博士学位論文

統計的決定理論に基づく
単語予測問題の理論的解析

Theoretical Analysis of Term Prediction Problems Based on Statistical Decision Theory

2014年2月

末永 高志

Takashi SUENAGA

博士学位論文

統計的決定理論に基づく
単語予測問題の理論的解析

Theoretical Analysis of Term Prediction Problems Based on Statistical Decision Theory

2014年2月

早稲田大学大学院 基幹理工学研究科
数学応用数理専攻 情報理論研究

末永 高志
Takashi SUENAGA

目 次

表 目 次

図 目 次

第1章　序論

1.1　はじめに

数理統計学研究の一つのアプローチとして統計的決定理論を基礎とした研究 [3, 42, 59, 4, 13, 63] がある．この研究では，データ解析 [63] を始め，情報理論における歪みのない情報源圧縮 [23]，誤り訂正符号 [32]，文書分類 [37, 38]，推薦システム [45] といった様々な分野での適用例が報告されている [63].

一方，自然言語処理の分野においては，言語的な特性を規則化したルールを用いた方法のみならず，統計的言語処理と呼ばれる大量の言語データを学習データとして用いた数理的な手法をもとに処理法を求めるアプローチが適用され，実データに対する様々な効果が知られている [22, 34, 64]．これは，ルールを用いた方法の課題であった解釈の可能性が複数存在する場合において，これを確率的な事象ととらえることで優位性を発揮してきたといえる．

統計的言語処理においては，言語モデルと呼ばれる $N-1$ 次マルコフモデルである N グラムモデルを基盤とした数理モデルを用いることで発展してきた．そして，N グラムモデルを基盤とし，これに自然言語の特性を数理モデルとして導入する様々な研究が行われ，実証的な観点による有効性が示されてきた．しかしながら，学習データをもとに数理モデルに設定されたパラメータを推定する方法等において，理論的な意味付けが明らかでないまま提案されてきたものも少なくない．

このように，統計的決定理論と統計的言語処理の両者に共通して，学習データを利用したデータ解析を行うことになるが，統計的言語処理の研究において，統計的決定理論のアプローチを採用した例はほとんど見受けられない．

本研究では，統計的言語処理の分野における代表的な問題である単語予測問題とその分野適応問題に統計的決定理論のアプローチを適用し，この問題の理論的な解析を行うことを目的とする．具体的には，

(1) 予測誤り率に関してベイズ基準のもとで最適な単語予測法を導出

(2) 導出される予測法の特性の考察

(3) N グラムモデルを基盤とした統計的言語処理の分野の従来法の，上記方法からの解釈

(4) 実データによる評価実験

を行う．

まず，(1) においては，統計的決定理論のアプローチにより予測誤り率を損失関数として設定し，評価基準をベイズ基準としたもとで最適となる予測法を導出する．次に，(2) においては，統計的言語処理の一般的アプローチで求められる方法は，数値計算の繰り返しやアドホックなパラメータチューニングを学習データを用いて行われることが多い一方，本研究において導出される予測法は解析解として与えられることを示す．さらに，(3) においては，統計的決定理論のアプローチにより導出された予測法をもとに，従来法との形式を比較することで，従来法に含まれるパラメータの算出方法に解釈を与える．最後に，(4) の実データによる実験により基盤的な方法の中で優位性が報告されている従来法と，ほぼ同等の予測精度が得られることを示す．

従来の統計的言語処理の一般的アプローチでは，実証的に有効な結果が多く報告されていることに対して，本研究の統計的決定理論のアプローチの適用による議論を行うことで，上記の (1)，(2)，(3) の単語予測法の性質に対する理論的な考察が初めて可能となった．

なお，本研究では統計的言語処理の問題に統計的決定理論のアプローチを適用することを特徴とする．そのため，本論文においては統計的決定理論の用語を用いて記述を行う．しかしながら，統計的決定理論および統計的言語処理では類似の用語が用いられ，同じ表記で別の意味で用いられている場合もあり混乱を招きやすい．なるべく，その都度説明を加え，誤解のないよう記述を試みる．

また，このような新しいアプローチを用いる場合，情報理論 [26, 9] などでは，まずは独立同分布 [54] であったりマルコフ過程 [55, 46] といった汎用的で簡易な確率モデルを仮定したもとで，理論的最適性や計算量，メモリー量，その他の性質など

を議論するのが一般的である．今回もそれに従い，最も汎用的で統計的言語処理にて基盤となる N グラムモデルを確率モデルとして採用し議論する．

まず，統計的決定理論のアプローチを採用することで理論的な最適性の議論が可能となる背景について，統計的言語処理の一般的アプローチとの違いを整理することから始める．ここで，本研究で対象とする単語予測問題を定義すると，予測の条件となる $N-1$ 個の系列データである $x_i^{N-1} \in X^{N-1}$ と，予測対象となる $y_i \in Y$ のデータについて，事例として与えられた n 個の学習データの対 $(x^{N-1}, y)^n$ と，ある観測データ x_p^{N-1} が与えられたもとで，未知の y_p を出力する決定関数 $D(x_p^{N-1}, (x^{N-1}, y)^n)$ を導出する問題となる．なお，この前提は単語予測問題に限定するものではなく，例えば，回帰分析であれば $X \in R$，$Y \in R$ であるし，単語予測や離散マルコフ過程の予測問題であれば，$|X| < \infty$，$|Y| < \infty$ となる．

統計的決定理論のアプローチと統計的言語処理の一般的アプローチの違いを整理すると，

(i) 確率モデルと決定関数の区別の有無

(ii) 最適の意味の違い

(iii) 設定する評価基準と理論的最適性の保証の違い

があげられる．

まず，(i) の確率モデルと決定関数の区別の有無であるが，統計的決定理論ではパラメトリックな確率モデル $P(y|x^{N-1}, \boldsymbol{\theta})$ を仮定し，これとは別に，予測を行うための決定関数 $\hat{y} = D(x_p^{N-1}, (x^{N-1}, y)^n)$ を定義する [1]．上記の確率モデルは，単語予測を例にすると，長さ $N-1$ の単語列 x^{N-1} の次に単語 y が発生する確率として，そのパラメータが $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ でパラメタライズされていることを意味する．また，決定関数は n 個の学習データの対 $(x^{N-1}, y)^n$ と単語列 x_p^{N-1} を入力に，y_p の予測のために $\hat{y}$ を出力する関数である．

予測といった意思決定の場面においてはデータの発生する確率モデルのクラスを仮定しても，確率モデルに対する完全な知識を有することは前提にできない．そのため，確率モデルと決定関数を区別し，確率モデルに対する既知なパラメータと未

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

知のパラメータ，および予測に利用できるデータを明確にし，ある基準から最適な決定関数 D を求めることを統計的決定理論では主要な問題としている．

一方，統計的言語処理の一般的アプローチでは，$P(y|x^{N-1}, \boldsymbol{\theta})$ に相当するものを言語モデルと表現するが，確率モデルではなく決定関数のように取り扱われており，確率モデルにおける未知のパラメータや，予測における利用可能なデータについて明確に整理して議論されているとはいえない．

次の (ii) の最適性の意味の違いであるが，これは，$P(y|x^{N-1}, \boldsymbol{\theta})$ における $\boldsymbol{\theta}$ が既知としたもとで，例えば，

$$D_1(x^{N-1}, \boldsymbol{\theta}) = \max_y P(y|x_p^{N-1}, \boldsymbol{\theta}) \tag{1.1}$$

という予測法が最適な方法であると論じられている [11, 33, 5] [2]．なお，この議論の方法は，統計的言語処理の一般的アプローチと類似するパターン認識・機械学習の分野にて特徴的に見受けられる．

しかしながら，実際は確率モデルのパラメータ $\boldsymbol{\theta}$ は未知であり，これをいかに決定するかが最重要な問題であるにもかかわらず，その観点なしに最適性の主張がされている．すなわち，$\boldsymbol{\theta}$ をどのように決定するかが論じられていなければ，予測誤り対する理論的な保証は何もないと言える．

予測法の導出のための具体的な方法については，確率モデルの未知なパラメータの決定法については明言せず，式 (1.1) の予測法に対して，その想定する決定関数のクラスを式 (1.1) に含まれる $P(y|x_p^{N-1}, \boldsymbol{\theta})$ に限定し，$\boldsymbol{\theta}$ を操作することで決定関数を決める研究が多く報告されている [1, 18, 16, 33, 43, 57]．すなわち，言語モデルとは決定関数の形式を設定したものととらえられる．

一方，統計的決定理論においては，次の (iii) で述べる方法で学習データの分布を陽に扱い，決定関数 $D(x_p^{N-1}, (x^{N-1}, y)^n)$ の最適化を行うことで，統計的言語処理の一般的アプローチと異なり，決定関数のクラスを限定せずに最適性の議論を行っている．

[2] 式 (1.1) のような予測法を x_p^{N-1} が与えられた元で，y の事後確率を最大にするということで，ベイズ決定と呼ばれる場合がある．これは，統計的決定理論を含むベイズ統計学におけるベイズ基準 [60] による最適な決定関数とはまったく異なっていることに注意が必要である．このあたり混乱のないよう注意されたい．

(iii) の設定する評価基準と理論的最適性の保証の有無であるが，アプローチの違いにより説明できる．まず，損失に対しては，例えば次のような予測誤り率をもとに，損失関数[3]

$$L(D(x_p^{N-1},(x^{N-1},y)^n),Y|\boldsymbol{\theta}) = \sum_{y_p \in Y} d(D,y_p)p(y_p|x_p^{N-1},\boldsymbol{\theta}) \tag{1.2}$$

を設定する．ただし，$d(D,y_p)$ は決定関数 $D(x_p^{N-1},(x^{N-1},y)^n)$ の出力と，予測対象の真の値 y_p のとの距離を表し，例えば，一致すれば 0，一致しなければ 1 を出力する $0-1$ 損失などが考えられる．統計的決定理論では，これを学習データに対して期待値をとった危険関数

$$\begin{aligned} &R(D(x_p^{N-1},(x^{N-1},y)^n),Y|\boldsymbol{\theta},\boldsymbol{\mu}) = \\ &\sum_{(x^{N-1})^n \in (X^{N-1})^n} \sum_{y^n \in Y^n} L(D,Y|\boldsymbol{\theta})P(y|x^{N-1},\boldsymbol{\theta})P(x^{N-1}|\boldsymbol{\mu}) \end{aligned} \tag{1.3}$$

を設定する．ただし，$\boldsymbol{\mu}$ は x^{N-1} の確率パラメータである．式 (1.3) の最適化を行うにあたっては様々な方法がとられるが，例えば以下の，パラメータ $\boldsymbol{\theta}$，$\boldsymbol{\mu}$ の事前分布にて期待値をとったベイズ危険関数

$$B_{\text{risk}}(D(x_p^{N-1},(x^{N-1},y)^n),Y) = \int_{\boldsymbol{\mu}}\int_{\boldsymbol{\theta}} R(D,Y|\boldsymbol{\theta},\boldsymbol{\mu})f(\boldsymbol{\theta})d\boldsymbol{\theta}f(\boldsymbol{\mu})d\boldsymbol{\mu} \tag{1.4}$$

に基づいて，これを最小とする決定関数 $D(x_p^{N-1},(x^{N-1},y)^n)$ を導出する．このように，未知のパラメータ $\boldsymbol{\theta}$ の近似的な操作を行わないことで，決定関数に対する最適性の理論的保証を行っている．

一方，統計的言語処理の一般的アプローチでは，式 (1.1) に示したとおり，確率モデルに含まれる未知であるはずのパラメータ $\boldsymbol{\theta}$ が残った形式にて決定関数の型が決まっていた．そのため，ほとんどの場合において学習データ $(x^{N-1},y)^n$ を用いて $\boldsymbol{\theta}$ をチューニングする近似的な方式を採用している．これは，実用性の観点から広く利用されているが，理論的な最適性の観点からは大数の法則で保証されるものの，有限のデータを用いる場合に最適であるかは明らかでない．

[3]統計的言語処理の一般的アプローチと類似するパターン認識・機械学習の分野では，式 (1.2) に含まれる $d(D,y_p)$ を損失関数やコスト関数，式 (1.2) を期待損失と呼ぶこともある [16, 57].

これまでは，統計的決定理論のアプローチと統計的言語処理の一般的アプローチの議論の方法の違いを見てきた．このような違いのある統計的決定理論のアプローチを採用することで，統計的言語処理の一般的アプローチで生じる問題が自然に解決されることを述べる．

なお，ここで想定する統計的言語処理の一般的アプローチで生じる問題として，

(a) ゼロ頻度問題

(b) N グラムの次数決定問題

を取り上げる．

(a) のゼロ頻度問題とは，統計的言語処理の一般的アプローチに基づき $\boldsymbol{\theta}$ を決定する場合，学習データに現れた単語の出現頻度が前提となるため，学習データには含まれないが存在することはわかっているデータの出現頻度を 0 と扱わざるをえなくなる．このようなデータについては，式 (1.1) の形式では予測できない．そのため，スムージングと呼ばれるアドホックなパラメータを用意し，学習データから推定することが行われている [21, 7, 34]．

一方，統計的決定理論のアプローチでは，ベイズ危険関数の導出の際に $\boldsymbol{\theta}$ の事前分布を導入することにより，学習データに含まれないデータであっても確率を与えることができる．そのため，ベイズ危険関数を最小にする決定関数を求めることにより自然とその問題は解決可能である．なお，統計的言語処理の一般的アプローチにおける，スムージングと呼ばれるパラメータにより未知の単語の出現頻度の調整を行うことは，事前分布の設定を行っている方法と解釈することができる．

(b) の N グラムの次数決定問題であるが，N グラムモデルは $N-1$ 次のマルコフモデルであり，高次のモデルが低次のモデルを含む入れ子型の階層モデル族 [56] なため，高次のモデルの方がより広いクラスの確率モデルを表現可能である．ただし，高次のモデルは低次のモデルよりもパラメータの数が多く，同数のサンプルから各パラメータの推定に用いられるサンプル数が相対的に少なくなるために，推定誤差は大きくなる傾向にあるといえる．

この問題に対して，統計的言語処理の分野では様々なアプローチがとられている．例えば，補完法と呼ばれる方法では，各次数の N グラムを重み付け加算した予測法が提案されている [14, 34, 17] ．この方法においては，EM アルゴリズム [10] を

用いて隠れ変数を推定する方法 [14] をはじめ，最大事後確率推定をもとに各次数の重みを推定する方法 [41, 49]，隠れ変数である各モデルの重みをマルコフ連鎖モンテカルロ法 [36] などを利用する，事後確率のサンプリング的な手法で推定する方法 [30, 29, 53] など様々な研究が報告されている．

他には，学習データから推定するパラメータの数を削減することで，モデルに含まれるパラメータの推定誤差を低減させることを狙いとした方法が提案されている．例えば，単語列 x^{N-1} に対して，表層的な単語ではなく，個々の単語が対応付く仮想的な単語クラスを設定したクラス N グラムモデル [6, 22, 34, 62] の有効性が広く知られている．同様に，想定する最大次数のすべてのパラメータを推定するのではなく，単語列によってはこの N の数を変動させる可変長 N グラムモデル [25, 34] や，さらに，これらを組み合わせたクラスに基づく可変長 N グラムモデル [47] が提案されている．

このように，統計的言語処理の一般的アプローチに基づく研究では，言語モデルと呼ばれる決定関数に相当するものに様々な形式を設定し，その決定関数に含まれるパラメータを学習データをもとにチューニングを行う方法が一般的である．統計的言語処理の研究においては，自然言語の特性に合わせた様々な形式を，問題にあわせて様々に取り込んだ数理モデルを提案し，実証的な評価を行うことで有効な結果が得られている．

一方，統計的決定理論のアプローチを適用した場合，N グラムの次数決定問題は，どの次数の確率モデルを用いればよいかという問題となり，モデル選択 [59, 31, 56] や，次数が未知のマルコフモデルに対するベイズ統計の枠組み [23, 59] 等で解決可能である．

なお，統計的言語処理の一般的アプローチにおいては，言語モデルと呼ばれる決定関数に相当するものを，パープレキシティ [34] と呼ばれる評価指標で評価している．これは，実データをもとに確率モデルの適否を決定するアドホックな方法と解釈することができ，予測誤り率を最小とする決定関数を求めることには直接的につながらない．一方，統計的決定理論のアプローチにおいては，仮定した確率モデルに対して，予測誤差に関する評価基準を最適化する決定関数が導出される．

統計的決定理論のアプローチにおいては，確率モデルと決定関数を分けて議論することから，統計的言語処理の従来研究で培われた様々な言語モデルを，確率モデ

ルとして利用することが可能である．しかしながら，本研究では統計的決定理論と統計的言語処理の一般的アプローチとの違いを議論することと，統計的決定理論のアプローチの特徴を明らかにすることが主要な目的であり，最も広く利用され簡便であり，統計的言語処理の基盤となる単語 N グラムモデル（以下，N グラムモデルと呼ぶ）に固定して議論をすすめる．この理由としては，本研究の主とする目的が単語予測における予測誤り率の改善ではなく，統計的決定理論のアプローチを単語予測問題に適用することで，従来のアプローチでは容易ではなかった理論的な最適性や，その他の優れた特性の議論が可能となることを示すことにあるからである．

1.2　本論文の構成

第 2 章では，統計的決定理論のアプローチを説明し本研究の位置づけを明確にする．ここでは，統計的決定理論のアプローチにおける議論の展開方法を詳述する．次に，統計的決定理論のアプローチの観点から統計的言語処理の一般的アプローチにおける最適性の議論との違いを整理する．さらに，統計的言語処理の分野にて研究されてきた，言語モデルと呼ばれる決定関数の従来研究を整理し，これらが，統計的言語処理の一般的アプローチに基づいた研究であることを示す．これらをもとに，本研究で採用する統計的決定理論のアプローチと統計的言語処理の一般的アプローチとの違いを明確にする．

次の，第 3 章，第 4 章では，提案する統計的決定理論のアプローチを，統計的言語処理の代表的な問題である単語予測問題と，その分野適応問題の二つの問題に適用し有効性を示す．

まず，第 3 章では，単語予測問題に対して統計的決定理論のアプローチを適用する．具体的には，N グラムモデルを確率モデルに仮定し，その次数および確率パラメータの双方を未知とする．さらに，決定関数を定義し，予測誤り率を損失関数と設定し，その評価基準のもとで最適となる予測法を導出する．

次に，導出された予測法の特性を整理する．具体的には，単語の発生する分布に条件付き多項分布を仮定した場合，予測法の最適解が解析解として求まることを示す．ここで，導出された予測法を実装したアルゴリズムについて，計算量，メモリー

使用量が統計的言語処理の分野での従来法と同等であることと，データの追加に対し加算増分により更新できる計算量的な優位点があることを示す．

あわせて，導出された予測法と従来法の形式を比較することで，統計的言語処理の分野にて従来法として知られていた予測法に対して，理論的な解釈が与えられることを示す．

最後に，導出された予測法が従来法とほぼ同等の予測誤り率となることを実験により示す．

第 4 章では，単語予測における分野適応問題 [15, 58, 48] に対して，統計的決定理論のアプローチを適用する．分野適応問題では，例えば，システム開発分野における文書作成支援を想定した場合，当該業務のデータ（以下，当該データと呼ぶ）は少量にしか得られないことが多く，類似する分野のデータとして，公開データである特許文書や，科学技術論文など，大量に入手可能なデータ（以下，類似データと呼ぶ）を活用することで，単語の予測精度を向上させることを考える．ここで，当該データと類似データは母集団が異なるため，利用される単語が類似することは期待されるものの同一の確率分布であるとは限らない．そのため，類似データを当該データにあわせるための，適応処理をいかに行うかが問題となる．

これに対して，当該データと類似データの類似性について，条件付き多項分布のパラメータが確率的に変化する数理モデルを用いた確率モデルを仮定する．そのもとで，統計的決定理論のアプローチを適用することで，予測誤り率を損失関数として設定し，この評価基準のもとで最適となる予測法を導出する．なお，ここで仮定する変化モデルは，多項分布のパラメータ変化をランダムウォークとして表現した Simple Power Steady Model (SPSM)[27] と呼ばれる確率モデルに対応している．

さらに，実現されるアルゴリズムの特性を整理し，統計的言語処理の分野にて従来法として知られていた予測法に対して，理論的な解釈が与えられることを示す．

また，導出する予測法は漸近的な一致性を持つことが示されている [50]．これは，類似データの事後分布のパラメータが確率的に変化した分布を，当該データの事前分布として利用する場合と利用しない場合において，当該データの増加に従い双方で求められる予測分布が一致することを意味する．そこで，当該データの増加に従い類似データの影響が小さくなる特徴を有すことを実験にて確認する．

最後に，第 5 章において以上の成果をまとめ，今後の展望を述べる．

第2章　本研究の位置づけ

2.1　はじめに

本章では，統計的決定理論のアプローチと統計的言語処理の一般的アプローチの対比をもとに，本研究の位置づけを明確にする．最初に本研究にて着目する統計的決定理論のアプローチを整理し，次に，統計的言語処理の一般的アプローチについて，統計的決定理論のアプローチとの対比を行い，違いを整理する．さらに，統計的決定理論のアプローチを採用することの効果を述べ，従来研究として各種検討されてきた具体的な課題が，統計的決定理論のアプローチを採用することで自然に解決されることを述べる．また，統計的言語処理の研究にて言語モデルと呼ばれる関数を用いた[1] 代表的な従来研究を整理し，提案する統計的決定理論のアプローチと，アプローチ法に違いがあることを明確にする．

なお，本研究においては，統計的言語処理の代表的な問題である単語予測問題を対象とする．本研究にて想定する単語予測問題を定義すると，予測の条件となる $x_i^{N-1} \in X^{N-1}$ と，予測対象となる $y_i \in Y$ のデータについて，事例として与えられた n 個の学習データの対 $(x^{N-1}, y)^n$ とある観測データ x_p^{N-1} が与えられたもとで，未知の y_p を出力する決定関数 $D(x_p^{N-1}, (x^{N-1}, y)^n)$ を導出する問題となる．ここで，n 個の学習データは独立同分布で発生し，学習データと予測の条件および対象となるデータは同一の分布から発生するものとする．また，単語予測問題の場合は単語の数を有限とし $|Y| < \infty$ とする．

[1]第 1 章で記載したとおり，これは統計的決定理論における決定関数に相当し，確率モデルを指しているとは限らないことに注意.

2.2　統計的決定理論のアプローチ

統計的決定理論に基づき理論的に最適な予測法を導出するアプローチの特徴を，

(1) 確率モデルの仮定

(2) 決定関数の定義

(3) 損失関数の設定

(4) 危険関数の設定

(5) ベイズ危険関数によるベイズ基準の設定

(6) 最適な決定関数の導出

と整理し，それぞれの具体的な内容を説明する．

(1) 確率モデルの仮定

データが発生する分布について，一般的には以下のパラメトリックな確率モデル，

$$P(x^{N-1}, y|\boldsymbol{\nu}) = P(y|x^{N-1}, \boldsymbol{\theta})P(x^{N-1}|\boldsymbol{\mu}) \tag{2.1}$$

を仮定する[2]．ここで，$\boldsymbol{\nu}$ は x^{N-1} と y の同時分布のパラメータ，$\boldsymbol{\theta}$ は x^{N-1} が与えられたもとでの y の出現する分布のパラメータ，$\boldsymbol{\mu}$ は x^{N-1} の出現する分布のパラメータである．

統計的決定理論においては，確率モデルの一部が未知であると仮定する．例えば，上記の $\boldsymbol{\theta}$ や $\boldsymbol{\mu}$ であったり，N グラムモデルのようなマルコフモデルにおいては，y がどの次数から出現しているかを確定することは困難であったりするため，真の次数の値を未知とすることが行われる．

[2]本研究においては，パラメータ $\boldsymbol{\mu}$ と $\boldsymbol{\theta}$ が独立であることを仮定している．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いることが，文献 [60] にて指摘されている．

(2) 決定関数の定義

データを観測したもとでの意思決定を行うための決定関数を

$$\hat{y} = D(x_p^{N-1}, (x^{N-1}, y)^n) \tag{2.2}$$

と定義する[3]．これは，n 個の学習データの対 $(x^{N-1}, y)^n$ と単語列 x_p^{N-1} を入力に，y_p の予測のために $\hat{y}$ を出力する関数である．

予測といった意思決定の場面においては，データの発生する確率モデルを仮定するものの，それに含まれるパラメータのいずれかは未知であり確率モデルに対する完全な知識を有することは前提にできない．そのため，データの発生する確率モデルと意思決定のための決定関数を明確に区別している．統計的決定理論では，予測に利用できるデータを明確にし，ある基準から最適な決定関数 D を求めることを主要な問題と考えている．

(3) 損失関数の設定

予測の損失を考慮するにあたり，予測した結果の正誤判定の距離を定義する．例えば，

$$d(\hat{y}, y_p) = \begin{cases} 0 & (\hat{y} = y_p) \\ 1 & (\hat{y} \neq y_p) \end{cases} \tag{2.3}$$

と定義される $0-1$ 損失などが考えられる．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．

このように定義された距離に対して，$y_p \in Y$ は確率変数であるため，真の分布 $\boldsymbol{\theta}$ で期待値をとることで，損失関数を，

$$L(D(x_p^{N-1}, (x^{N-1}, y)^n), Y|\boldsymbol{\theta}) = \sum_{y_p \in Y} d(D, y_p) p(y_p|x_p^{N-1}, \boldsymbol{\theta}) \tag{2.4}$$

と設定する．ここで，式 (2.3) で定義された距離関数を利用する場合は予測誤り率を意味する．しかしながら，上記の (2) で説明した通り学習データ $(x^{N-1}, y)^n$ は母集団から確率的に発生したものであり，そのままでは最適化ができない．そのため，次に示す危険関数を設定する．

[3]以後，決定関数 $D(x_p^{N-1}, (x^{N-1}, y)^n)$ と明らかな場合は D と省略する．

(4) 危険関数の設定

危険関数は，損失関数 $L(D(x_p^{N-1},(x^{N-1},y)^n),Y|\boldsymbol{\theta})$ に対して $(x^{N-1},y)^n$ の学習データによる期待値であり，

$$R(D(x_p^{N-1},(x^{N-1},y)^n),Y|\boldsymbol{\theta},\boldsymbol{\mu}) = \sum_{(x^{N-1})^n\in(X^{N-1})^n}\sum_{y^n\in Y^n} L(D,Y|\boldsymbol{\theta})P(y|x^{N-1},\boldsymbol{\theta})P(x^{N-1}|\boldsymbol{\mu}) \tag{2.5}$$

となる．ここで，危険関数には未知のパラメータ $\boldsymbol{\theta}$ と $\boldsymbol{\mu}$ が含まれるため，そのままでは最適化ができない．そこで，ベイズ統計学における一つの方法として，次に示すようにそれぞれのパラメータの事前分布を仮定し，その事前分布に対する平均化を行ったベイズ危険関数を，最適化を行うための評価基準とする方法がある．

(5) ベイズ危険関数によるベイズ基準の設定

ベイズ危険関数は，上記の危険関数 $R(D,Y|\boldsymbol{\theta},\boldsymbol{\mu})$ に対して，パラメータ $\boldsymbol{\theta}$ と $\boldsymbol{\mu}$ の事前分布 $f(\boldsymbol{\theta})$ と $f(\boldsymbol{\mu})$ により平均化したものであり，

$$B_{\mathrm{risk}}(D(x_p^{N-1},(x^{N-1},y)^n),Y) = \int_{\boldsymbol{\mu}}\int_{\boldsymbol{\theta}} R(D,Y|\boldsymbol{\theta},\boldsymbol{\mu})f(\boldsymbol{\theta})d\boldsymbol{\theta}f(\boldsymbol{\mu})d\boldsymbol{\mu} \tag{2.6}$$

となる．これにより，未知のパラメータが消去された形式となり，これに対して最適化を行う．

(6) 最適な決定関数の導出

(5) で設定したベイズ危険関数を最小とする基準をベイズ基準とよび，その基準に対して

$$D_{\mathrm{br}}(x_p^{N-1},(x^{N-1},y)^n) = \arg\max_{D} B_{\mathrm{risk}}(D,Y) \tag{2.7}$$

となる最適な決定関数を求める．

2.3　統計的言語処理の一般的アプローチ

自然言語処理の分野にて単語が確率的に発生することを前提とした統計的言語処理 [34] の一般的アプローチについて，上述した統計的決定理論のアプローチの観点と対比することで基本的な考えの違いを整理する．なお，具体的な研究としては，分類誤りを考慮した研究 [1, 18, 16, 43, 33, 57] や，決定関数に言語モデルと呼ばれる自然言語の制約を数理モデルで表現し，与えられた学習データを用いて数理モデルに含まれるパラメータをチューニングする方法の研究 [14, 41, 49, 30, 29, 53] があげられる．

ここでは，前節の (1) から (6) で整理した観点との差異を整理する．なお，統計的言語処理の一般的アプローチは，パターン認識・学習理論のアプローチに類似しているところがあり，適宜，補足としてこの分野についても触れることがある．

(1) 確率モデルの仮定と (2) 決定関数の定義

データが発生する確率と決定関数は明確に区別されず，まず，$\boldsymbol{\theta}$ を既知とした場合の予測法として

$$D_1(x_p^{N-1}, \boldsymbol{\theta}) = \arg\max_y P(y|x_p^{N-1}, \boldsymbol{\theta}) \tag{2.8}$$

と形式を定める．次に，本来未知である $\boldsymbol{\theta}$ を，n 個の学習データの対 $(x^{N-1}, y)^n$ 等を利用してチューニングを行い予測法を求める．最も単純な方法は学習データにおける単語の出現頻度を利用した $\hat{\theta}_{y|x_p^{N-1}} = c(y|x_p^{N-1})$ とする方法である．ただし，$c(y|x_p^{N-1})$ は学習データに含まれる x_p^{N-1} に後続する y の出現頻度である．このように，$\boldsymbol{\theta}$ をどのようにチューニングするかは予測法の性能を左右する重要な問題となる．

なお，パターン認識・機械学習の分野では，決定関数という表現は用いないが，例えば，線形関数を仮定した場合は，

$$\begin{aligned} \hat{y} &= f(x_p^{N-1}, \boldsymbol{\theta}') \\ &= \theta_0 + \sum_{i=1}^{N-1} \theta_i' x_i \end{aligned} \tag{2.9}$$

の形式にて予測を行う，ただし，$\boldsymbol{\theta}' = \{\theta'_1, \theta'_2, \cdots, \theta'_{N-1}\}$ である．なおこの形式は，回帰分析においては回帰式に対応し [55]，パターン認識の分野では式 (2.9) を識別関数 [33]，学習理論の分野では学習モデルや学習機械とも呼ばれる．

この $\boldsymbol{\theta}'$ は，統計的決定理論における決定関数のチューニングパラメータと考えることができる．注意しなければならない点として，統計的決定理論で仮定した確率モデルのパラメータ $\boldsymbol{\theta}$ とは異なるものを指していることである．このように，決定関数に相当するものは考慮されているものの，一般的に未知である確率モデルの直接的な取り扱いは意識されていない．

(3) 損失関数の設定，(4) 危険関数の設定，(5) ベイズ危険関数によるベイズ基準の設定

予測や分類に対する損失として，

$$l(y, \hat{y}) = I(y \neq D_1(x_p^{N-1}, \boldsymbol{\theta})) \tag{2.10}$$

で表現される $0-1$ 損失を定義した例で説明する．ただし，$I(\cdot)$ は $\cdot$ が真の場合 1 を，偽の場合 0 を返す関数である．統計的言語処理の一般的アプローチにおいては，

$$L(Y, X^{N-1}, \boldsymbol{\theta}) = \sum_{x_p^{N-1} \in X^{N-1}} \sum_{y \in Y} l(y, D(x_p^{N-1}, \boldsymbol{\theta})) P(y, x_p^{N-1} | \boldsymbol{\theta}) \tag{2.11}$$

と表現される関数が損失として設定される．

なお，パターン認識・機械学習の分野においては，式 (2.10) のことを損失関数もしくはコスト関数，式 (2.11) のことを期待損失と呼ばれたりする [5, 57]．

統計的決定理論のアプローチと対比すると，式 (2.11) が式 (2.4) で示した損失関数に類似しているととらえることができる．また，損失関数を学習データで期待値をとった危険関数や，確率モデルの未知なパラメータへの対応を考慮したベイズ危険関数に相当するものはない．

(6) 最適な決定関数の導出

式 (2.11) を最小化する決定規則として，確率パラメータ $\boldsymbol{\theta}$ が既知の場合においては，

$$D_1(x_p^{N-1}, \boldsymbol{\theta}) = \arg\max_y P(y|x_p^{N-1}, \boldsymbol{\theta}) \tag{2.12}$$

とすることが最適な予測法となる．これは，パターン認識・機械学習の分野の研究にて広く知られている考え方である [11, 33, 5].

この考え方は，統計的決定理論のアプローチで採用されるパラメトリックな確率モデルを仮定し，その未知のパラメータが何であるかを明確にする議論とは大きな違いがある．また，この決定規則では，決定関数の形式が $P(y|x_p^{N-1}, \boldsymbol{\theta})$ で与えられるものと決めておき，以下の手続きでパラメータのチューニングを行っている．

具体的には，$P(y|x_p^{N-1}, \boldsymbol{\theta})$ に対する完全な情報を保持していることは一般にはないため，その代替として，学習データ $(x^{N-1}, y)^n$ を用いた経験損失

$$L_{\mathrm{emp}}(D_1(x_p^{N-1}, \boldsymbol{\theta}), (x^{N-1}, y)^n) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, D(x_i^{N-1}, \boldsymbol{\theta})) \tag{2.13}$$

で近似した基準が設定される [1, 18, 16, 33, 43, 5, 57].

このように，統計的言語処理の一般的アプローチでは，学習データを用いて $\boldsymbol{\theta}$ をチューニングして，経験損失を最小化したものを決定関数として利用しているにすぎない．そのため，学習データが有限であることの考慮はなく，有限のデータに対する理論的な性質は明らかでない．

2.4　統計的決定理論のアプローチの特徴

これまで，統計的決定理論のアプローチをもとに，統計的言語処理の一般的アプローチとの違いを見てきた．本節では，これらの違いをもとに本研究で採用する統計的決定理論の特徴をもとに，従来研究にて課題とされ様々な対策が行われていたことが，統計的決定理論のアプローチを採用することで自然に解決されることを示す．

まず，統計的言語処理の一般的アプローチの課題として，

(a) ゼロ頻度問題

(b) N グラムの次数決定問題

を取り上げる．

(a) ゼロ頻度問題とは，学習データに含まれないが，存在することはわかっているデータをいかに取り扱うかの問題である．統計的決定理論のアプローチでは，事前分布を仮定した議論を行っており，ゼロ頻度問題においては，事前分布に対する頻度の調整を行う問題と解釈することができる．これにより，より体系的な観点による対応が可能となる．

一方，統計的言語処理の一般的アプローチでは，加算スムージング [8, 34] と呼ばれる出現頻度を調整するパラメータの設定方法が個別に検討されている．

(b)N グラムの次数決定問題とは N グラムモデルが高次のモデルが低次のモデルを含むモデルとなっていることに起因する問題である．N グラムモデルは $N-1$ 次のマルコフモデルであり，高次のモデルが低次のモデルを含む入れ子型の階層モデル族 [56] なため，高次のモデルの方がより広いクラスの確率モデルを表現可能である．ただし，高次のモデルは低次のモデルよりもパラメータの数が多く，同数のサンプルから各パラメータの推定に用いられるサンプル数が相対的に少なくなるために，推定誤差は大きくなる傾向にあるといえる．

統計的決定理論の分野においては，この仮定した確率モデルに対して真のモデルがただ一つ存在するとし，その次数を未知の場合において，与えられた学習データに対して，どの確率モデルを利用するかという問題になる．その観点においては，統計的モデル選択問題 [31, 56] や次数が未知のマルコフモデルにおけるベイズ統計の枠組み [23, 59, 60, 50] にて解決法が提案されている．これは，情報理論における歪みのない情報源圧縮の成果 [23] をもとに，誤り訂正符号 [32]，分類 [37, 38]，推薦システム [45] といった様々な分野での適用例が報告され，本研究の課題である単語予測問題においても有効なアプローチであると考えられる．

一方，統計的言語処理のアプローチでは，上記のような問題設定ではなく，次節にて紹介する補完法と呼ばれる複数の次数のモデルを足しあわせた形式による対応や，高次モデルのパラメータを学習データに合わせて低減させる方法が検討される．

2.5　単語予測問題の従来研究

統計的言語処理研究の基礎となる N グラムモデルは，$N-1$ 個の単語で構成された単語列に依存してその次に続く単語が確率的に生成することを前提としている．これは，N の値が大きくなるに従いパラメータの数が増加し，同数のサンプルの場合はパラメータに対するサンプル数が相対的に少なくなる．自然言語のデータを対象とした場合，単語の種類が非常に多いため結果的に単語列として観測されないゼロ頻度のサンプルが多く存在することとなる．

この対策として，統計的言語処理の研究では，複数の次数のモデルを足しあわせた形式の決定関数を用意することで，単語を予測することが行われている [22, 34, 17]．統計的言語処理の分野ではこの重み付けし足しあわせる方法を補完法と呼ぶことがある．

具体的には，履歴となる単語列 $x^{N-1} \in X^{N-1}$ に対して，ある次数 m（ただし $m < N$）から次に続く単語 $y \in Y$ が発生する確率の推定値 $P(y|x_p^{N-1}, \boldsymbol{\theta}_m)$ を $w(m)$ で重みづけた予測式

$$D_{\text{inter}}(x_p^{N-1}, \boldsymbol{\theta}) = \sum_{m \in \{N\}} P(y|x_p^{N-1}, \boldsymbol{\theta}_m) w(m) \tag{2.14}$$

が提案されている．この $\boldsymbol{\theta}_m$ や $w(m)$ の算出においては，EM アルゴリズムを適用した方法 [14] や，最大事後確率推定をもとに各次数の重みを推定する方法 [41, 49] など様々な方法が提案されている．

また，上記の重み付けの考え方を踏襲しつつ，学習データに含まれる単語の出現頻度を調整するパラメータを設定した方法が提案されている [19, 7]．例えば，

$$\begin{aligned} D_{\text{kn}}(x^m, \boldsymbol{\theta}) \;\; = \;\; & \frac{\max\{c(y|x^m) - d_{\text{discount}}, 0\}}{\sum_{y \in Y} c(y|x^m)} \\ & + \frac{d_{\text{discount}}}{\sum_{y \in Y} c(y|x^m)} \big|\{y : c(y|x^m) > 0\}\big| D_{\text{kn}}(x^{m-1}, \boldsymbol{\theta}) \end{aligned} \tag{2.15}$$

といった形式が提案されている．ただし，d_{discount} は $d_{\text{discount}} \geq 0$ とする割引係数，$c(y|x^m)$ は学習データに含まれる x^m に後続する y の出現頻度，$|\cdot|$ は集合の要素数，すなわち，x^m の後に出現した y の種類を意味する．

この他に，学習データから推定するパラメータの数を削減することで，モデルに含まれるパラメータの推定誤差を低減する試みが行われている．例えば，マルコフ連鎖を行う単語列の条件に対して，表層的な単語ではなく仮想的な単語クラスを設定したクラス N グラムモデル [6, 22, 34, 62] が提案されている．これは，単語列 $x_p^{N-1} = x_{p,1}x_{p,2}\cdots x_{p,N-1}$ ではなく，個々の単語 $x_{p,i}$ があるクラス $c_{p,i}$ に属すようにクラスタリングの手法をもとにクラスタ分けし，クラスの係列 $c_p^{N-1} = c_{p,1}c_{p,2}\cdots c_{p,N-1}$ を用いて $P_{\rm class}(y|c_p^{N-1})$ と予測する方法である．これは，多くの場合式 (2.14) にて示した補完法と組み合わせて使われる [47].

同様に，N グラムモデルのパラメータの数を調整するアプローチとして，想定する最大次数のすべてのパラメータを推定するのではなく，単語列によってこの N の値を変動させる可変長 N グラムモデル [25, 34] や，さらに，これらを組み合わせたクラスに基づく可変長 N グラムモデル [47] が提案されている．

また，N グラムモデルは直前の系列 x^{N-1} のみに依存することを仮定したモデルであるが，この依存する情報に x^{N-1} 以外のものを追加する方法が提案されている．これは，自然言語の一般的な傾向として，記載される話題によって利用される単語の出現分布の傾向が異なることに着目している．具体的には，ある単語 y は単語列 x^{N-1} に含まれる M 個の単語列 x_{N-M}^{N-1} により強く依存して出現すると想定し，その直近の情報 x_{N-M}^{N-1} を短期記憶するとなぞらえた，キャッシュモデルと呼ばれる方法が提案されている [24]．これは，短期記憶用の $P_c(y|x_{N-M}^{N-1}, \boldsymbol{\theta}_c)$ を用意し，$P(y|x^{N-1}, \boldsymbol{\theta})$ に重み付け加算を行う予測式が提案されている．他には，特定の文書においてある単語と単語の共起関係を考慮した，トリガーモデルと呼ばれる方法が提案されている [20]．具体的には，キャッシュモデルと同様に，系列 x^{N-1} に含まれる M 個の単語列 x_{N-M}^{N-1} に対して後続する単語 y の共起関係を考慮した指標として

$$P_{\rm T}(y|x_{N-M}^{N-1}) = \frac{1}{M}\sum_{m=1}^{M} t(y, x_{n-m}) \tag{2.16}$$

を用意し, $P(y|x^{N-1}, \boldsymbol{\theta})$ に重み付け加算が行われた予測式が提案されている．ただし，$t(y, x_{n-m})$ は単語 x_{n-m} と y の共起の強さを意味し，学習データをもとに算出されるものである．この値は，例えば，相互情報量 [9] といった指標を利用して求められる．

ここに示した従来研究の評価においては，多くの場合，パープレキシティと呼ばれる指標の中で，特に評価用データ $(x_i^{N-1}, y_i), i = 1, 2, \cdots, n_t$ を用いて算出する，テストセット・パープレキシティ

$$P_{\text{perplexity}} = \prod_{i=1}^{n_t} D_{\text{nl}}(x^{N-1}, \boldsymbol{\theta})^{-\frac{1}{n_t}} \tag{2.17}$$

が利用されている [34]．ただし，$D_{\text{nl}}(x^{N-1}, \boldsymbol{\theta})$ は，上記の様々な形式にて提案されている言語モデルと呼ばれる予測式に対応する．これは後続する単語の平均分岐数を表し，評価データに対する当てはまりの良さを表す指標である．しかしながら，この評価方法は，実データをもとに確率モデルの適否を決定するアドホックな方法と解釈することができ，予測誤り率を最小とする決定関数を求めることには直接的につながらない

このように，単語予測の問題に対して多数の方法論のもとで，様々な決定関数の形式が提案されている．その形式に対して，設定されたパラメータを学習データをもとにチューニングを行うアプローチがとられている．これらは，問題にあわせて適切に組み合わせることで，実証的な観点での有効性が報告されている [35, 64]．

2.6　統計的決定理論のアプローチと統計的言語処理の一般的アプローチとの違い

本研究では統計的決定理論のアプローチを採用するが，統計的言語処理の従来研究で採用されるアプローチとの違いは以下のように整理される．

統計的言語処理の従来研究においては，様々な方法論をもとに自然言語の特性を考慮した言語モデルと呼ばれる決定関数の形式が提案されている．このアプローチ法は，確率モデルと決定関数を明確に区別していない．一方で，提案する統計的決定理論のアプローチでは，確率モデルを仮定し，未知であるパラメータを明確にする．その次に決定関数を定義する．そのもとである損失関数を設定し，危険関数，ベイズ危険関数といった評価基準に対して対して最適な決定関数を求めることを主要な問題としている．

統計的言語処理の一般的アプローチとの明らかな違いは，決定関数の形式を最初に決めておくのではなく，上記の手続きをもとに導出することである．これは，2.2 節に示した通り，決定関数に $D((x^{N-1}, y)^n, x_p^{N-1})$ と関数に含まれる変数の宣言のみで定義していることから明らかである．この点が提案するアプローチとの違いの一つの重要なポイントである [4] ．

2.7　統計的決定理論のアプローチを適用する具体的な問題

本研究にて採用する統計的決定理論のアプローチを，次の二つの問題に適用する．まず，第 3 章においては，問題 1 とした単語予測問題に統計的決定理論のアプローチを適用する．ここでは，自然言語の単語が確率的に発生することを前提とした単語の予測問題に対して，予測誤り率を損失関数とした基準を設定し，その基準に関して最適な単語予測法を導出しその性質を考察する．次に，第 4 章においては，問題 2 とした単語予測における分野適応問題に統計的決定理論のアプローチを適用する．ここでは，予測の対象とする当該分野のデータ（以下，当該データ）が少量しか得られない場合，類似した分野のデータ（以下，類似データ）を用意しそれを組み合わせて利用する問題である [5] ．これにより本アプローチの拡張性を確認する．

問題 1 の議論の内容であるが，確率モデルに N グラムモデルを仮定し，単語予測問題に統計的決定理論のアプローチを適用することで，単語予測の誤り率を理論的に最小とする単語予測法を導出する．ここで仮定する N グラムモデルは単語の出現する分布を表すパラメータと，依存する単語の発生次数のパラメータを含む．ここでの議論においては，最初に簡単のためモデルの次数は既知で確率分布のパラメータは未知とした条件で始め，次に双方が未知とした条件で議論する．また，導出された方法を従来研究と比較することで，従来研究で提案された方法に理論的な解釈を与える．

[4] これは，本研究においては，N グラムモデルを確率モデルと仮定しているが，決定関数として N グラムモデルを利用することにならないことに注意．

[5] 機械学習の分野では転移学習 [58] と呼ばれることもある．

問題 2 の議論の内容であるが，単語予測における分野適応の問題を対象とする．分野適応の問題とは，当該データが少量しか得られない場合において，類似データを追加することで，より精度の高い予測を実現することを狙ったものである．統計的決定理論の観点で見ると，ここに示す類似データは事前知識と解釈することができる．しかしながら，類似データと当該データは異なる母集団から発生したものである．そのため，類似データと当該データの類似性を，条件付き多項分布のパラメータが変化する数理モデルで表現し，それを確率モデルとして仮定したもとで，統計的決定理論のアプローチに基づいて単語予測法を導出する．なお，ここで仮定する変化モデルは，多項分布のパラメータ変化をランダムウォークとして表現した Simple Power Steady Model (SPSM)[27] と呼ばれる確率モデルを利用する．また，導出された予測法を統計的言語処理の分野で従来法として知られている方法と比較することで，従来法に理論的な解釈を与える．

第3章　問題1：単語予測問題に対する統計的決定理論アプローチの適用

3.1　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 [44]，業務文書の表現の統一 [66]，オフショア開発といった日本語非母語者向け入力支援 [40] などの文書の作成支援において，入力の省力化や表記ゆれの低減といった要求に対して，入力済みの単語列をもとに後続する単語を予測し候補として提示することが求められている．この課題に対して，単語が確率的に発生することを仮定した確率モデルを用いた単語の予測方法を考える．これは，単語の生成という不確かに発生する事象に対して，後続する単語を予測する決定のための方法を議論することが目的である．その方法論として統計的決定理論 [3, 42, 59, 4, 13, 63] に基づくアプローチを採用する．

3.2　従来研究

本節では，統計的言語処理の分野にて言語モデルと呼ばれる形式で提案されてきた，数理モデルの基盤となる N グラムモデルと，N グラムモデルの各次数のモデルを重み付けし加算する補完法と呼ばれる予測式の形式を説明する．さらに，この形式に対して，その重みパラメータを隠れ変数とみなし算出する方法と，この補完法に，学習データから算出される単語の出現頻度の値を調整する，アドホックなパラメータを追加した形式の決定関数を利用するニーザー・ナイ法，および，基盤となる N グラムモデルを自然言語の特性を考慮し拡張した数理モデルを利用する方法を

説明する．

N グラムモデル

本研究で対象とする N グラムモデルは，$N-1$ 個の単語で構成された単語列 $x^{N-1} = x_1 x_2 \cdots x_{N-1}$ に依存してその次に続く単語 y が確率的に生成すると仮定したモデルである．経験的には N がある程度の大きさであることが単語予測の精度に寄与すると報告されている [26, 51]．しかしながら，N の値が大きくなるに従いパラメータの数が指数的に増加するため，パラメータの数と比較すると得られるデータは相対的に少なくなる．自然言語のデータを対象とした場合，単語の種類が非常に多いため結果的にゼロ頻度となる単語列が数多く存在することになる．

これに対して，統計的言語処理の分野においては，このような性質をもつ N グラムモデルに対して低次のモデルを高次のモデルに加算する，補完法と呼ばれる方法が適用されてきた [22, 34, 17]．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 $x^{N-1} \in X^{N-1}$ が与えられたときに，次に続く単語 $y \in Y$ の確率として，N グラムモデルの低次のモデル m とパラメータ $\boldsymbol{\theta}_m$，さらにモデルの重み $w(m)$ とし，低次のモデルから高次のモデルまで重みで加算する予測式

$$D_{\mathrm{inter}}(x^{N-1}, \boldsymbol{\theta}) = \sum_{m=1}^{N} p(y|x^{m-1}, \boldsymbol{\theta}_m) w(m) \tag{3.1}$$

が提案されている．ただし，$m \leq N$，$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_m,)$，$\sum_{m=1}^{N} w(m) = 1$ である．

隠れ変数を仮定した方法

式 (3.1) に含まれるモデル m の重み $w(m)$ について，学習データが N グラムのいずれの次数で発生したかを未観測の隠れ変数ととらえ，その欠損値を $z \in Z$ としたもとで，学習データからその重みを推定することが行われている．なお，これは各次数のモデルが混合した分布から単語が生起すると仮定していると解釈することもできる．この重みについて，例えば，EM アルゴリズム [10, 5] を用いた算出法が広く知られている [14, 22, 34, 17]．この方法は学習データに対する尤度関数の最大化

の近似であり，以下の E ステップと呼ばれる時点 t でのパラメータの推定値 $\hat{\boldsymbol{\theta}}_t$ のもとで，各次数の重みの期待値を計算する

$$Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}_t) = E_{Z|(x^{N-1},y)^n,\boldsymbol{\theta}}[\log L_{\mathrm{ml}}(\hat{\boldsymbol{\theta}}; (x^{N-1}, y)^n, Z)] \tag{3.2}$$

の処理と，M ステップと呼ばれるその次数の重みの期待値のもとでパラメータを最大化する

$$\hat{\boldsymbol{\theta}}_{t+1} = \arg\max_{\hat{\boldsymbol{\theta}}} Q(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{\theta}}_t) \tag{3.3}$$

の処理による繰り返し計算により求められる．ただし，$L_{\mathrm{ml}}(\hat{\boldsymbol{\theta}}; (x^{N-1}, y)^n, Z)$ は隠れ変数 z の尤度関数である．

なお，EM アルゴリズムでは，一般的に尤度関数の凸性が成立するとは限らないため，大域的最適解が保証されない [5]．また，N グラムモデルのようなマルコフモデルは入れ子型階層モデルであるため，最尤法により近似を求めると，多くの場合，最高次のモデルの対数尤度が最大となってしまう [56]．そのため，それぞれのステップ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，パラメータの推定精度を向上させるために様々なデータの分割方法も提案されている．また，最大事後確率推定をもとに各次数の重みを推定する方法 [41, 49] など様々な方法が提案されている．

ニーザー・ナイ法

この方法は，前節で示した補完法の形式に，学習データから算出される単語の出現頻度の値を調整する，アドホックなパラメータを追加した形式の決定関数を利用することが特徴である．具体的な形式としては，予測式

$$\begin{aligned} D_{\mathrm{kn}}(x_p^m, \boldsymbol{\theta}) &= \frac{\max\{c(y|x_p^m) - d_{\mathrm{discount}}, 0\}}{\sum_{y\in Y} c(y|x_p^m)} \\ &\quad + \frac{d_{\mathrm{discount}}}{\sum_{y\in Y} c(y|x_p^m)} \big|\{y : c(y|x^m) > 0\}\big| D_{\mathrm{kn}}(x_p^{m-1}, \boldsymbol{\theta}) \end{aligned} \tag{3.4}$$

が提案されている．ただし，d_{discount} は $d_{\mathrm{discount}} \geq 0$ とする割引係数，$c(y|x^m)$ は学習データに含まれる x^m に後続する y の出現頻度，$|\cdot|$ は集合の要素数を表し，x^m の後に出現した y の種類を意味する．

d_{discount} はいくつかの算出法が検討されているが，長さ m の単語列に対して，$c_{m,1}$ 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，$c_{m,2}$ 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，

$$d_{\text{discount},m} = \frac{c_{m,1}}{c_{m,1} + 2c_{m,2}} \tag{3.5}$$

のように算出する方法が提案されている [19, 7]．しかしながら，これらの算出法等の理論的な意味は明らかでない．

近年，この形式は最大事後確率推定の方法を発展させたピットマン–ヨー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 [28, 29, 53]．しかしながら，単語予測の評価にあたっては，評価データに対するテストセット・パープレキシティの指標による実験的な検証が中心で，単語予測の誤り率に対する理論的な解析とはなっていない．

その他の研究

本研究で利用する N グラムモデルは，単語列を前提としたもので単語 N グラムモデルとも呼ばれる．統計的言語処理の研究においては，単語 N グラムモデルを発展させ，自然言語の特性を数理モデルとして表現する様々な方法が提案されている．

例えば，それぞれの単語をより上位概念のクラスに対応付けるクラス N グラムモデル [6] や，単語列ごとに利用する N の値を調整する可変長 N グラムモデル [25] が提案されている．

N グラムモデルは高次のモデルが低次のモデルを含む入れ子型階層モデル族 [56] なため，高次のモデルの方がより広いクラスの確率モデルを表現可能である．ただし，高次のモデルは低次のモデルよりもパラメータの数が多く，同数のサンプルの場合，各パラメータの推定に用いられるサンプル数が相対的に少なくなる．そのため，パラメータの推定誤差は大きくなる傾向にあるといえる．これらの方法は，学習データをもとに推定するパラメータの数を低減させることで，上記の推定誤差を抑えることを狙ったものといえる．

また，N グラムモデルで前提とする局所的な依存関係に対して，より広い範囲の依存関係を考慮するために，直近に出現する単語列に含まれる一部の単語列の影響

をより強めるキャッシュモデル [24] や，単語の共起関係を事前に求めそれを予測の際に加算するトリガーモデル [20] などが提案されている．

このように，自然言語の特性を言語モデルと呼ばれる決定関数に取り入れることで，実証的な観点による有効性が報告されている．

従来研究のアプローチ

上記のとおり確率モデルと決定関数を明確に分けて議論されていない．ここに示した方法は，それぞれ単語予測を行うための決定関数の形式を最初に決め，学習データから何かしらの方法で，その決定関数に含まれるパラメータをチューニングすることが行われている．これらの点が，統計的決定理論のアプローチとの大きな違いである．

3.3　統計的決定理論のアプローチの適用

本節では，統計的決定理論のアプローチを単語予測問題に適用し，確率モデルとして N グラムモデルを仮定し，予測誤り率を損失関数と設定した評価基準のもとで最適な単語予測法を導出する．

最初に，仮定する確率モデルを整理する．次に，予測誤り率を損失関数として設定しベイズ基準のもとで最適な単語予測法を導出する．最後に，導出された予測法のアルゴリズムの実現方法を示す．

3.3.1　仮定する確率モデル

最初にパラメトリックな確率モデルを仮定する．具体的には，N グラムモデルにおける次数が $m = 1, 2, \cdots, N$ のパラメータを $\boldsymbol{\theta}_m$ とし，履歴となる単語列 x^{m-1} が与えられたもとで，単語 y が出現する確率を

$$p(y|x^{m-1}) = p(y|x^{m-1}, \boldsymbol{\theta}_m) \tag{3.6}$$

とする．統計的決定理論のアプローチにおいては，仮定する確率モデルのパラメータ $\boldsymbol{\theta}_m$ や N グラムモデルの次数を表す m について，これらの一部，もしくはすべてを未知のものとして議論を行う．なお，N グラムモデルの次数 m はただ一つ存在し，この次数のことを真の次数，もしくは真のモデルの次数と呼ぶ．

3.3.2　単語予測法の導出

単語予測のための決定関数は，履歴となる単語列 $x^{N-1} \in X^{N-1}$ とその次に続く単語 $y \in Y$ の n 個の対である学習データ $(x^{N-1}, y)^n$ と，単語列 x_p^{N-1} が得られたもとで x_p^{N-1} の次に続く単語 $y_p \in Y$ を予測することになる．これは，

$$\hat{y} = D(x_p^{N-1}, (x^{N-1}, y)^n) \tag{3.7}$$

と定義できる．

以上の準備のもと統計的決定理論のアプローチにもとづき，以下のように決定関数を導出する．まず，式 (3.7) で示される決定関数を用いて，予測誤り率を損失関数として設定する．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 (3.6) で示した真のモデルの分布で期待値を取った危険関数を求める．この危険関数を最小にする決定関数を求めるのだが，真のモデルの次数とそのパラメータは未知でありそのままでは最適化ができない．ここでは，これらに事前分布を仮定し，その事前分布で平均化したベイズ危険関数を最小化することで決定関数を導出する．なお，このベイズ危険関数を最小化する基準は，ベイズ基準と呼ばれる．

本研究では，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．

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

$$d(\hat{y}, y_p) = \begin{cases} 0 & (\hat{y} = y_p) \\ 1 & (\hat{y} \neq y_p) \end{cases} \tag{3.8}$$

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，$y_p \in Y$ 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設定すると [1]，

$$L(D(x_p^{N-1}, (x^{N-1}, y)^n), Y|\boldsymbol{\theta}) = \sum_{y_p \in Y} d(D, y_p) p(y_p|x_p^{N-1}, \boldsymbol{\theta}) \tag{3.9}$$

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求めると，

$$\begin{aligned} & R(D(x_p^{N-1}, (x^{N-1}, y)^n), Y|\boldsymbol{\theta}, \boldsymbol{\mu}) \\ = & \sum_{(x^{N-1})^n \in (X^{N-1})^n} \sum_{y^n \in Y^n} L(D, Y) p(y^n|(x^{N-1})^n, \boldsymbol{\theta}) p((x^{N-1})^n|\boldsymbol{\mu}) \end{aligned} \tag{3.10}$$

となる．ただし，$\boldsymbol{\mu}$ は 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

[1] 以下，決定関数 $D(x_p^{N-1}, (x^{N-1}, y)^n)$ 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いることが文献 [60] にて指摘されている．

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，

$$
\begin{aligned}
& B_{\mathrm{risk}}(D(x_p^{N-1},(x^{N-1},y)^n),Y) \\
= & \int_{\boldsymbol{\mu}}\int_{\boldsymbol{\theta}} R(D,Y|\boldsymbol{\theta},\boldsymbol{\mu})f(\boldsymbol{\theta})d\boldsymbol{\theta}f(\boldsymbol{\mu})d\boldsymbol{\mu} \\
= & \sum_{(x^{N-1})^n\in(X^{N-1})^n}\sum_{y^n\in Y^n}\sum_{y_p\in Y}\int_{\boldsymbol{\mu}}\int_{\boldsymbol{\theta}} d(D,y_p)p(y_p|x_p^{N-1},\boldsymbol{\theta}) \\
& p(y^n|(x^{N-1})^n,\boldsymbol{\theta})p((x^{N-1})^n|\boldsymbol{\mu})f(\boldsymbol{\theta})d\boldsymbol{\theta}f(\boldsymbol{\mu})d\boldsymbol{\mu} \\
= & \sum_{(x^{N-1})^n\in(X^{N-1})^n}\sum_{y^n\in Y^n}\sum_{y_p\in Y}\int_{\boldsymbol{\mu}}\int_{\boldsymbol{\theta}} d(D,y_p)p(y_p|x_p^{N-1},\boldsymbol{\theta}) \\
& f(\boldsymbol{\theta}|(x^{N-1})^n,y^n)d\boldsymbol{\theta}p(y^n|(x^{N-1})^n)p((x^{N-1})^n|\boldsymbol{\mu})f(\boldsymbol{\mu})d\boldsymbol{\mu} \\
= & \sum_{(x^{N-1})^n\in(X^{N-1})^n}\sum_{y^n\in Y^n}\sum_{y_p\in Y}\int_{\boldsymbol{\mu}}\Big\{1-\int_{\boldsymbol{\theta}} I_D(y_p)p(y_p|x_p^{N-1},\boldsymbol{\theta}) \\
& f(\boldsymbol{\theta}|(x^{N-1},y)^n)d\boldsymbol{\theta}\Big\}p(y^n|(x^{N-1})^n)p((x^{N-1})^n|\boldsymbol{\mu})f(\boldsymbol{\mu})d\boldsymbol{\mu} \qquad (3.11)
\end{aligned}
$$

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，(3.11) 式に含まれる

$$1-\int_{\boldsymbol{\theta}} I_D(y_p)p(y_p|x_p^{N-1},\boldsymbol{\theta})f(\boldsymbol{\theta}|(x^{N-1},y)^n)d\boldsymbol{\theta} \qquad (3.12)$$

を最小化することで得られる．すなわち，

$$\hat{y}=\arg\max_y \int_{\boldsymbol{\theta}} p(y|x_p^{N-1},\boldsymbol{\theta})f(\boldsymbol{\theta}|(x^{N-1},y)^n)d\boldsymbol{\theta} \qquad (3.13)$$

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

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

が成立することから，

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

の関係を利用した．

真の次数が未知の場合

次に，真のモデルの次数が未知のもとで単語予測法を導出する．なお，距離関数は式 (3.8) を定義する．

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 x_p^{N-1} に含まれる長さ $m-1$ の単語の履歴を 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，

$$L_{\mathrm{h}}(D(x_p^{N-1},(x^{N-1},y)^n),Y|m,\boldsymbol{\theta}_m)=\sum_{y_p\in Y}d(D,y_p)p(y_p|x_p^{m-1},\boldsymbol{\theta}_m) \tag{3.14}$$

となる．

この損失関数に対する危険関数は，

$$\begin{aligned}&R_{\mathrm{h}}(D(x_p^{N-1},(x^{N-1},y)^n),Y|m,\boldsymbol{\theta}_m,\boldsymbol{\mu})\\ =&\sum_{(x^{N-1})^n\in(X^{N-1})^n}\sum_{y^n\in Y^n}L_h(D,Y|m,\boldsymbol{\theta}_m)\\ &p(y^n|(x^{N-1})^n,\boldsymbol{\theta}_m)p((x^{N-1})^n|\boldsymbol{\mu})\end{aligned} \tag{3.15}$$

となる．

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

$$\begin{aligned}&B_{\mathrm{h,risk}}(D(x_p^{N-1},(x^{N-1},y)^n),Y)\\ =&\int_{\boldsymbol{\mu}}\sum_{m=1}^{N}p(m)\int_{\boldsymbol{\theta}_m}R_h(D,Y|m,\boldsymbol{\theta}_m,\boldsymbol{\mu})f(\boldsymbol{\theta}_m)d\boldsymbol{\theta}_m f(\boldsymbol{\mu})d\boldsymbol{\mu}\\ =&\sum_{(x^{N-1})^n\in(X^{N-1})^n}\sum_{y^n\in Y^n}\sum_{y_p\in Y}\int_{\boldsymbol{\mu}}\sum_{m=1}^{N}p(m)\int_{\boldsymbol{\theta}_m}d(D,y_p)p(y_p|x_p^{m-1},\boldsymbol{\theta}_m)\\ &p(y^n|(x^{N-1})^n,\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m)d\boldsymbol{\theta}_m p((x^{N-1})^n|\boldsymbol{\mu})f(\boldsymbol{\mu})d\boldsymbol{\mu}\end{aligned} \tag{3.16}$$

となる．ベイズ危険関数の最小値は (3.16) 式に含まれる，

$$\begin{aligned}
& \sum_{m=1}^{N} p(m) \int_{\boldsymbol{\theta}_m} d(D, y_p) p(y_p | x_p^{m-1}, \boldsymbol{\theta}_m) \\
& p(y^n | (x^{N-1})^n, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\
= \; & 1 - \sum_{m=1}^{N} \int_{\boldsymbol{\theta}_m} I_D(y_p) p(y_p | x_p^{m-1}, \boldsymbol{\theta}_m) \\
& f(\boldsymbol{\theta}_m | y^n, (x^{N-1})^n) d\boldsymbol{\theta}_m p(m | (x^{N-1}, y)^n) / p(y^n | (x^{N-1})^n)
\end{aligned} \tag{3.17}$$

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

$$\begin{aligned}
\hat{y} \; = \; & \arg\max_{y} \sum_{m=1}^{N} p(m | (x^{N-1}, y)^n) \\
& \int_{\boldsymbol{\theta}_m} p(y | x_p^{N-1}, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m | (x^{N-1}, y)^n) d\boldsymbol{\theta}_m
\end{aligned} \tag{3.18}$$

となる $\hat{y}$ を出力することになる．

条件付き多項分布を仮定した単語予測法

ここで，式 (3.18) に含まれる予測分布の算出に必要な積分計算は，単語の出現する分布に条件付き多項分布，そのパラメータ $\boldsymbol{\theta}$ の事前分布 $f(\boldsymbol{\theta})$ にディレクレ分布を仮定することで，自然共役の関係から，

$$\begin{aligned}
& \int_{\boldsymbol{\theta}_m} p(y | x_p^{N-1}, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m | (x^{N-1}, y)^n) d\boldsymbol{\theta}_m \\
= \; & \frac{c(y | x_p^{m-1}) + \alpha(y | x_p^{m-1})}{\sum_{y \in \boldsymbol{Y}} c(y | x_p^{m-1}) + \sum_{y \in \boldsymbol{Y}} \alpha(y | x_p^{m-1})}
\end{aligned} \tag{3.19}$$

により容易に求められる [4, 5]．ただし，$\alpha(y|x^{m-1})$ は，$p(y|x_p^{m-1}, \boldsymbol{\theta}_m)$ に対応するディレクレ分布のパラメータ，$c(y|x^{m-1})$ は (3.4) 式と同様に学習データに含まれる x^{m-1} に後続する y の出現頻度をそれぞれ表す．

また，モデルの事後確率 $p(m|(x^{N-1},y)^n)$ は，ベイズの定理より

$$\begin{aligned} p(m|(x^{N-1},y)^n) &\propto p((x^{N-1},y)^n|m)p(m) \\ &= p(m)\prod_{i=1}^{n} p(x_i^{N-1},y_i|m) \\ &= p(m)\prod_{i=1}^{n}\int_{\boldsymbol{\theta}_m} p(y_i,x_i^{N-1},\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m \\ &\propto p(m)\prod_{i=1}^{n}\int_{\boldsymbol{\theta}_m} p(y_i|x_i^{N-1},\boldsymbol{\theta}_m)f(\boldsymbol{\theta}_m|m)d\boldsymbol{\theta}_m \end{aligned} \tag{3.20}$$

から容易に求まる．

このように，単語予測のための最適解が閉じた形式で求まり，学習データの追加に対して加算増分により更新できる．そのため，オンライン処理にも適用可能である点において計算量的な優位性を有した予測法である．

なお，式 (3.18) は，N グラムモデルを確率モデルとして仮定し，そのパラメータと N グラムモデルの次数を未知のもと，予測誤り率に関するベイズ基準のもとで最適な予測法を決定関数として導出したものである．そのため，N グラムモデルを決定関数として条件付き多項分布のパラメータをチューニングしたものとは異なる決定関数が求まっている．

3.3.3　アルゴリズムの実現方法

本研究で対象としている N グラムモデルでは，履歴となる $N-1$ 個の単語列 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 [34] を利用した単語予測法（以下，提案法とする）の実現方法を説明する．

まず，接尾木は図 3.1 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．例えば，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 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．

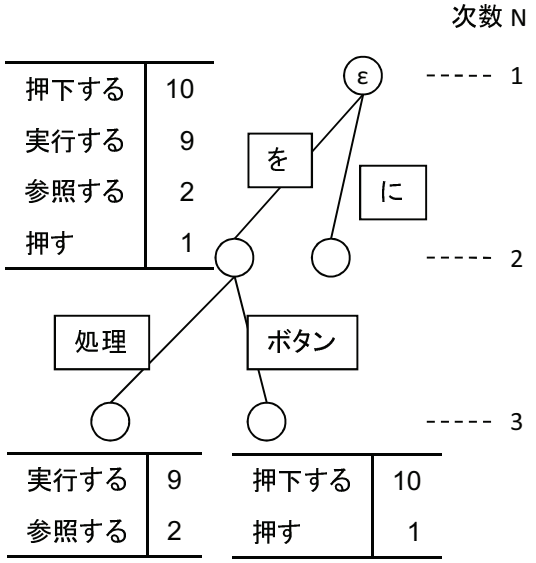

図 3.1: 単語予測に用いる接尾木の例

学習を行うに当たり最初に接尾木を構築する．具体的には，学習データの 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 (3.19) をもとに各節点に対応する予測分布を求める．さらに，式 (3.20) 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，(3.18) 式に含まれる，

$$\sum_{m=1}^{M} p(m|(x^{N-1},y)^n) \int_{\boldsymbol{\theta}_m} p(y|x_p^{N-1},\boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|(x^{N-1},y)^n) d\boldsymbol{\theta}_m \tag{3.21}$$

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ーに必要な量は節点の数と，各節点で保持する単語の種類の和に比例する [4] ．

予測を行う場合は，与えられた 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が最大である単語の予測値 $\hat{y}$ を出力する．なお，x_p^{N-1} と

[4]なお，メモリーを最大に利用する場合は，$|Y||X|^{N-1}$ 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，自然言語の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．

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，本研究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，3.2 節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にて予測値を出力するため，使用するメモリー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．

上記にて説明したとおり，学習の手順は以下の，

1. 接尾木の構築
2. 各節点の単語の予測分布の算出
3. 次数の重みの算出
4. 次数を階層的に加算

の四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 (3.19) 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類の和に比例する．(3) は，学習データごとに式 (3.20) 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の和に比例する．

EM アルゴリズムを用いる方法が提案法と異なる点としては，学習データを二つに分割し，(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．従って，モデルの重み $w(m)$ の算出の繰り返し計算の部分が，提案法よりも増加する．

また，ニーザー・ナイ法を用いる場合は (3) のステップが異なる．これは，式 (3.4) に含まれる d_{discount} を学習データ全体から算出する．d_{discount} 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，提案法のメモリー量および計算量は従来法とほぼ同等であると言える．

3.4　提案法の視点から解釈される従来法の特徴

ニーザー・ナイ法として知られている式 (3.4) は，高次の項 $\frac{c(y|x^m)}{\sum_{y\in Y} c(y|x^m)}$ と，低次の項 $D_{\mathrm{kn}}(x^{m-1},\boldsymbol{\theta})$ に $\frac{d_{\mathrm{discount}}}{\sum_{y\in Y} c(y|x^m)}\left|\{y : c(y|x^m) > 0\}\right|$ の重み付けを行ったものとの和を算出していることに対応する．

提案法のアルゴリズムである式 (3.18) では，ここに含まれるモデルの次数 m の予測分布と事後確率の積を最大次数まで和をとった形式となっている．この和の算出においては，(3.4) 式と類似した漸化式の形式でも記述が可能である．具体的には，モデルの次数 m の予測分布を

$$F(m) = \int_{\boldsymbol{\theta}_m} p(y|x_p^{N-1},\boldsymbol{\theta}_m,m) f(\boldsymbol{\theta}_m|(x^{N-1},y)^n,m) d\boldsymbol{\theta}_m \tag{3.22}$$

予測分布の漸化式を $F_{\mathrm{r}}(m)$，各次数のモデルの事後確率を

$$G(m) = P(m|(x^{N-1},y)^n) \tag{3.23}$$

とすると，

$$F_{\mathrm{r}}(m) = F(m) + \frac{G(m-1)}{G(m)} F_{\mathrm{r}}(m-1) \tag{3.24}$$

となる．これを，式 (3.18) に代入することで同様の予測法が実現可能である．

このように，従来法と提案法で類似した記述が可能であり，このことから従来法の効果を解釈すると以下のようになる．まず，従来法の低次の重みは $\left|y : c(y|x^m) > 0\right|$ に比例するが，これは単語列 x^m に後続する単語 y の種類を意味し，その種類が多い場合は低次の項の重みを増すことになる．

このことを予測および事後確率の視点から整理すると，y の種類が多い節点は予測が困難で事後確率は小さくなる傾向があるといえ，そのため低次の節点の重みを増していると解釈できる．このように，従来法であるニーザー・ナイ法は，提案法で示された予測分布を事後確率の重みで足し合わせることを近似していると解釈できる．

3.5　文書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，既存の隠れ変数を仮定した EM アルゴリズムを用いる方法，ニーザー・ナイ法，提案法のそれぞれで，学習データから予測法を求め，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法の比較を行い，提案法が実用的にも有効であることを示す．

3.5.1　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]．

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 2,553 種類，システム開発文書が 1,989 種類であった．

3.5.2　単語予測実験の結果

単語予測の正答率の評価にあたっては，従来法として 3.2 節で説明した，EM アルゴリズムを用いる方法，ニーザー・ナイ法を用いた．提案は式 (3.18) にもとづく予測法を用いた．それぞれの方法に対して，学習データをもとに予測法を求め，検証データによる単語予測の実験を行った．

[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．

EM アルゴリズムを用いる方法の学習フェーズでは，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，もう一方を次数の重みの算出に用いた EM アルゴリズム [34] 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，次数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，式 (3.1) に従い各次数を加算した．

ニーザー・ナイ法は，現在，高性能として広く知られている修正ニーザー・ナイ法 [7] を用いた．この方法は，式 (3.4) に含まれる d_{discount} の算出に対して，式 (3.5) で示した方法を拡張した，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レ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通り設定した．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法で，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]．ディレ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，3.5.1 節で示した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，各方法に対して，履歴となる単語列が到達できる最高次の接点をもとに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，予測法により算出された値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，予測法により算出された値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 3.1，表 3.2 に示す．表中の N はモデルの単語列の最大長の上限値であり，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，予測法により算出された値が最上位の単語を一つ出力した場合，上位 5 位は，予測法により算出された値の大きい単語を五つ出力した場

[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 3.1: 各方法による特許文書に対する単語予測結果の正答率（単位 %）

	最上位			上位 5 位を出力		
N	EM	MKN	提案	EM	MKN	提案
3	37.07	44.51	**44.52**	61.07	64.83	**65.00**
4	46.47	46.68	**46.70**	65.88	66.14	**66.16**
5	49.92	52.48	**52.49**	67.55	**68.99**	**68.99**
6	52.76	53.51	**53.57**	68.20	**69.14**	**69.14**

表 3.2: 各方法によるシステム開発文書に対する単語予測結果の正答率（単位 %）

	最上位			上位 5 位を出力		
N	EM	MKN	提案	EM	MKN	提案
3	46.34	**52.17**	**52.17**	71.06	78.58	**78.73**
4	64.16	**64.43**	64.41	83.53	**83.93**	83.83
5	67.86	73.16	**73.19**	84.73	**86.26**	86.17
6	75.48	**75.84**	75.65	86.00	**86.57**	86.52

合を表す．EM は隠れ変数を仮定した EM アルゴリズムを用いる方法，MKN は修正ニーザー・ナイ法，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，すべての方法で 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・ナ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・ナ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．

3.6　本章のまとめ

本章では，まず，単語予測の問題に対して，従来研究で提案されている様々な方法が，統計的言語処理の一般的アプローチを基に検討されていることを示した．

次に，確率モデルに統計的言語処理で基盤となる N グラムモデルを仮定し N グラムモデルの真の次数が未知の場合において，N グラムモデルの確率パラメータ $\boldsymbol{\theta}$ の事前分布と，N グラムモデルの次数 m のモデルの事前分布を用いたベイズ基準のもとで最適となる単語予測法を導出した．

この単語予測法は，与えられた学習データ $(x^{N-1}, y)^n$ に対して，次数 m ごとのパラメータの事後分布の積分により求められる予測分布をモデルの事後確率により重み付けする形式となっている．この形式は従来の式 (3.1) で表現される重み付け法と形式的に類似しているが，本研究により $p(y|x^{N-1}, \boldsymbol{\theta}_m)$ を m の予測分布，$w(m)$ を m の事後確率としたものが，単語の予測誤り確率を損失関数として設定したベイズ基準のもとで最適な単語予測法であることが明らかになった．

この予測法に対して，単語の発生する分布に条件付き多項分布，そのパラメータの事前分布にディレクレ分布を仮定した場合の具体的なアルゴリズムを示した．このアルゴリズムは，提案する基準に対して最適解が閉じた形式で求まり，学習データの追加に対して加算増分により更新できる．オンライン処理にも適用可能である点において計算量的な優位性を有したアルゴリズムであることを示した．また，アルゴリズムの具体的な実現方法の説明をもとにアルゴリズムに要するメモリー量および計算量を考察し，従来法とほぼ同等であることを示した．

また，従来研究にてアドホックなパラメータを設定する方法であるニーザー・ナイ法について，形式的な類似性をもとに議論した．ここでは，低次の次数と高次の次数を加算する重みパラメータについて，$m-1$ 次と m 次のモデルの事後確率の比を近似したものと解釈できることを示した．

最後に，実データである文書データによる単語予測実験を行った．これにより，実証的な検討から有効性が知られていた修正ニーザー・ナイ法と比較して，ほぼ同等の単語予測の正答率を持つといえ実用的にも適正に動作することを示した．

第4章　問題2：分野適応問題に対する統計的決定理論のアプローチの適用

4.1　はじめに

自然言語で記述される文書データにおいては，記載される話題の分野により用いられる用語が異なるため，単語の出現頻度の分布の傾向が異なることが想定される．単語予測を実施するにあたっては，対象とする分野に適応した予測が行えるようにすることが望ましいと考えられる．

しかしながら，一般には単語予測の対象とする分野に対応したデータが大量に得られることはまれである．そのため，他の類似する分野のデータ（以下，類似データと呼ぶ）を追加し，予測精度を向上させることが検討されている．これは，例えば，システム開発分野における業務文書作成の支援を想定した単語予測法を開発する場合，当該業務のデータ（以下，当該データと呼ぶ）は少量にしか与えられないものの，類似の分野に属した文書データとして，例えば，公開データである特許文書や，科学技術論文など，大量に入手可能なデータを活用することを意味する．

このような異なる母集団から得られるデータを用いて当該分野に適応するように単語予測法を検討する問題は，分野適応と呼ばれ様々な研究が行われている．ここで課題となるのは，当該データと類似データの母集団が異なるため，出現する単語が類似することは期待されるものの，同一の確率分布から発生するとは限らないため何かしらの適応処理が必要となる．

これに対して，当該データと類似データの出現頻度の推定値をもとに，双方の重み付けを行うことで分野適応を行う方法が提案されている [15]．また，これの簡易

な方法として，当該データと類似データのそれぞれに含まれる単語の出現頻度に対して，何かしらの重み付けを行い加算する方法が提案されている [2, 48]．その他の方法として，類似データを事前知識と見なし，最大事後確率推定を前提とした単語予測法 [41] をもとに，その形式に含まれる事前分布について，類似データを利用して設定する方法が提案されている [39, 2]．これらの方法は，実用的な観点による有効性が報告されており [51, 48]，分野適応におけるひとつの標準的な方法であるといえる．

本研究においては，この問題に統計的決定理論のアプローチを適用する．具体的には，当該データと類似データの類似性を，確率パラメータの変化する数理モデルで表現した確率モデルを仮定し，予測誤り率を損失関数と設定した評価基準のもとで最適な予測法を導出する．あわせて，導出された予測法の形式をもとに，従来研究にて提案されてきた方法に，理論的な解釈が与えられることを確認する．

なお，提案するアプローチにおいては，事前分布を仮定することから類似データを事前情報として活用することとなる．本研究のアプローチにより求められる予測法は，漸近的な一致性を持つことが示されている [50]．この結果を分野適応問題にあてはめると，類似データを事前分布として利用する場合と利用しない場合の双方で，学習データとして用いる当該データを増加させた場合，双方で求められた予測法が一致することを意味する．このような性質は，例えば，利用者に提供する際は当該データが少量しか与えられないものの，文書作成を継続することで当該データの学習用データが蓄積でき，適宜予測法を更新することにより類似データの影響を小さくできることを意味する．これを想定した実験的な検証を行うことで，この理論的な性質を支持する結果が得られることを確認する．

4.2　従来研究

まず，分野適応問題の統計的言語処理の分野の従来法として，予測式を

$$D_{\text{da-inter}}(x_p^{N-1}, \boldsymbol{\theta}) = w_f P(y|x_p^{N-1}, \boldsymbol{\theta}_f) + w_t P(y|x_p^{N-1}, \boldsymbol{\theta}_t) \tag{4.1}$$

とする方法が提案されている．ただし，$\boldsymbol{\theta}_t$ および w_t はそれぞれ当該データの予測のためのパラメータとその重み，$\boldsymbol{\theta}_f$ および w_f はそれぞれ類似データの予測のための

パラメータとその重みである．ここで示した重みは，例えば削除補完法 [15] と呼ばれる隠れ変数を仮定した推定法により決定される．

この形式を計算が簡易になるように変更した予測式

$$D_{\text{da-simple}}(x_p^{N-1}, \boldsymbol{\theta}) = \frac{c_t(y|x_p^{N-1}) + w\{c_f(y|x_p^{N-1})\}}{\sum_{y\in Y} c_t(y|x_p^{N-1}) + w\{\sum_{y\in Y} c_f(y|x_p^{N-1})\}} \tag{4.2}$$

が提案されている [2, 48]．これは，類似データに含まれる単語の出現頻度に一定の重み w をかけたものである．ただし，$c_t(y|x^{N-1})$ は当該データに含まれる x^{N-1} に後続する単語 y の出現頻度，$c_f(y|x^{N-1})$ は類似データに含まれる x^{N-1} に後続する単語 y の出現頻度，w は,$0 < w < 1$ となる定数である．その他の方法として，最大事後確率推定を前提とした単語予測法では，

$$\hat{\boldsymbol{\theta}}_{\text{map}} = \arg\max_{\theta} p((x^{N-1}, y)^n|\boldsymbol{\theta}) f(\boldsymbol{\theta}) \tag{4.3}$$

となるよう $\hat{\boldsymbol{\theta}}_{\text{map}}$ を推定し，確率モデルに $\hat{\boldsymbol{\theta}}_{\text{map}}$ をビルドインし，その確率の推定値が最大となる単語を出力することで予測を行う．ただし，$(x^{N-1}, y)^n$ は n 個の学習データの対，$f(\boldsymbol{\theta})$ は確率モデルのパラメータ $\boldsymbol{\theta}$ の事前分布である．これを前提とした分野適応の研究においては，この $f(\boldsymbol{\theta})$ の設定方法に対する様々な方法が提案されている [39, 51] [1]．式 (4.1) や式 (4.2) で示した方法は，確率モデルと決定関数を明確に区別せずに決定関数の形式を最初に決める，第 2 章で示した統計的言語処理の一般的アプローチといえる．また，式 (4.3) にて表現される方法は，確率モデルと決定関数の区別はなく，提案する統計的決定理論のアプローチとは異なる．これらの研究においては，学習データを元にしたテストセット・パープレキシティの指標を用いた有効性の議論が中心である．そのため，単語予測に対する理論的な性質は明らかでない．

4.3　統計的決定理論のアプローチの適用

本節では，統計的決定理論のアプローチに基づき単語予測の分野適応問題に対する予測法の議論を行う．

[1] なお，事前分布の設定方法の研究においては，どのようなデータを利用するかもひとつの課題となっており，分野という観点のみならず，新聞記事のような話題が経時的に変化するような課題に対する様々な方法が議論されている [51].

最初に，類似データと当該データの類似性を仮定した確率モデルを整理する．次に，この確率モデルを仮定し，予測誤り率を損失関数とし，設定した基準のもとで最適な単語予測法を導出する．また，導出された予測法のアルゴリズムの実現方法を示し，最後に，統計的言語処理の分野にて従来法として知られる方法について，理論的な解釈が与えられることを示す．

4.3.1　仮定する確率モデル

まず，類似データと当該データの類似性に対して，これらのデータの双方が条件付き多項分布で出現するとし，この分布に含まれるパラメータが確率的な変化を起こす数理モデルを考える．これは，ある単語の発生確率の大小変化について，指向性を持たせない変化パターンを仮定していることになる．

なお，ここで示す確率的パラメータの変化を扱った数理モデルは，統計学の分野を中心に古くから研究されている．たとえば，ガウス過程のランダムウォーク型モデルであるカルマンフィルター [12] や，これをガウス過程以外の指数型分布族のクラスまで拡張した Simple Power Steady Model (SPSM)[27] などが挙げることができる．この確率的パラメータの変化は，実数定数の超パラメータひとつのみを利用して表現されるランダムウォーク型のクラスとなっていることが特徴である．また，このクラスによって，この超パラメータの値が既知であれば，確率モデルに含まれるパラメータの推定や，単語予測が解析的に，かつそれらの四則演算のみで可能となる．

これを定式化すると以下のようになる．まず，類似データ，当該データのそれぞれについて，パラメータ，

$$\boldsymbol{\theta}_f(x^{N-1}, m) = (\theta_f(y_1|x^{N-1}, m), \cdots, \theta_f(y_{|Y|}|x^{N-1}, m)) \tag{4.4}$$

$$\boldsymbol{\theta}_t(x^{N-1}, m) = (\theta_t(y_1|x^{N-1}, m), \cdots, \theta_t(y_{|Y|}|x^{N-1}, m)) \tag{4.5}$$

をもつ条件付き多項分布に従って単語 y が発生し，このパラメータ $\boldsymbol{\theta}_f(x^{N-1}, m)$ が，$\boldsymbol{\theta}_t(x^{N-1}, m)$ に確率的に変化すると仮定する．ここで，以下は議論は単語列 x^{N-1} や，N グラムモデルの次数 m に依存しないため，式 (4.4) に含まれる $\boldsymbol{\theta}_f(x^{N-1}, m)$ を $\boldsymbol{\theta}_f$，

$\theta_f(y_1|x^{N-1}, m)$ を $\theta_{f,y}$，同様に，式 (4.5) に含まれる $\boldsymbol{\theta}_t(x^{N-1}, m)$ を $\boldsymbol{\theta}_t$，$\theta_t(y_1|x^{N-1}, m)$ を $\theta_{t,y}$ と簡潔に表現する．

これらのパラメータが確率的に変化する数理モデルとして，$\theta_{f,y}$ と $\theta_{t,y}$ は，それぞれ，

$$\theta_{f,y} = \frac{A_{f,y}}{\sum_{y\in Y} A_{f,y}}, y \in Y \tag{4.6}$$

$$\theta_{t,y} = \frac{A_{t,y}}{\sum_{y\in Y} A_{t,y}}, y \in Y \tag{4.7}$$

で表現されるものとする．ここで，$A_{f,y}$，$A_{t,y}$ は

$$A_{t,y} = A_{f,y}U_{f,y} \tag{4.8}$$

という関係によりパラメータが確率的に変化する．ただし，$U_{f,y}$ は，

$$f(U_{f,y}) = \frac{\Gamma(\rho\alpha'_{f,y} + (1-\rho)\alpha'_{f,y})}{\Gamma(\rho\alpha'_{f,y})\Gamma((1-\rho)\alpha'_{f,y})} U_{f,y}^{\rho\alpha'_{f,y}-1}(1-U_{f,y})^{(1-\rho\alpha'_{f,y})-1} \tag{4.9}$$

$$= Beta(\rho\alpha'_{f,y}, (1-\rho)\alpha'_{f,y}) \tag{4.10}$$

で表現されるベータ分布で，$\rho \in (0, 1)$，$\alpha'_{f,y}$ は

$$\alpha'_{f,y} = c_f(y|x^{N-1}) + \alpha_{f,y} \tag{4.11}$$

である．また，$A_{f,y}$ の事前分布としては，

$$f(A_{f,y}) = \frac{1}{\Gamma(\alpha_{f,y})} A_{f,y}^{\alpha_{f,y}-1} \mathrm{e}^{-A_{f,y}}$$

$$= Ga(\alpha_{f,y}) \tag{4.12}$$

で表現されるガンマ分布を仮定する．

ここで，式 (4.12) の仮定は，$\boldsymbol{\theta}_f$ の事前分布に

$$f(\boldsymbol{\theta}_f) = Dir(\alpha_{f,y_1}, \alpha_{f,y_2}, \cdots, \alpha_{f,y_{|Y|}}) \tag{4.13}$$

となるディレクレ分布を仮定していることと等価になる [13, 52]．

このように，条件付き多項分布のパラメータが確率的に変化する数理モデルにより，類似データと当該データの類似性を表現した確率モデルを仮定する．なお，この変化モデルは，多項分布のパラメータ変化をランダムウォークとして表現した Simple Power Steady Model (SPSM) と呼ばれる確率モデルに対応している．

4.3.2　単語予測法の導出

4.3.1 節では，仮定する確率モデルの定式化を行った．本節では，この定式化をもとに，式 (4.8) で変換される確率変数 $A_{t,y}$ の分布を求める．これは，条件付き多項分布で発生する類似データの事後分布が，パラメータ変化モデルにより変化したパラメータによる分布である．すなわち，これは当該データの事前分布に対応することになる．この事前分布を式 (3.19) で表される 3.3 節で導出した予測法に含まれる予測分布に代入することで，分野適応問題を想定した単語予測法を導出する．

はじめに，式 (4.13) で仮定したパラメータ $\boldsymbol{\theta}_f$ をもつディレクレ分布の事前分布に対して，類似データ $(x^{N-1}, y)^{n_f}$ が与えられた場合の事後分布を求める．これは，

$$\begin{aligned} f(\boldsymbol{\theta}_f|(x^{N-1},y)^{n_f}) &= Dir(c_f(y|x^{N-1})+\alpha_{f,y_1}, c_f(y|x^{N-1})+\alpha_{f,y_2}, \\ &\quad \cdots, c_f(y|x^{N-1})+\alpha_{f,y_{|Y|}}) \\ &= Dir(\alpha'_{f,y_1}, \alpha'_{f,y_2}, \cdots, \alpha'_{f,y_{|Y|}}) \end{aligned} \tag{4.14}$$

となる．この結果は，式 (4.12) と式 (4.13) で示したと関係と同様に，

$$\begin{aligned} f(A_{f,y}|(x^{N-1},y)^{n_f}) &= \frac{1}{\Gamma(\alpha'_{f,y})} A_f^{\alpha'_{f,y}-1} e^{-A_{f,y}} \\ &= Ga(\alpha'_{f,y}) \end{aligned} \tag{4.15}$$

と等価である．さらに，式 (4.15) と式 (4.12) を式 (4.8) に代入すると，

$$f(A_{t,f}|(x^{N-1},y)^{n_f}) = \frac{1}{\Gamma(\rho\alpha'_{f,y})} A_{f,y}^{\alpha'_{f,y}-1} \mathrm{e}^{-A_{f,y}} \tag{4.16}$$

となる [2]．

この結果に対して，式 (4.12) と式 (4.13) で示した関係から，

$$f(\boldsymbol{\theta}_t|(x^{N-1},y)^{n_f}) = Dir(\rho\alpha'_{f,y_1}, \rho\alpha'_{f,y_2}, \cdots, \rho\alpha'_{f,y_{|Y|}}) \tag{4.17}$$

となる．

[2] 導出の詳細は付録 A を参照．

結局，式 (4.17) は，当該データの事前分布として仮定したディレクレ分布のパラメータに，$\rho\alpha'_{f,y}, y \in Y$ を利用することを意味する．これを，式 (3.19) の予測分布に代入すると，

$$\frac{c_t(y|x_p^{m-1}) + \rho\{c_f(y|x_p^{m-1}) + \alpha_f(y|x_p^{m-1})\}}{\sum_{y\in Y} c_t(y|x_p^{m-1}) + \rho\{\sum_{y\in Y} c_f(y|x_p^{m-1}) + \sum_{y\in Y} \alpha_f(y|x_p^{m-1})\}} \tag{4.18}$$

が導出される．すなわち，この予測分布をもとに単語 y の確率値を推定し，その値が最大となる $\hat{y}$ を出力することになる．

これは，真のモデルの次数が未知の場合も同様であり，結局，決定関数としては，

$$\begin{aligned} & D_{\mathrm{da}}(x_p^{N-1}, (x^{N-1}, y)^{n_t}, (x^{N-1}, y)^{n_f}) \\ = & \sum_{m=1}^{N} P(m|(x^{m-1}, y)^{n_t}) \\ & \frac{c_t(y|x_p^{m-1}) + \rho\{c_f(y|x_p^{m-1}) + \alpha_f(y|x_p^{m-1})\}}{\sum_{y\in Y} c_t(y|x_p^{m-1}) + \rho\{\sum_{y\in Y} c_f(y|x_p^{m-1}) + \sum_{y\in Y} \alpha_f(y|x_p^{m-1})\}} \end{aligned} \tag{4.19}$$

が導出される．

このように，単語予測の分野適応問題に拡張したとしても，もとの問題の確率モデルに拡張部分の確率モデルを加えて仮定することで，予測誤り率に関するベイズ基準のもとで最適な決定関数が導出できる．すなわち，確率モデルの親和性を保つようにすると，もとの決定関数を拡張した形式の決定関数が求まる．

4.3.3　アルゴリズムの実現方法

これまでの説明の通り，導出された予測法（以下，提案法と呼ぶ）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 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パラメータの事前分布として利用することで，単語予測のための分野適応の対応がアルゴリズムを変更することなく実施できることを示す．

まず，ディレ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 [4, 5] 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，

ディレクレ分布のパラメータとして利用することになる．これは，分野適応の対応方法として示した，式 (4.18) に含まれる，$\rho\{c_f(y|x_p^{m-1}) + \alpha_f(y|x_p^{m-1})\}$ をディレクレ分布のパラメータとして利用することを意味する．これは，ρ の値を決定すれば求めることは容易である．次に，式 (3.20) を算出することになるが，形式は変わらないため算出方法に変更はない．

アルゴリズムをまとめると，類似データの予測対象となる単語の出現頻度を当該データの事前分布であるディレ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当該データの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となる．

なお，ρ の決定においてはいくつかの方法がある [48] が，本研究では以下のようにする．まず，類似データは当該データよりも大量に入手可能であることが想定される．各節点が保持する出現頻度をそのまま当該データ用の事前分布であるディレクレ分布のパラメータに利用すると，当該データの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にて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する．

具体的には，類似データで出現した履歴となる単語列 x^{m-1} のそれぞれに対して，式 (4.18) に含まれる $\sum_{y\in Y} c_f(y|x^{m-1}) + \sum_{y\in Y} \alpha_f(y|x^{m-1})$ が ρ となるように，$c_f(y|x^{m-1}) + \alpha_f(y|x^{m-1})$ 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．

4.4　提案法の視点から解釈される従来法の特徴

分野適応のために式 (4.19) を用いて対応することは，4.3.1 節で示した類似データと当該データの確率分布を仮定し，予測誤り率に関するベイズ基準のもとで最適な単語予測法であることが示された．ここで，従来法である式 (4.2) と提案するアプローチにより N グラムモデルの次数が既知のもとで導出された (4.18) を比較すると，類似した形式であるといえる．

具体的には，双方とも，ある単語に対して当該データに含まれる出現頻度に，類似データに含まれる単語の出現頻度を調整するパラメータを乗算し加算している．違

いとしては，類似データの事前分布の考慮の有無である．本研究のアプローチにより求められる予測法は，漸近的な一致性を持つことが示されている [50] ため，類似データは大量に得られることを想定すると，この違いによる差は微少であるといえる．すなわち，従来法として知られる式 (4.2) の形式による分野適応を想定した単語予測法は，4.3.1 節で定式化された確率モデルを仮定し，N グラムモデルの次数が既知，すなわちある値に固定したもとで，適切な予測法であることが明らかになった．

4.5　実データによる分野適合による単語予測実験

本節では，学習データとして利用できる当該データが少量しか得られない場合を想定し，提案アプローチにより導出された予測法に対して，類似データを事前分布として利用することの効果を検証する．

具体的には，事前分布の設定として，事前知識がない場合を無情報事前分布 [42]，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用いてそれぞれ予測法を求め，検証データによる単語予測実験をもとに予測の正答率の比較を行う．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．これは，事前分布の設定を，無情報事前分布とした場合と，類似データによる事後分布のパラメータを変動させた値を事前分布として導入した場合のそれぞれで，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．

4.5.1　文書データの条件

実験に用いるデータは，3.5 節で利用したデータをもとに事前知識とする類似データを特許文書のデータ，少量の当該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たものを利用した．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，3.5.2 節の実験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，3.5.2 節の実験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 4.3.3 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，学習データによる予測分布の更新が行われないため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当該データの双方に含まれる場合は当該データの影響を高め，類似データの影響としては当該データに含まれない x^m に対する予測の効果を狙い $\rho = 0.01$ とした [3] ．

4.5.2　単語予測実験の結果

$N = 3$ とした実験結果を表 4.1，4.2 に示す．表中の，学習データの件数は当該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 [65] 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，当該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分布の影響を受け，データの増加に従い事前分布の影響が小さくなる性質をもつといえる．

ここで示した方法は漸近的な一致性を持つことが示されている [50]．これは，本実験で使用した二種類の事前分布の設定方法に対して当該データの増加に従い，双方で求められた決定関数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，

[3] なお，本実験にて 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，類似データと当該データの学習データの双方で出現した 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 x^m に対する単語の正答数を上回ったためであった．

表 4.1: $N = 3$ での事前知識の利用有無での単語予測結果の比較（最上位の場合，単位 %）

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
232	22.47	23.70	**有為**	**有為**
464	26.86	27.75	**有為**	**有為**
928	34.31	34.15	無為	無為
1,856	37.31	37.14	無為	無為
3,712	41.37	41.45	無為	無為
7,424	45.05	45.09	無為	無為
14,848	48.10	48.14	無為	無為

[50] の解析結果を支持した結果となっている．

実応用を想定すると，当該データが少量しか得られない場合でも，業務を継続しそのデータを蓄積することで当該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，上記の結果から考察される予測法の性質として，データ量の増加に従い予測法を更新することで，類似データの影響よりも追加された当該データの影響が大きくなることが好ましいといえる．提案法は，システム提供者が用意した類似データが適用先と類似しなかったとしても，事前分布の再調整を行うことなしに，データの追加による予測法の更新が可能となる望ましい性質をもつことが示唆される．

4.6　本章のまとめ

本章では，まず，単語予測における分野適応の問題に対して，従来研究で提案されている様々な方法が統計的言語処理の一般的アプローチに基づくことを示した．次に，統計的決定理論のアプローチをもとに，類似データと当該データの類似性に対して，これらのデータの双方で，条件付き多項分布で出現するとし，この分布に含まれるパラメータが確率的な変化を起こす数理モデルを確率モデルとして仮定し，予

表 4.2: $N = 3$ での事前知識の利用有無での単語予測結果の比較（上位 5 位の場合，単位 %）

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
232	38.31	39.63	**有為**	**有為**
464	44.99	45.45	**有為**	無為
928	52.29	52.09	無為	無為
1,856	57.53	57.33	無為	無為
3,712	62.02	62.08	無為	無為
7,424	67.37	67.35	無為	無為
14,848	72.15	72.12	無為	無為

測誤り率に関するベイズ基準のもとで，理論的に最適となる単語予測法を導出した．さらに，従来研究でよく知られている方法は，導出された予測法を近似した形式であることを示し，従来法に対する理論的な解釈を与えられることを示した．

また，提案法にて類似データの事後分布のパラメータが確率的に変化した分布を，当該データの事前分布として利用した単語予測実験を行い，学習データの量が少ない場合に単語予測の正答率が向上することを示した．さらに，学習データ量を追加した場合の実験を行い，理論的な解析結果により提案法は漸近的な一致性を持つことが知られていることに対して，この結果を支持する結果となることを示した．

第5章　結論

5.1　まとめ

本研究では，統計的決定理論のアプローチの特徴をもとに，統計的言語処理の一般的アプローチとの違いを整理した．さらに，統計的言語処理の分野で代表的な単語予測問題，および，それを拡張した分野適応問題に統計的決定理論のアプローチをそれぞれ適用し，予測誤り率に関するベイズ基準のもとで最適な予測法を導出した．これは，統計的決定理論のアプローチの，新たな応用分野を付け加えることの礎となる結果と考えられる．

第 2 章では統計的決定理論のアプローチを整理し，従来研究である統計的言語処理の一般的アプローチとの差異を整理し，本研究の位置づけを明確にした．統計的言語処理の一般的アプローチにおいては，確率モデルと決定関数を明確に区別せず，決定関数の形式を最初に決めたもとで，学習データによる決定関数に含まれるパラメータのチューニングの方法を議論すること中心であることを示した．提案する統計的決定理論のアプローチでは，確率モデルと決定関数を明確に区別し，決定関数の形式を最初に決めず，予測誤差を損失関数として設定した基準のもとで，最適な決定関数を導出することを特徴とすることを示した．

第 3 章では単語予測問題に統計的決定理論のアプローチを適用した．具体的には，N グラムモデルを確率モデルとし真の次数と確率パラメータの双方を未知とした場合において，N グラムモデルの確率パラメータ $\boldsymbol{\theta}$ の事前分布と，N グラムモデルの次数 m のモデルの事前確率を利用し，予測誤り率に関するベイズ基準のもとで最適な単語予測法を導出した．

この単語予測法は，与えられた学習データ $(x^{N-1}, y)^n$ に対して，次数 m ごとのパラメータの事後分布の積分により求められる予測分布を，モデルの事後確率により重み付けする形式となっている．この形式は従来の式 (3.1) で表現される重み付け

法と類似しているが，本研究により $p(y|x_p^{m-1}, \boldsymbol{\theta}_m)$ を m の予測分布，$w(m)$ を m の事後確率としたものが，単語の予測誤り率を最適にする単語予測法であることが明らかになった．さらに，単語の発生する分布に条件付き多項分布，そのパラメータの事前分布にディレクレ分布を仮定した場合の単語予測法を提案した．この単語予測法は，予測誤り率に関するベイズ基準に対して最適解が閉じた形式で求まり，学習データの追加に対して加算増分により更新できるため，オンライン処理にも適用可能である点において計算量的な優位性を有したアルゴリズムであることを示した．また，アルゴリズムの具体的な実現方法の説明をもとにアルゴリズムに要するメモリー量および計算量を考察し，従来法とほぼ同等であることを示した．

また，予測式に含まれるパラメータの算出方法の，理論的な意味が明確でないニーザー・ナイ法について，予測式の形式の比較を行い理論的な観点で解釈を与えた．ここでは，重み付けのパラメータは $m-1$ 次と m 次のモデルの事後確率の比を近似したものと解釈できることを示した．

最後に，実データである文書データによる単語予測実験を行った．単語予測の正答率の点で，基盤的なモデルの中では性能が高いことが知られている修正ニーザー・ナイ法と，ほぼ同等の単語予測精度であることを示した．

第 4 章では単語予測の分野適合問題に統計的決定理論のアプローチを適用した．確率モデルとして，当該データと類似データの条件付き多項分布のパラメータが確率的に変化する数理モデルを仮定し，予測誤り率に関するベイズ基準のもとで最適な単語予測法を導出した．さらに，統計的言語処理の分野にて，一般的に知られる分野適応の従来法に対して，導出された予測法との比較を行い理論的な観点で解釈を与えた．ここでは，従来法は導出された予測法の近似であり，実用的にはほぼ差がないことを示した．

最後に，提案法にて類似データによる事後分布にパラメータを変化させた確率モデルを仮定し，その変化させたパラメータを事前分布として利用した単語予測実験を行い，学習データの量が少ない場合に，単語予測の正答率が向上することを示した．また，学習データ量を増加させた場合の実験を行い，提案するアプローチにより求められた単語予測法が理論的な解析により漸近的な一致性を持つことが示されていたことに対して，この結果を支持する結果となることを示した．

5.2 今後の発展

統計的決定理論のアプローチは，確率モデルと決定関数を明確に区別して議論を行うという特徴を持つ．そのため，対応可能な確率モデルは N グラムモデルに限定したものではない．今後の発展としては，統計的言語処理の分野にて提案されている，単語の品詞も利用した形式や，クラス N グラムモデルといった決定関数に相当するものを，提案するアプローチにて再検討することが可能である．

これを統計的決定理論のアプローチにて整理すると，以下のようになる．まず，統計的言語処理の研究において言語モデルとして提案されている統計的決定理論における決定関数の様々な形式に対して，確率モデルとして書き直すことを行う．その確率モデルを仮定し，単語予測を含めそれぞれの目的に適した損失関数で表現する．最後に，統計的決定理論のアプローチに基づいた基準のもとで，最適となる決定関数を導出する．

これは，今までの統計的言語処理で提案されてきた様々な言語モデルの従来研究と対峙するものではなく，双方の成果を生かしていくことのできる新しいアプローチを統計的言語処理の分野に付け加えることになったと考えられる．

謝辞

本論文をまとめるにあたり主査として御指導頂いた，早稲田大学基幹理工学部応用数理学科 松嶋敏泰教授に感謝いたします．松嶋研究室に配属され，卒業後も社会人博士課程の学生として快く引き受けていただき数え切れないほど多くの有益な御指導，御助言を承りました。

また，副査として大変貴重なお時間をいただき御指導頂いた，早稲田大学基幹理工学部応用数理学科 大石進一教授，早稲田大学基幹理工学部応用数理学科 匂坂芳典教授，早稲田大学理工学術院総合研究所 平澤茂一名誉教授に感謝いたします．

松嶋研究室の諸先輩方には，多くの御指導，御助言を頂き感謝いたします．特に，横浜商科大学 浮田善文教授には学部時代より常に親身なアドバイスを頂きました．さらに，早稲田大学メディアネットワークセンター 須子統太博士，早稲田大学メディアネットワークセンター 堀井俊佑博士，株式会社 NTT ドコモ 桑田修平博士，早稲田大学理工学研究科博士後期課程 安田豪毅氏，には研究に関する多くの議論や助言を頂きました．また，早稲田大学理工学研究科博士後期課程 宮希望氏をはじめとする松嶋研究室に所属する学生一同には，審査に関わる諸々の作業について多くのご支援をいただき感謝いたします．

さらに，私の上司として御指導を頂いた（株）NTT データの中村太一博士（現東京工科大学），関根純博士（現専修大学），上島康司氏，城塚音也氏，松永努博士，石打智美氏（現 NTT 知的財産センタ），坂野鋭博士（現 NTT コミュニケーション科学基礎研究所），中川慶一郎博士，高木徹博士，松本良平氏，に感謝いたします．特に，関根純博士の強い後押しのもと，この機会をいただけたことに深く感謝いたします．また，同僚として日ごろから有益な議論を頂いた（株）NTT データの同僚諸氏，特に北内啓氏（現システマジック），山中啓之氏，鈴木賢一郎氏，高橋彰子氏，佐藤新氏，東陽子氏，佐治美歩氏，米森力氏，矢野順子氏，野村雄司氏，大木

環美氏に感謝いたします.

本論文は以上をはじめとする，多くの方々の御指導，御支援の賜物です．お世話になった方々に心より御礼申し上げます.

最後に，健康面，精神面の支えになってくれた父近志，母博子，妻弘恵，息子佑司に感謝いたします.

2014 年 2 月

末永　高志

付 録A　確率変数の変換

A.1　確率変数変換の定理

確率密度関数 $f_x(x_1, \cdots, x_n)$ を持つ連続値の n 次元確率ベクトル $\boldsymbol{X} = (X_1, \cdots, X_n)$ に対して，$(X_1, \cdots, X_n)$ から別の n 次元確率ベクトル $\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ への 1 対 1 変換を $\boldsymbol{\phi}$，その逆変換を $\boldsymbol{\psi}$ とする．このとき，$\boldsymbol{Y} = (Y_1, \cdots, Y_n)$ の確率密度関数 $f_y(y_1, \cdots, y_n)$ は

$$f_y(y_1, \cdots, y_n) = f_x(x_1, \cdots, x_n)|J|, (\boldsymbol{x} = \boldsymbol{\psi}(\boldsymbol{y})) \tag{A.1}$$

となる [61, 13]．ここで，$J = J(\boldsymbol{\psi} : \boldsymbol{y})$ は変換 $\boldsymbol{\psi}$ に対応する関数行列式（ヤコビアン）

$$J = \frac{\partial(x_1, \cdots, x_n)}{\partial(y_1, \cdots, y_n)} = \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{y}} \tag{A.2}$$

を表し，各点 $\boldsymbol{y}$ で $J \neq 0$ とする．

A.2　ガンマ分布に従う変数とベータ分布に従う変数の積

X_1 を確率密度関数

$$f(x_1) = \frac{1}{\Gamma(\alpha)} x_1^{\alpha-1} \mathrm{e}^{-x_1} \tag{A.3}$$

とするガンマ分布に従う確率変数，X_2 を確率密度関数

$$f(x_2) = \frac{\Gamma(\alpha)}{\Gamma(\rho\alpha)\Gamma((1-\rho)\alpha)} x_2^{\rho\alpha-1} (1 - x_2)^{(1-\rho)\alpha-1} \tag{A.4}$$

とするベータ分布に従う確率変数とする．これらの変数に対して，$Y_1 = X_1X_2$，$Y_2 = X_1(1 - X2)$ という変数変換 $\boldsymbol{\phi}$ を設定すると，その逆変換である $\boldsymbol{\psi}$ は $X_1 = Y_1 + Y_2$，$X_2 = \frac{Y_1}{Y_1+Y_2}$ となる．

$\boldsymbol{\psi}$ に対する関数行列式を求めると，

$$J = \begin{vmatrix} \frac{\partial X_1}{\partial Y_1} & \frac{\partial X_1}{\partial Y_2} \\ \frac{\partial X_2}{\partial Y_1} & \frac{\partial X_2}{\partial Y_2} \end{vmatrix} \tag{A.5}$$

$$= -\frac{1}{Y_1 + Y_2} \tag{A.6}$$

となる．

従って，式 (A.1) により，

$$f(y_1, y_2) = f(x_1, x_2)|J| \tag{A.7}$$

$$= \frac{1}{\Gamma(\alpha)} x_1^{\alpha-1} \mathrm{e}^{-x_1} \frac{\Gamma(\alpha)}{\Gamma(\rho\alpha)\Gamma((1-\rho)\alpha)} x_2^{\rho\alpha-1}(1-x_2)^{(1-\rho)\alpha-1}|J| \tag{A.8}$$

$$= \frac{1}{\Gamma(\alpha)} x_1^{\alpha-1}(x_1x_2)^{\rho\alpha-1}\mathrm{e}^{-x_1x_2}\{x_1(1-x_2)\}^{(1-\rho)\alpha-1}\mathrm{e}^{-x_1(1-x_2)} \tag{A.9}$$

$$= \frac{1}{\Gamma(\rho\alpha)\Gamma((1-\rho)\alpha)} y_1^{\rho\alpha-1}\mathrm{e}^{-y_1} y_2^{1-\rho)\alpha-1}\mathrm{e}^{-y_2} \tag{A.10}$$

となる．

求めたいのは Y_1 の関数であり，

$$f(y_1) = \int_{y_2=0}^{\infty} f(y_1, y_2)dy_2 \tag{A.11}$$

$$= \frac{1}{\Gamma(\rho\alpha)\Gamma((1-\rho)\alpha)} y_1^{\rho\alpha-1}\mathrm{e}^{-y_1} \int_0^{\infty} y_2^{1-\rho)\alpha-1}\mathrm{e}^{-y_2}dy_2 \tag{A.12}$$

$$= \frac{1}{\Gamma(\rho\alpha)\Gamma((1-\rho)\alpha)} y_1^{\rho\alpha-1}\mathrm{e}^{-y_1}\Gamma((1-\rho)\alpha) \tag{A.13}$$

$$= \frac{1}{\Gamma(\rho\alpha)} y_1^{\rho\alpha-1}\mathrm{e}^{-y_1} \tag{A.14}$$

となる．これは，パラメータを $\rho\alpha$ とするガンマ分布を意味する．

参考文献

[1] Shunichi Amari. A theory of adaptive pattern classifiers. *IEEE Transaction on Electronic Computers*, Vol. EC-16, No. 3, pp. 299–307, 1967.

[2] Michiel Bacchiani and Brian Roark. Unsupervised language model adaptation. In *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP '03)*, pp. 224–227, 2003.

[3] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1980.

[4] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, 2000.

[5] Christopher M. Bishop. 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.

[6] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, Vol. 18, No. 4, pp. 467–479, 1992.

[7] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of ACL*, pp. 310–318, 1996.

[8] K. W. Church and W. A. Gale. A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of english bigrams. *Computer Speech and Language*, Vol. 5, No. 5, pp. 19–54, 1991.

[9] T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley, 1991.

[10] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1, pp. 1–38, 1977.

[11] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[12] Andrew C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1989.

[13] Robert V. Hogg, Joseph W. McKean, and Allen T. Craig. *Introduction to Mathematical Statistics*. Pearson, 2004.

[14] Frederick Jelinek and Robert L. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pp. 381–397, 1980.

[15] Frederick Jelinek, Robert L. Mercer, and Salim Rouks. *Principles of Lexical Language Modeling for Speech Recognition*. Dekker Publishers, New York, 1991.

[16] Biing-Hwang Juang and Shigeru Katagiri. Discriminative learning for minumum error classification. *IEEE Transaction on Signal Processing*, Vol. 40, No. 12, pp. 3043–3054, 1992.

[17] Daniel Jurafsky and James H. Martin. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2008.

[18] Shigeru Katagiri, Chin-Hui Lee, and Biing-Hwang Juang. A generalized decent method. 日本音響学会講演論文集, pp. 141–142. 日本音響学会, 1990.

[19] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *Proceedings of ICASSP*, Vol. 1, pp. 181–184. Association for Computational Linguistics Morristown, NJ, USA, 1995.

[20] R. Lau, R. Rosenfeld, and S. Roukos. Trigger-based language models: a maximum entropy approach. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 45–48, 1993.

[21] David J. C. MacKay and Linda C. Bauman Peto. A hierarchical dirichlet language model. *Natural Language Engineering*, Vol. 1, No. 3, pp. 289–307, 1995.

[22] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

[23] Toshiyasu Matsushima, Hiroshige Inazumi, and Shigeichi Hirasawa. A class of distortionless codes designed by bayes decision theory. *IEEE Transactions on Information Theory*, Vol. 37, No. 5, pp. 1288–1293, 1991.

[24] Kuhn R. and Renato de Mori. A cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 6, pp. 570–583, 1990.

[25] D. Ron, Y. Singer, and N. Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, Vol. 25, pp. 117–149, 1996.

[26] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, Vol. 27, pp. 379–423, 623–656,, 1948.

[27] J. Q. Smith. A generalization of the bayesian steady forecasting model. *Journal of the Royal Statistical Society, Series B*, Vol. 41, pp. 375–387, 1979.

[28] Yee Whye Teh. A bayesian interpretation of interpolated kneser-ney. Technical report, NUS School of Computing Technical Report, 2006.

[29] Yee Whye Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of COLING/ACL 2006*, pp. 985–992, 2006.

[30] Yee Whye Teh, Jordanm Michael I., Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, Vol. 101, pp. 1566–1581, 2006.

[31] 下平英寿, 久保川達也, 竹内啓, 伊藤秀一. 統計科学のフロンティア 3 モデル選択–予測・検定・推定の交差点. 岩波書店, 2004.

[32] 塚本健一朗, 松嶋敏泰, 平澤茂一. 適応型 arq における制御パラメータ決定方式について. 電子情報通信学会論文誌, Vol. J81-B-1, No. 6, pp. 391–400, 1998.

[33] 石井健一郎, 上田修功, 前田英作, 村瀬洋. わかりやすいパターン認識. オーム社, 1998.

[34] 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.

[35] 北研二, 中村哲, 永田昌明. 確率・統計モデルの音声言語処理への応用. 人工知能学会誌, Vol. 10, No. 2, pp. 189–196, 1995.

[36] 伊庭幸人, 種村正美, 大森裕浩, 和合肇, 大森裕浩, 佐藤整尚, 高橋明彦. 統計科学のフロンティア 12 計算統計 II–マルコフ連鎖モンテカルロ法とその周辺. 岩波書店, 2005.

[37] 前田康成, 小原永. 統計的決定理論に基づく電報分類方法に関する一考察. 情報処理学会論文誌, Vol. 43, No. 10, pp. 3119–3126, 2002.

[38] 前田康成, 吉田秀樹, 鈴木正清, 松嶋敏泰. 学習データが少量しかない場合の文書分類方法に関する一考察. 電気学会論文誌 C, Vol. 131, No. 8, pp. 1459–1466, 2011.

[39] 政瀧浩和, 匂坂芳典, 久木和也, 河原達也. 最大事後確率推定による n-gram 言語モデルのタスク適応. 電子情報通信学会論文誌 D-II, Vol. J81-D-II, No. 11, pp. 2519–2525, 1998.

[40] 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, pp. 6–9. 言語処理学会, 2012.

[41] 川端豪, 田本真詞. 二項事後分布に基づく n-gram 言語モデルの back-off 平滑化. 情報処理学会研究報告, Vol. 1995-SP-95, pp. 87–92, 1995.

[42] 繁桝算男. ベイズ統計入門. 東京大学出版会, 1985.

[43] 片桐滋. 一般化確率的降下法の展開. 日本音響学会誌, Vol. 55, No. 8, pp. 563–568, 1999.

[44] 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, pp. 1095–1098. 言語処理学会, 2011.

[45] 桑田修平, 前田康成, 松嶋敏泰, 平澤茂一. 推薦システムのための状態遷移確率の構造を未知としたマルコフ決定過程. 情報処理学会論文誌. 数理モデル化と応用, Vol. 6, No. 1, pp. 20–30, 2013.

[46] 尾崎俊治. 確率モデル入門. 朝倉書店, 1996.

[47] 森信介. クラスに基づく可変長記憶マルコフモデル. 情報処理学会論文誌, Vol. 43, No. 1, pp. 34–43, 2002.

[48] 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.

[49] 渡部晋治, 堀貴明. ベイズアプローチによる n-gram 言語モデリング. 日本音響学会講演論文集, pp. 79–80. 日本音響学会, 2003.

[50] 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. PhD thesis, 早稲田大学大学院理工学研究科, 2000.

[51] 中川聖一, 赤松裕隆, 西崎博光. 音声認識用言語モデルのためのタスク適応化と定型表現の利用. 自然言語処理, Vol. 6, No. 2, pp. 97–115, 1999.

[52] 蓑谷千凰. すぐに役立つ統計分布. 東京図書, 1998.

[53] 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.

[54] 東京大学教養学部統計学教室編. 統計学入門. 東京大学出版会, 1991.

[55] 東京大学教養学部統計学教室編. 自然科学の統計学. 東京大学出版会, 1992.

[56] 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.

[57] 佐藤敦. パターン認識問題の数理. 電子情報通信学会 基礎・境界ソサイェティ Fundamental Review, Vol. 5, No. 4, pp. 302–311, 2012.

[58] 神嶌敏弘. 転移学習. 人工知能学会誌, Vol. 25, No. 4, pp. 572–580, 2010.

[59] 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.

[60] 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.

[61] 鈴木武, 山田作太郎. 数理統計学–基礎から学ぶデータ解析–. 内田老鶴圃, 1996.

[62] 匂坂芳典. 音声認識のための普遍的言語制約モデルを目指して. 白井克彦（編）, 情報システムとヒューマンインタフェース, pp. 91–100. 早稲田大学出版, 2010.

[63] 松原望. 入門ベイズ統計–意思決定の理論と発展. 東京図書, 2008.

[64] 金明哲, 村上征勝, 永田昌明, 大津起夫, 山西健司. 統計科学のフロンティア 10 言語と心理の統計–ことばと行動の確率モデルによる分析. 岩波書店, 2003.

[65] 永田靖. 入門統計解析法. 日科技連, 1992.

[66] 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, pp. 1107–1110. 言語処理学会, 2011.

研究業績

種類別	題名，発表・発行掲載誌名，発表・発行年月，連名者（申請者含む）
1. ○論文	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 情報処理学会論文誌数理モデル化と応用 Vol. 6，No. 1，pp. 102-110,（2013-3） 末永高志，松嶋敏泰
2. 論文	単語の重要度評価基準の検討と医療関連文書への適用評価 情報処理学会論文誌数理モデル化と応用 Vol.3，No. 2，pp. 108-118,（2010-3） 末永高志，松永務，関根純，村松正明
3. 論文	A Framework for Business Data Analysis IEEE International Conference on e-Business Engineering, pp. 703-708, (2008-10) Takashi Suenaga, Shoko Takahashi, Miho Saji, Junko Yano, Keiichiro Nakagawa, Jun Sekine
4. 論文	業務データ分析のためのデータ分析フレームワークの開発 情処論文誌データベース，Vol. 1，No. 2，pp. 15-25,（2008-9） 末永高志，山中啓之，高橋彰子，東陽子，佐治美歩，矢野順子，中川慶一郎，関根純
5. 論文	Cluster Discriminant analysis for feature space visualization Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies, pp.146-150, (2002-9） Takashi Suenaga, Arata Sato, Hitoshi Sakano

種類別	題名，発表・発行掲載誌名，発表・発行年月，連名者（申請者含む）
6. 論文	クラスタ構造に着目した特徴空間の可視化－クラスタ判別法－ 電子情報通信学会論文誌 D-II，Vol.J85-No.5，pp.785-795,（2002-5） 末永高志，佐藤新，坂野鋭
7. ○講演	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 情報処理学会研究報告，MPS-90,（2012-9） 末永高志，松嶋敏泰
8. ○講演	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,（2012-3） 末永高志，松嶋敏泰
9. 講演	テキストマイニングのためのドメイン別単語辞書の構築方法 情報処理学会研究報告，MPS-76,（2009-12） 末永高志，松永務，関根純，村松正明
10. 講演	顔画像検出におけるデータ採取地の影響について 電子情報通信学会総合大会講演論文集,（2003-3） 末永高志，坂野鋭，松永務
11. 講演	顔画像認識におけるデータ採取地の影響について 電子情報通信学会技術研究報告，PRMU2002-147，pp.7-12（2002-12） 末永高志，坂野鋭，松永務
12. 講演	クラスタ判別法による文字特徴データ解析 電子情報通信学会総合大会講演論文集,（2002-3） 末永高志，佐藤新，坂野鋭
13. 講演	クラスタ判別法による顔画像データ解析 電子情報通信学会ソサイエティ大会講演論文集,（2001-3） 末永高志，佐藤新，坂野鋭

種類別	題名，発表・発行掲載誌名，発表・発行年月，連名者（申請者含む）
14. 講演	分布の構造に着目した特徴空間の可視化–クラスタ判別法– 電子情報通信学会技術報告，PRMU-2001-44,（2001-7） 末永高志，佐藤新， 坂野鋭
15. その他（論文）	ビジネス・インテリジェンス・システムにおける情報要求の抽出手法 情報処理学会論文誌，Vol. 50，No. 12，pp. 2990-3000,（2009-12） 関根純，末永高志，矢野順子，中川慶一郎，山本修一郎
16. その他（論文）	外部ソースを活用したウェブ・マーケティングのための分析フレームワークの提案 オペレーションズ・リサーチ，Vol. 53，No. 2,（2008-2） 矢野順子， 加藤元英， 末永高志， 生田目崇
17. その他（総説）	高次元データの可視化技術 画像電子学会誌，Vol. 32，No. 3，pp. 251-257,（2003-5） 坂野鋭，末永高志
18. その他（講演）	相補的な素性選択基準の関係を考慮した文書分類のための素性選択方式 情報処理学会研究報告，MPS-73，No. 9,（2009-3） 末永高志，松永努，関根純
19. その他（講演）	Web アクセスログデータの系列情報を利用したサービスの関連性の分析 電子情報通信学会技術研究報告，PRMU2005-17，pp.25-28,（2005-6） 末永高志，岡田崇，石打智美
20. その他（講演）	アクセス履歴に基づく Web ページ利用傾向の可視化法 電子情報通信学会総合大会講演論文集,（2005-3） 岡田崇，末永高志，石打智美

種類別	題名，発表・発行掲載誌名，発表・発行年月，連名者（申請者含む）
21. その他 （講演）	分析シナリオに注目したアイテム分析システムの提案 教育システム情報学会研究報告，Vol. 19，No. 1，pp. 77-82，(2004-5) 末永高志， 大内学， 石打智美
22. その他 （講演）	テスト結果を用いた多面的分析方法の提案 電子情報通信学会技術研究報告，ET-103，No. 368，pp.37-40，(2003-10) 大内学，末永高志，石打智美
23. その他 （講演）	部分空間比較による変量選択法 電子情報通信学会技術研究報告，PRMU-103，No. 295，(2003-9) 米森力，末永高志，原正巳，松永務
24. その他 （講演）	クラスタ判別法の医療データ解析への応用 知識ベースシステム研究会，Vol. 54，pp. 237-242，(2001-11) 佐藤新，末永高志，坂野鋭
25. その他 （講演）	トレリス符号を用いた有歪みデータ圧縮の一考察 電子情報通信学会技術研究報告，IT97-33，(1997-7) 末永高志，松嶋敏泰，平澤茂一