかが課題である．従来は混合分布の仮定や，ごく少数にしか出現しない単語列を考慮した割引係数をもとに，各次数のモデルを重みづけ足し合わせることが行われていたが，予測に対する理論的な保証はない．本稿では，これに対し真の次数が未知の統計問題ととらえ，ベイズ決定理論に基づいて，単語の予測誤りの損失に対しベイズ基準のもとで最小となることが保証された予測法を導出する．さらに，日本語文書データでの単語予測の実験を行い，提案法が実用的にも有効であることを示す．

キーワード：次数未知の N グラムモデル，ベイズ決定理論，ベイズ基準，予測入力

An Optimal Prediction Method Using Hierarchical N-gram Based on Bayesian Decision Theory

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

Received: August 22, 2012, Revised: October 11, 2012,
Accepted: October 29, 2012

Abstract: Predictive word is an input technology showing candidate words which a system predict by user partial input. We treat predictive methods using an N-gram model. The model is generally produced by analyzing train data. The data is more sparse in proportion to an N–gram order, because of enormous combinations of words in the sequences. An issue of producing the model is how to combine a lower order model into a higher order one. Many researchers proposed models composed of weighed each-order one, such as a mixture distribution or an interpolation created by discount parameters considering about extremely lower frequent sequence. But these methods have no theoretical guarantee about prediction errors. In this paper, we treat the issue as a statistical problem that the model order is unknown, and discuss prediction errors from a point of view about Bayesian decision theory. We present that an optimal prediction method with reference to the Bayes criterion for minimizing the errors. Experimental results using Japanese documents show that our method performs good predictive words.

Keywords: unknown order of N–gram model, Bayesian decision theory, Bayes criterion, 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]，業務文書の表現の統一 [2]，オフショア開発といった日本語非母語者向け入力支援 [3] 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

[1] 株式会社 NTT データ技術開発本部
Research and Development Headquarters, NTT DATA CORPORATION, Koto, Tokyo 135–8671, Japan
[2] 早稲田大学基幹理工学部応用数理学科
Department of Applied Mathematics, School of Fundamental Science and Engineering, WASEDA UNIVERSITY, Shinjuku, Tokyo 169–8555, Japan
[a)] suenagatk@nttdata.co.jp

© 2013 Information Processing Society of Japan

列に対して，N グラムモデルを用いて予測した単語の候補を提示することを考えると，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，統計的に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，平滑化の処理が行われるが，この処理は，ゼロ頻度の場合のみ低次の情報で補間するバックオフ [4] や想定するモデルよりも低次なモデルをつねに一定の割合で足し合わせる内挿 [5], [6] により行われる．モデルの構築にあたっては，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いえる．

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 回程度のごく低頻度の単語列の出現回数により算出された割引係数を用いて，重みの調整が行われていた [4]．この中でも，ニーザー・ネイ法 [5], [6] 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，単語の予測に対する理論的な説明や性能に対する保証はない．

ここで，N グラムの確率モデルの構築は，N の長さが未知である単語生成モデルの統計問題といえる．これは，ある単語の生起する確率が直前の何個の単語列に依存しているかが，モデル構築のさいに明確でないことを意味している．

本稿では，N の長さが未知である問題に対してベイズ決定理論に基づく考察を行い，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．これにより，モデルの予測分布 [8] に対してモデルの事後確率で重みづけして足し合わせた，従来研究と形式的に類似した方式が，ベイズ基準のもとで最適な予測法であることと，計算が容易でアルゴリズム化しやすいことを示す．

また，予測法の導出では，モデルの事前確率と各次数のモデルが持つ単語の生成に関する分布のパラメータに対して事前分布を仮定する．このことから，類似のデータにより学習した結果を，モデル構築時に事前知識として利用することが容易である．学習データが少量の場合には，類似データの利用により単語予測の正答率の向上が期待できる．

提案に対して，日本語の文書データによる入力支援技術への応用を想定した実験を行い，単語予測の正答率を用いて，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予測の正答率が向上することを示す．

2 章では 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 章では入力支援を想定した単語の予測に対し，ベイズ決定理論に基づく最適予測法を提案する．4 章では 3 章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 章はまとめである．

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の予測では，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されると仮定し，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となる高次のモデルであるほうが予測の性能が期待できる．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 [8] 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 [4]．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，各次数のモデルが混合した分布から単語が生起されると仮定して，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 [4], [6]．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つねに最尤となるため，$\boldsymbol{\theta}_m$ 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，ごく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

© 2013 Information Processing Society of Japan

数を表す．すなわち，$\boldsymbol{x}^m$ の後に出現した y の種類数を意味する．

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，

$$D_m = \frac{m_1}{m_1 + 2m_2} \tag{3}$$

のように算出する方法が提案されている [5], [6]．

この形式は Pitman–Yor 過程と呼ばれる確率過程の近似となっていることが指摘されている [9], [10], [11] が，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. 真の次数を未知とした N グラムモデル構築法

本章では，N グラムモデルに対して，N の長さが未知である単語生成モデルを構築する問題ととらえ，ベイズ決定理論に基づいて，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，単語が生成される真の確率を，

$$p^*(y|\boldsymbol{x}^{N-1}) = p(y|\boldsymbol{x}^{N-1}, \boldsymbol{\theta}_{m^*}) \tag{4}$$

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，

$$\hat{y} = D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n) \tag{5}$$

と定義できる．

以上の準備のもとベイズ決定理論に基づく予測法を以下のように考察する．まず，式 (5) 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式 (4) で示された真のモデルの分布で期待値をと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がベイズ基準と呼ばれる．

本稿では，最初に簡単のため真のモデルの次数が既知の場合で議論し，次に真のモデルの次数が未知の場合を議論する．

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

$$d(\hat{y}, y_p) = \begin{cases} 0 & (\hat{y} = y_p) \\ 1 & (\hat{y} \neq y_p) \end{cases} \tag{6}$$

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]，

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

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，

$$\begin{aligned} &R(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p|\boldsymbol{\theta}, \boldsymbol{\mu}) \\ &= \sum_{\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n} \sum_{y^n \in \boldsymbol{Y}^n} L(D, \boldsymbol{Y}_p) \\ &\quad p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}) p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) \end{aligned} \tag{8}$$

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 $\boldsymbol{\mu}$ と $\boldsymbol{\theta}$ が独立であること[*2]と，事前分布 $f(\boldsymbol{\mu})$，$f(\boldsymbol{\theta})$ の存在を仮定し，危険関数を平均化したベイズ危険関数を導出すると，

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

[*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．

[*2] この仮定は多くの学習理論研究にて暗黙のうちに前提となってることが，文献 [7] において指摘されている．

© 2013 Information Processing Society of Japan

$$
\begin{aligned}
&= \sum_{\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n} \sum_{y^n \in \boldsymbol{Y}^n} \sum_{y_p \in \boldsymbol{Y}_p} \\
&\int_{\boldsymbol{\mu}} \Big\{ 1 - \int_{\boldsymbol{\theta}} I_D(y_p) p(y_p|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) \\
&f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \Big\} p(y^n|\{\boldsymbol{x}^{N-1}\}^n) \\
&p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) f(\boldsymbol{\mu}) d\boldsymbol{\mu} \qquad (9)
\end{aligned}
$$

となる[*3]．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，式 (9) に含まれる

$$
\begin{aligned}
&1 - \int_{\boldsymbol{\theta}} I_D(y_p) p(y_p|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) \\
&f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \qquad (10)
\end{aligned}
$$

を最小化することで得られる．すなわち，

$$
\begin{aligned}
&\hat{y} = \arg\max_y \\
&\int_{\boldsymbol{\theta}} p(y|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \qquad (11)
\end{aligned}
$$

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，式 (11) 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，

$$
\begin{aligned}
&\int_{\boldsymbol{\theta}} p(y|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \\
&= \frac{c(y|\boldsymbol{x}^{N-1}) + \alpha(y|\boldsymbol{x}^{N-1})}{\sum_{y \in \boldsymbol{Y}} c(y|\boldsymbol{x}^{N-1}) + \sum_{y \in \boldsymbol{Y}} \alpha(y|\boldsymbol{x}^{N-1})} \qquad (12)
\end{aligned}
$$

により容易に求められる [12], [13]．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，$c(y|\boldsymbol{x}^{N-1})$ は式 (2) と同様に学習データに含まれる $\boldsymbol{x}^{N-1}$ の後に出現する y の頻度をそれぞれ表す．

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，ベイズ決定理論に基づく最適な予測法を導出する．なお，距離関数は式 (6) 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，

$$
\begin{aligned}
&L_h(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p|m, \boldsymbol{\theta}_m) \\
&\quad = \sum_{y_p \in \boldsymbol{Y}_p} d(D, y_p) p(y_p|\boldsymbol{x}_p^{m-1}, \boldsymbol{\theta}_m, m) \qquad (13)
\end{aligned}
$$

となる．

この損失関数に対する危険関数は，

$$
\begin{aligned}
&R_h(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p|m, \boldsymbol{\theta}_m, \boldsymbol{\mu}) \\
&= \sum_{\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n} \sum_{y^n \in \boldsymbol{Y}^n} L_h(D, \boldsymbol{Y}_p|m, \boldsymbol{\theta}_m) \\
&\qquad p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}_m, m) p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) \qquad (14)
\end{aligned}
$$

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

$$
\begin{aligned}
&B_{h,risk}(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p) \\
&= \int_{\boldsymbol{\mu}} \sum_{m \in \{N\}} p(m) \int_{\boldsymbol{\theta}_m} \\
&R_h(D, \boldsymbol{Y}_p|m, \boldsymbol{\theta}_m, \boldsymbol{\mu}) f(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m f(\boldsymbol{\mu}) d\boldsymbol{\mu} \\
&= \sum_{\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n} \sum_{y^n \in \boldsymbol{Y}^n} \sum_{y_p \in \boldsymbol{Y}_p} \int_{\boldsymbol{\mu}} \sum_{m \in \{N\}} \\
&p(m) \int_{\boldsymbol{\theta}_m} d(D, y_p) p(y_p|\boldsymbol{x}_p^{m-1}, \boldsymbol{\theta}_m, m) \\
&p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}_m, m) f(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\
&p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) f(\boldsymbol{\mu}) d\boldsymbol{\mu} \qquad (15)
\end{aligned}
$$

となる．ベイズ危険関数の最小値は式 (15) に含まれる，

$$
\begin{aligned}
&\sum_{m \in \{N\}} p(m) \int_{\boldsymbol{\theta}_m} d(D, y_p) p(y_p|\boldsymbol{x}_p^{m-1}, \boldsymbol{\theta}_m, m) \\
&p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}_m, m) f(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\
&= 1 - \sum_{m \in \{N\}} \int_{\boldsymbol{\theta}_m} I_D(y_p) p(y_p|\boldsymbol{x}_p^{m-1}, \boldsymbol{\theta}_m) \\
&f(\boldsymbol{\theta}_m|y^n, \{\boldsymbol{x}^{N-1}\}^n, m) d\boldsymbol{\theta}_m p(m|\{\boldsymbol{x}^{N-1}\}^n, y^n) \\
&/ p(y^n|\{\boldsymbol{x}^{N-1}\}^n) \qquad (16)
\end{aligned}
$$

を最小化することで得られる．結局，ベイズ基準のもとでの最適な予測法は，$p(y^n|\{\boldsymbol{x}^{N-1}\}^n)$ が定数で無視できるため，

$$
\begin{aligned}
&\hat{y} = \arg\max_y \sum_{m \in \{N\}} \int_{\boldsymbol{\theta}_m} p(y|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}_m, m) \\
&f(\boldsymbol{\theta}_m|\{\boldsymbol{x}^{N-1}\}^n, y^n, m) d\boldsymbol{\theta}_m \\
&p(m|\{\boldsymbol{x}^{N-1}\}^n, y^n) \qquad (17)
\end{aligned}
$$

となる $\hat{y}$ を出力することになる．

これは，従来研究の式 (1) と形式的に類似している．す

[*3] ここで，$\{\boldsymbol{x}^{N-1}\}^n$ と $\boldsymbol{\theta}$ が独立で，

$$
\begin{aligned}
&f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) p(\{\boldsymbol{x}^{N-1}\}^n, y^n) \\
&= p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}) p(\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}) \\
&= p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}) p(\{\boldsymbol{x}^{N-1}\}^n) f(\boldsymbol{\theta})
\end{aligned}
$$

が成立することから，

$$
\begin{aligned}
&p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}) f(\boldsymbol{\theta}) \\
&= f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) p(y^n|\{\boldsymbol{x}^{N-1}\}^n)
\end{aligned}
$$

の関係を利用した．

© 2013 Information Processing Society of Japan

なわち，式 (1) に対して，$p(y|\boldsymbol{x}^{m-1}, \boldsymbol{\theta}_m, m)$ を m の予測分布，$w(m)$ を m の事後確率としたものが，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，予測分布は式 (12) と同様の形式で，モデルの事後確率はベイズの定理より，

$$\begin{aligned} & p(m|\{\boldsymbol{x}^{N-1}\}^n, y^n) \\ \propto\ & p(\{\boldsymbol{x}^{N-1}\}^n, y^n|m)p(m) \\ =\ & p(m)\prod_{i=1}^{n} p(\{\boldsymbol{x}^{N-1}\}_i, y_i|m) \\ =\ & p(m)\prod_{i=1}^{n}\int_{\boldsymbol{\theta}} p(y_i, \{\boldsymbol{x}^{N-1}\}_i, \boldsymbol{\theta}|m)d\boldsymbol{\theta} \\ \propto\ & p(m)\prod_{i=1}^{n}\int_{\boldsymbol{\theta}} p(y_i|\{\boldsymbol{x}^{N-1}\}_i, \boldsymbol{\theta}, m)f(\boldsymbol{\theta}|m)d\boldsymbol{\theta} \qquad (18) \end{aligned}$$

から容易に求まる．

3.3　提案法の実現方法

本稿で対象としている N グラムモデルでは，履歴となる $N-1$ 個の単語列 $\boldsymbol{x}^{N-1}$ に依存して単語 y が生成する．このような構造に対して広く利用されている，接尾木 [4] を利用した提案法の実現方法を説明する．

まず，接尾木は**図 1** に示すように単語列を節点とし，各枝には対応する単語，各節点には生成した単語の情報を保持する構造となっている．たとえば，図中の四角で囲まれた「を」,「に」,「処理」,「ボタン」は，履歴となる単語列を構成する単語を意味し，ある節点から最上位にある根の節点へ至る経路をたどることで，その節点に対応する単語列 $\boldsymbol{x}^{N-1}$ は復元できる．なお，根の節点 ϵ は空の単語列を表している．また，この図では各節点から生成した単語の頻度の情報を保持した例を示している．

モデルの構築にあたっては，最初に，接尾木を構築する．具体的には，学習データの $\boldsymbol{x}^{N-1}$ をもとに根の節点から対応する枝をたどり，たどれない場合は枝と節点を追加する．また，各節点に対応する単語列の次に出現した単語の頻度を保持する．次に，式 (12) をもとに各節点に対応する確率分布を求める．さらに，式 (18) をもとに学習データから各次数の事後確率を算出する．最後に，構築した接尾木を根からすべての節点をたどり，各節点に対して，その次数を M として，順次，式 (17) に含まれる，

$$\begin{aligned} & \sum_{m\in\{M\}}\int_{\boldsymbol{\theta}_m} p(y|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}_m, m) \\ & f(\boldsymbol{\theta}_m|\{\boldsymbol{x}^{N-1}\}^n, y^n, m)d\boldsymbol{\theta}_m p(m|\{\boldsymbol{x}^{N-1}\}^n, y^n) \quad (19) \end{aligned}$$

を計算し，節点に対応する単語の情報として保持する．このように，接尾木を用いることから使用するメモリに必要な量は節点の数と，各節点で保持する単語の種類数の和に比例する[*4]．

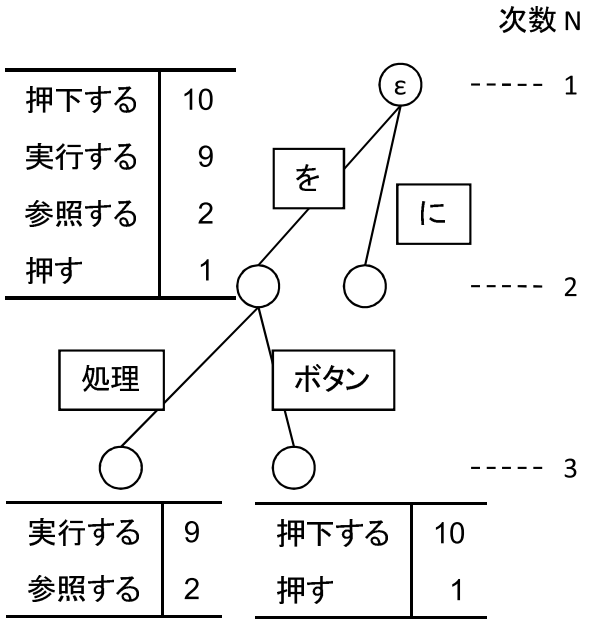

図 1　単語予測に用いる接尾木の例
Fig. 1　An example of term prediction suffix tree.

予測を行う場合は，与えられた $\boldsymbol{x}_p^{N-1}$ を根の節点から単語が一致する枝をたどり，到達した節点が保持した値をもとに確率値が最大である予測値 y_p を出力する．なお，$\boldsymbol{x}_p^{N-1}$ と完全に一致する単語列が学習データに含まれない場合も存在し，その場合はいくつかの対応法が考えられるが，本稿では到達できる最高次数の節点を用いて予測することとする．

なお，2 章で取り上げた従来法も同様の接尾木で実現でき，学習に要する処理量は異なるが，予測は提案法と同様の処理で予測値を出力するため，使用するメモリ量および計算量は提案法と同等である．

次に，学習に要する処理量の評価にあたり，手順をステップに分け提案法の計算量をそれぞれ評価し，従来法と差分となるステップについて比較する．

上記で説明したとおり，学習の手順は以下の，

(1) 接尾木の構築
(2) 各節点の単語の出現確率の算出
(3) 補間に使う重みの算出
(4) 階層的に補間

の 4 つのステップに分けられる．

(1) は，学習データに含まれる単語列ごとに構築済みの接尾木の節点をたどり，節点が存在しない場合は追加の処理を行う．そのため，計算量は学習データ数 n と最大次数 $N-1$ に比例する．(2) は，各節点に対して式 (12) を計算するので計算量は節点の数と各節点に保持された単語の種

[*4] なお，メモリを最大に利用する場合は，$|\boldsymbol{Y}||\boldsymbol{X}|^{N-1}$ に比例する量が必要となる．ただし，学習データにて出現した単語のみ保持すればよく，文書データは疎なことが多いため，最大量を必要とすることはまれである．

© 2013 Information Processing Society of Japan

類数の和に比例する．(3) は，学習データごとに式 (18) を計算するので，計算量は学習データ数 n と最大次数 $N-1$ に比例する．(4) は，高次の節点に対して低次の節点の加算を行うので，計算量は節点の数と各節点に保持された単語の種類数の和に比例する．

EM アルゴリズムを用いる方式の異なる点としては，学習データを 2 つに分割し，(1) と (2) のステップを一方のデータで行い，(3) のステップはもう一方のデータで，モデルの重み $w(m)$ の値が収束するまで計算を行う点である．したがって，モデルの重み $w(m)$ の算出の繰返し計算の部分が，提案法よりも増加する．

また，ニーザー・ネイ法を用いる場合は (3) のステップが異なる．これは，式 (2) に含まれる D を学習データ全体から算出する．D の算出は学習データ数 n と最大次数 $N-1$ に比例するが，これは提案法と同等の計算量といえる．

以上から，提案法のメモリ量および計算量は従来法とほぼ同等であるといえる．

3.4　事前知識の利用法

3.2 節の説明のとおり，提案法はモデルの事前確率とパラメータ $\boldsymbol{\theta}_m$ の事前分布を仮定している．$\boldsymbol{\theta}_m$ の事前分布は，接尾木の各節点，すなわち各々の $\boldsymbol{x}^m$ に対応するパラメータであり，類似データにより学習した結果を，モデル構築時に事前知識として利用することが容易である．

これを行う利点としては，類似データと学習データの双方で出現する単語列 $\boldsymbol{x}^m$ に対して，その次に生成された単語の類似データの情報や，予測を行う $\boldsymbol{x}_p^m$ が類似データにのみ存在する場合の情報を利用できることにある．

具体的には，式 (12) のディリクレ分布のパラメータは事前に観測した単語の出現頻度に相当する [12], [13] ことから，類似データをもとに接尾木を構築し観測された単語の出現頻度を，ディリクレ分布のパラメータとして利用できる．次に，式 (18) をもとに算出した類似データのモデルの事後確率を，学習データのモデルの事前確率として利用できる．

より詳細には，類似データの予測対象となる単語の出現頻度を学習データの事前分布であるディリクレ分布のパラメータに加算し，類似データにて構築済みの接尾木に対して学習データをもとに接尾木および各節点の頻度を更新することで，アルゴリズムを修正することなく行うことができる．

なお，類似データは学習データよりも大量に入手可能であることが想定される．そのため，各節点が保持する出現頻度をそのまま学習データ用の事前分布であるディリクレ分布のパラメータに利用すると，学習データで観測された値が反映されなくなる可能性が高い．この場合は，類似データの各節点で保持された出現頻度の和をある値 $\rho > 0$ に制限し，類似データの影響を弱めるよう調整することで対応可能である．具体的には，類似データで出現した履歴となる単語列 $\boldsymbol{x}^{N-1}$ のそれぞれに対して，式 (12) に含まれる $\sum_{y\in\boldsymbol{Y}} c(y|\boldsymbol{x}^{N-1}) + \sum_{y\in\boldsymbol{Y}} \alpha(y|\boldsymbol{x}^{N-1})$ が ρ となるように，$c(y|\boldsymbol{x}^{N-1}) + \alpha(y|\boldsymbol{x}^{N-1})$ を調整する．これにより，ρ を小さくするに従い事前知識として利用する類似データの影響は小さくなる．

4.　文書データによる単語予測実験

本章では，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，提案法が実用的にも有効であることを示す．

また，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類似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 [14], [15]．提案法は，3.4 節で示したとおり，式 (17) 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 [16]，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をそれぞれ用いてモデルを構築し，検証データによる単語予測の実験を行う．これにより，事前知識として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[*5]．

日本語の文書として，特許文書とシステム開発文書を利用した．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などを含む 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．

[*5] 文書データに対する単語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．

© 2013 Information Processing Society of Japan

表 1　各方式による単語予測結果の正答率（単位 %）
Table 1　Percentages of correct 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

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 2,553 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，従来法として 2 章で説明した，混合分布を仮定したモデル，ニーザー・ネイ法を用いた．提案は式 (17) に基づく予測法を用いた．それぞれの方式に対して，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った．

混合分布のモデル構築は，学習データを 2 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 [4] 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 2 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，式 (1) 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 [6] を用いた．この方式は式 (2) に含まれる D の算出に対して，式 (3) 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 回から 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 の事前確率と，パラメータ $\boldsymbol{\theta}_m$ の事前分布とするディリクレ分布のパラメータを，無情報事前分布となるよう以下のとおり設定した．まず，m の事前確率は，データ圧縮の分野で広く使われている方式で，m の値が大きくなるに従い値が小さくなるように 2^{-m} で与えた．ただし $m = N$ の場合は 2^{N-1} とした[*6]．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する動詞に対して，4.1 節で示した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．

利用場面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[*7]．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 1 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 2 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，予測の候補として確率が最上位の単語を 1 つ出力した場合，上位 5 位は，予測の候補として確率が大きい単語を 5 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 ポイント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正答率を持つといえ，理論的な最適性に加え実用的にも有効といえる．

4.3　事前知識の利用の効果

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の効果を検証する．具体的には，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でモデルを構築し，単語予測の実験を行う．これは，事前知識を利用しない場合と，利用する場合にそれぞれ相当する．

事前知識とする類似データは特許文書のすべてのデータを用いた．学習に用いる少量データはシステム開発文書の

[*6] モデルの事前確率の影響は小さく等確率で与えても傾向に大きな違いはなかった．

[*7] 複数候補を提示する場合でも，損失関数の期待値のとり方を修正することでアルゴリズムの導出が可能である．

© 2013 Information Processing Society of Japan

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 similar and train data with just train one in $N = 3$.

学習データ件数	最上位				上位 5 位を出力			
	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
232	22.47	23.70	有為	有為	38.31	39.63	有為	有為
464	26.86	27.75	有為	有為	44.99	45.45	有為	無為
928	34.31	34.15	無為	無為	52.29	52.09	無為	無為
1,856	37.31	37.14	無為	無為	57.53	57.33	無為	無為
3,712	41.37	41.45	無為	無為	62.02	62.08	無為	無為
7,424	45.05	45.09	無為	無為	67.37	67.35	無為	無為
14,848	48.10	48.14	無為	無為	72.15	72.12	無為	無為

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データに対して，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 3.4 節で説明した ρ で調整したものを，学習データの無情報事前分布に加算する形式で利用した．なお，類似データにしか存在しない単語の場合はそのままの値を利用することになる．本実験では，履歴となる単語列が類似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し，類似データは学習データに含まれなかった $\boldsymbol{x}^m$ に対する予測の効果を狙い $\rho = 0.01$ とした[*8]．

$N = 3$ 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 2 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 [17] 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，最上位による単語予測では学習データが少量の 232 件の場合は 1.23 ポイント，増加させた 14,848 件の場合は 0.04 ポイント，上位 5 件を出力する単語予測では学習データが少量の 232 件の場合は 1.32 ポイント，増加させた 14,848 件の場合は 0.03 ポイントと，学習データ量が増えるに従い小さくなる傾向がみられた．このことから，データが少量の場合は事前知識の影響を受け，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持つといえる．

なお，提案法は漸近的な一致性を持つことが示されている [18]．これは，本実験で使用した 2 種類の事前分布の設定方法に対して学習データの増加に従い，双方で構築したモデルが一致することを意味する．本実験により示された性質は，文献 [18] の解析結果を支持した結果となっている．

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，真の次数が未知の統計問題ととらえ，ベイズ決定理論に基づいて，単語の予測誤りの損失の最小化に対しベイズ基準のもとで最適となる予測法を導出した．これにより，モデルの予測分布に対してモデルの事後確率で重みづけして足し合わせた，従来研究と形式的に類似した方式が，ベイズ基準のもとで最適な予測法であることを示した．

文書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与え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持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応や業務適応を想定した場合，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，データの追加によるモデルの再構築の可能性が示唆される．

[*8] なお，本実験で ρ を大きな値にすると，全般的に正答率が低下した．これは，類似データと学習データの双方で出現した $\boldsymbol{x}^m$ に対して，類似データで出現した単語を出力した結果に誤りが多く含まれ，類似データのみで出現した $\boldsymbol{x}^m$ に対する単語の正答数を上回ったためであった．

© 2013 Information Processing Society of Japan

また，本提案法は，考えうるモデルに対してモデルの予測分布をモデルの事後確率で重み付けを行うという特徴を持つ．このことから，階層 N グラムモデルに限定したものではなく，言語モデルの分野で提案されている単語列だけでなく単語の品詞も利用したモデル，類似した単語列を同一種類の系列と見なしてまとめあげるクラス・モデル [4], [14], [15] といった，様々な確率的なモデルに対して適用できると考えられる．

今後の課題としては，階層 N グラムモデルに限定しない様々なモデルの導入による，単語予測の正答率向上の取り組みがあげられる．また，学習データが少量な場合の事前知識の導入においては，類似データの選定方法や事前知識の影響を適切に調整する方法があげられる．

また，今回の実験で行った動詞の予測だけでなく，複合名詞で構成される専門用語の予測など，入力支援を行う範囲の拡張によりユーザの利便性をより向上させることも課題の 1 つである．

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] 海野裕也，坪井祐太：頻出文脈に基づく分野依存入力支援，言語処理学会第 17 回年次大会，言語処理学会，pp.1107–1110 (2011).
[3] 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).
[4] 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).
[5] Kneser, R. and Ney, H.: Improved backing-off for m-gram language modeling, *Proc. ICASSP*, Vol.1, pp.181–184, Association for Computational Linguistics Morristown, NJ, USA (1995).
[6] Chen, S.F. and Goodman, J.: An Empirical Study of Smoothing Techniques for Language Modeling, *Proc. ACL*, pp.310–318 (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] 須子統太，鈴木　誠，浮田善文，小林　学，後藤正幸：確率統計学，オーム社 (2010).
[9] Teh, Y.W.: A Bayesian Interpretation of Interpolated Kneser-Ney, Technical report, NUS School of Computing Technical Report (2006).
[10] Teh, Y.: A Hierarchical Bayesian Language Model based on Pitman-Yor Processes, *Proc. COLING/ACL 2006*, pp.985–992 (2006).
[11] 持橋大地，隅田英一郎：階層 Pitman-Yor 過程に基づく可変長 n-gram 言語モデル，情報処理学会論文誌，Vol.48, No.12, pp.4023–4032 (2007).
[12] Bernardo, J.M. and Smith, A.F.M.: *Bayesian Theory*, Wiley (2000).
[13] 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).
[14] Jelinek, F., Mercer, R.L. and Rouks, S.: *Principles of Lexical Language Modeling for Speech Recognition*, Dekker Publishers, New York (1991).
[15] 森　信介：自然言語処理における分野適応，人工知能学会誌，Vol.27, No.4, pp.365–372 (2012).
[16] 繁桝算男：ベイズ統計入門，東京大学出版会 (1985).
[17] 永田　靖：入門統計解析法，日科技連 (1992).
[18] 後藤正幸：ベイズ統計理論に基づく確率モデルの推定と予測の漸近的評価に関する研究，博士論文，早稲田大学大学院理工学研究科 (2000).

末永 高志　（正会員）

平成 9 年早稲田大学理工学部経営システム工学科卒業．平成 11 年同大学大学院理工学研究科修士課程修了．同年株式会社 NTT データ入社．パターン認識，データ分析技術，自然言語処理技術の実用化研究に従事．人工知能学会，電子情報通信学会，言語処理学会各会員．

松嶋 敏泰　（正会員）

昭和 53 年早稲田大学理工学部工業経営学科卒業．昭和 55 年同大学大学院修士課程修了．同年日本電気（株）入社．昭和 61 年早稲田大学大学院理工学研究科博士後期課程入学．平成元年横浜商科大学講師．平成 3 年同大学助教授．平成 4 早稲田大学理工学部工業経営学科助教授．平成 9 年同大学教授．平成 19 年早稲田大学基幹理工学部応用数学科教授，現在に至る．知識情報処理および情報理論とその応用に関する研究に従事．博士（工学）．平成 13 年ハワイ大客員研究員．平成 23 年カリフォルニア州立大学バークレイ校客員教授．IEEE，電子情報通信学会，人工知能学会，OR 学会，日本経営工学会等各会員．

© 2013 Information Processing Society of Japan