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

末永 高志　　松嶋 敏泰
株式会社NTTデータ 技術開発本部　早稲田大学 基幹理工学部 応用数理学科
suenagatk@nttdata.co.jp, toshi@mgmt.waseda.ac.jp

1　はじめに

システム開発分野では，競争力強化のため一部業務の海外発注が増加している．海外発注する業務はプログラムの製造が主であるが，近年，日本語による設計書作成の依頼も期待されるようになってきた．しかしながら，日本語が母語でない開発者の作成する日本語文書には日本語の誤りが多く含まれ，文書の確認，修正に多くの時間がとられているのが現実である．

日本語非母語者の作成する日本語文には特に助詞の用法誤りの多さが指摘されている [1]．この支援を想定して，助詞の用法理解にフォーカスした用例検索支援方式 [3]，助詞の用法をチェックする誤用判定方式 [1] が提案されている．これらは日本語の文法の理解や訂正を支援することが目的であるが，本稿ではより適切な単語の選定を支援するための入力支援方式を検討する．具体的には，名詞句や助詞が与えられたもとでその後に続く動詞などの単語をNグラムモデルをもとに予測し，単語の候補を提示することを考える．

Nグラムモデルの構築の際は，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となるため，平滑化の処理が行われる．平滑化処理は，極低頻度に出現するデータの影響を制御するディスカウント処理と，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処理の二種類に分けられる [2].

本稿では，上記の階層的な補完処理に着目し，履歴となる単語列とその次に続く単語の対により構成される学習データをもとに，モデルを構築するための方法を検討する．具体的には，上記に示した学習問題をベイズ決定理論にもとづく定式化 [4] をもとに考察し，ここで検討される基準に対して最適となる予測法を導出する．最後に，日本語による設計書データを用いて入力支援技術への応用を検討する．

2節ではNグラムモデルとそれを用いた予測法を説明する．3節ではベイズ決定理論にもとづく最適予測法を導出する．4節では3節で導出した予測法の日本語入力支援への応用を検討する．5節は考察およびまとめである．

2　Nグラムモデルによる単語予測

Nグラムモデルによる単語予測では，名詞句や助詞で構成された履歴となる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が大きい高次なモデル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，パラメータの数と比較すると得られるデータは一般に疎，すなわち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なることが課題である．

これに対して，Nグラムモデルが階層モデル族 [5] 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 [2]．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．

3　ベイズ決定理論にもとづく最適予測法

本節では，学習用の n 個のデータの対 $\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n$，$y^n \in \boldsymbol{Y}^n$ と，履歴となる単語系列 $\boldsymbol{x}_p^{N-1}$

が与えられた条件で，ベイズ決定理論にもとづく最適な予測法を検討する．

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，

$$p(y|\boldsymbol{x}^{N-1}) = p(y|\boldsymbol{x}^{N-1}, \boldsymbol{\theta}_{m^*}) \tag{2}$$

という確率で単語が生成されると仮定する．また，どのような意志決定を行うかを整理すると，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 $\{\boldsymbol{x}^{N-1}\}^n$，$y^n$ 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．これを決定関数として定式化すると，

$$\hat{y} = D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n) \tag{3}$$

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考察する手順は以下の通りである．まず，(3) 式で示される決定関数をもとに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もののため，学習データに対して (2) 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で階層化が不要と仮定した場合の N グラムモデルで議論を行い，次にモデルの次数が未知の場合の階層 N グラムモデルを対象とする．

3.1 N グラムモデル

まず，予測した結果の正誤判定に対して距離

$$d(\hat{y}, y_p) = \begin{cases} 0 & (\hat{y} = y_p) \\ 1 & (\hat{y} \neq y_p) \end{cases} \tag{4}$$

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]，

$$\begin{aligned} & L(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p) \\ &= \sum_{y_p \in \boldsymbol{Y}_p} d(D, y_p) p(y_p|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) \end{aligned} \tag{5}$$

[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，

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

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，

$$\begin{aligned} & B_{risk}(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p) \\ &= \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\theta}} R(D, \boldsymbol{Y}_p|\boldsymbol{\theta}, \boldsymbol{\mu}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} f(\boldsymbol{\mu}) d\boldsymbol{\mu} \\ &= \sum_{\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n} \sum_{y^n \in \boldsymbol{Y}^n} \sum_{y_p \in \boldsymbol{Y}_p} \\ & \quad \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\theta}} d(D, y_p) p(y_p|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) \\ & \quad p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}) p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) \\ & \quad f(\boldsymbol{\theta}) d\boldsymbol{\theta} f(\boldsymbol{\mu}) d\boldsymbol{\mu} \\ &= \sum_{\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n} \sum_{y^n \in \boldsymbol{Y}^n} \sum_{y_p \in \boldsymbol{Y}_p} \\ & \quad \int_{\boldsymbol{\mu}} \Big\{ 1 - \int_{\boldsymbol{\theta}} I_D(y_p) p(y_p|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) \\ & \quad f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \Big\} p(\{\boldsymbol{x}^{N-1}\}^n, y^n) \\ & \quad p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) f(\boldsymbol{\mu}) d\boldsymbol{\mu} \end{aligned} \tag{7}$$

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，(7) 式に含まれる

$$\int_{\boldsymbol{\theta}} p(y_p|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \tag{8}$$

を最大化することで得られる．すなわち，

$$\begin{aligned} & \hat{y} = \arg\max_y \\ & \int_{\boldsymbol{\theta}} p(y|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}) f(\boldsymbol{\theta}|\{\boldsymbol{x}^{N-1}\}^n, y^n) d\boldsymbol{\theta} \end{aligned} \tag{9}$$

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．

3.2 階層 N グラムモデル

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く．なお，距離関数は (4) 式を仮定する．

まず，階層 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，

$$\begin{aligned} & L_h(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p|m) \\ &= \sum_{y_p \in \boldsymbol{Y}_p} d(D, y_p) p(y_p|\boldsymbol{x}_p^{m-1}, \boldsymbol{\theta}_m, m) \end{aligned} \tag{10}$$

となる．

この損失関数に対する危険関数は，

$$\begin{aligned} & R_h(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p|m, \boldsymbol{\theta}_m, \boldsymbol{\mu}) \\ &= \sum_{\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n} \sum_{y^n \in \boldsymbol{Y}^n} L_h(D, \boldsymbol{Y}_p|m) \\ & \quad p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}_m, m) p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) \end{aligned} \tag{11}$$

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

$$\begin{aligned} & B_{h,risk}(D(\boldsymbol{x}_p^{N-1}, \{\boldsymbol{x}^{N-1}\}^n, y^n), \boldsymbol{Y}_p) \\ &= \int_{\boldsymbol{\mu}} \sum_{m \in \{N\}} p(m) \int_{\boldsymbol{\theta}_m} \\ & \quad R_h(D, \boldsymbol{Y}_p|m, \boldsymbol{\theta}_m, \boldsymbol{\mu}) d\boldsymbol{\theta}_m d\boldsymbol{\mu} \\ &= \sum_{\{\boldsymbol{x}^{N-1}\}^n \in \{\boldsymbol{X}^{N-1}\}^n} \sum_{y^n \in \boldsymbol{Y}^n} \sum_{y_p \in \boldsymbol{Y}_p} \int_{\boldsymbol{\mu}} \sum_{m \in \{N\}} \\ & \quad p(m) \int_{\boldsymbol{\theta}_m} d(D, y_p) p(y_p|\boldsymbol{x}_p^{m-1}, \boldsymbol{\theta}_m, m) \\ & \quad p(y^n|\{\boldsymbol{x}^{N-1}\}^n, \boldsymbol{\theta}_m, m) f(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\ & \quad p(\{\boldsymbol{x}^{N-1}\}^n|\boldsymbol{\mu}) f(\boldsymbol{\mu}) d\boldsymbol{\mu} \end{aligned} \tag{12}$$

となる．ベイズ危険関数の最小値は (12) 式の，

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

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

$$\begin{aligned} \hat{y} &= \arg\max_y \sum_{m \in \{N\}} \int_{\boldsymbol{\theta}_m} p(y|\boldsymbol{x}_p^{N-1}, \boldsymbol{\theta}_m, m) \\ & \quad f(\boldsymbol{\theta}_m|\{\boldsymbol{x}^{N-1}\}^n, y^n, m) d\boldsymbol{\theta}_m \\ & \quad p(m|\{\boldsymbol{x}^{N-1}\}^n, y^n) \end{aligned} \tag{14}$$

となる $\hat{y}$ を出力することになる．これは，N グラムの低次のモデルごとに予測分布 [5] を求め，それらをモデルの事後確率で重み付けることで予測を行うことを意味している．

4 日本語入力支援への応用

本節ではシステム開発文書を対象に，提案する予測法の日本語入力支援への応用を検討する．具体的には，階層化しない N グラムモデルと階層 N グラムモデルで予測の精度比較を行い，提案する予測法が適切であることを確認する．

精度の比較方法は，履歴となる単語系列と予測する単語の対となるデータをもとに学習データを用いて予測分布およびモデルの事後確率を推定し，その推定結果と検証データの履歴となる単語系列をもとに予測した単語と，検証データのその次に続く単語との正答率で行う．

4.1 検証の条件

対象とするデータは文書から名詞，助詞，動詞と連続する単語系列を抽出し，さらに履歴のデータとしてこの単語系列よりも前に出現する単語列を品詞を区別することなく抽出した．また，助詞を含めそれより前の系列を履歴となる単語系列，予測対象とする単語を動詞とした[2]．このデータから無作為に学習データと検証データに分割しそれぞれ 59,621 件，59,634 件を利用した．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 1,989 種類であった．モデル m の事前確率は 2^{-m}，ただし $m = N$ の場合は 2^{N-1} で与えた[3]．

比較する N グラムモデルの N は 3 から 6 まで行い，3.1 節で示した階層化しない N グラムモデルと，3.2 節で示した階層 N グラムモデルの双方で，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．また，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[4]．検証にあたっては予測に使う単語の確率が最上位であった単語と検証データの正解となる単語を比較した正答率と，確率が大きいものから上位 5 件の単

[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．

[3] ただし，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．

[4] 複数候補を提示する場合でも，損失関数の期待値の取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．

表 1: 予測結果の正答率

N	最上位		上位 5 位を出力	
	なし	あり	なし	あり
3	0.519	0.522	0.775	0.787
4	0.626	0.643	0.793	0.837
5	0.688	0.731	0.773	0.861
6	0.702	0.756	0.759	0.865

語を抽出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の評価を行った．

4.2 検証結果

学習データをもとに構築したモデルを用いて，検証データの予測を行い正答率を算出した結果を表 1 に示す．まず，予測に使う単語が最上位のものの場合を比較すると，二つの方式ともに N の増加とともに正答率が向上しているが，N が 3 と 6 の時の正答率の向上度合いを確認すると，階層化しない場合は 0.183，階層化する場合は 0.234 と正答率の向上する度合いに差が見られた．また，上位 5 位まで出力した場合は，階層化しない場合は N が 4 の場合に正答率が最大値をとりその後，低下する傾向が見られたが，階層化する場合は N が増加するに従い精度が上昇している．このことから，高次の N グラムモデルを利用するだけでなく，提案した階層 N グラムモデルの予測法を適用することで，複数候補を提示する場合にも正答率の向上に寄与することが示された．

次に，表 2 に階層 N グラムモデルの N が 6 の場合の各次数におけるモデルの事後確率を示す．このように，モデルの次数が高くなるに従いモデルの事後確率が高く予測に対する寄与が大きくなることを示している．しかしながら，階層化しない場合のように履歴情報の到達した最高次のモデルによる予測を行うのではなく，提案したモデルの事後確率による重み付けを行うことで正答率の向上が見られたと考えられる．

5 おわりに

本稿では，N グラムモデルを用いた単語の予測に対して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を提案し，実データによる予測の正答率を検証した．本提案法は，考え得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

表 2: 階層 N グラムにおける $N = 6$ での各次数のモデルの事後確率

モデルの次数 m	事後確率
2	0.008
3	0.110
4	0.209
5	0.320
6	0.353

という特徴をもつ．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，その他の言語モデルの分野で提案されている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．

今後の課題は，階層 N グラムモデルに限定しない様々なモデルに提案法の効果を検証し，適用可能な範囲を明らかにすることである．

参考文献

[1] 大木環美, 大山浩美, 北内啓, 末永高志, 松本裕治. 非日本語母国話者の作成するシステム開発文書を対象とした助詞の誤用判定. 言語処理学会第 17 回年次大会. 言語処理学会, 2011.

[2] 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.

[3] 加藤宏紀, 葛原和也, 加藤芳秀, 松原茂樹. Escort-jacs: 外国人による日本語文作成を支援する用例文検索システム. 信学技法, pp. 13–18, 2010.

[4] 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–. 情報論的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.

[5] 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.