早稲田大学大学院 基幹理工学研究科

博士論文審査報告書

論　文　題　目

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

申　請　者

末永	高志
Takashi	SUENAGA

数学応用数理専攻　情報理論研究

2014 年 2 月

統計的決定理論のアプローチは，データ解析を始め，情報理論における歪みのない情報源圧縮，誤り訂正符号，文書分類，推薦システムといった様々な分野で適用例が報告されている．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，近年，大量の言語データを学習データとして用い，数理的な手法をもとに処理法を求めるアプローチが行われており，その実証的な有効性が知られている．しかし，自然言語処理における学習データの取り扱いにおいて，統計的決定理論のアプローチを用いた研究はあまり見受けられない．

本研究では，自然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，この問題を理論的に解析することを目的とする．具体的には，

(1) 単語予測問題の汎用的な基盤モデルである N グラムモデル（$N-1$ 次マルコフモデル）を確率モデルと仮定し，予測誤り率に関してベイズ基準のもとで最適な予測法（決定関数）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自然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

を行っている．

単語予測問題とは，単語 $x_i \in X$ の $N-1$ 個を連接した $x^{N-1} = x_1 x_2 \cdots x_{N-1}$ に後続する単語 $y \in Y$ を予測する問題である．x^{N-1} のもとで y の発生する確率を，確率パラメータ $\boldsymbol{\theta}$ でパラメタライズされた条件付き多項分布 $P(y|x^{N-1}, \boldsymbol{\theta})$ で表現する．ただし，$N=1$ の場合，x^{N-1} は空列 ϕ とする．

第 1 章は序論で，第 2 章では，自然言語処理におけるこの問題に関する従来研究とその一般的アプローチ法に対する，統計的決定理論のアプローチ法との違いを述べ，本研究の位置づけを明らかにしている．自然言語処理のこの問題に対する一般的アプローチでは，$\boldsymbol{\theta}$ は既知と考え，まず予測法として，$D_1(x^{N-1}, \boldsymbol{\theta}) = \arg\max_y P(y|x^{N-1}, \boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$ とする方法で，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．

また，予測法を確定するためには，次数 N の決定も必要となる．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の階層モデル族で，高次のモデルの方がより広いクラスの確率モデルが表現可能となる．しかし，高次のモデルは低次のモデルよりもパラメータ数が多いため，各パラメータ推定に用いられるサンプル数が相対的に少なくなり，推定誤差は大きくなる傾向にある．そのような性質を考慮した，パラメータの推定方法が必要となってくる．

自然言語処理の分野においては，このような問題に対して様々なアプローチが行われている．例えば，補完法と呼ばれる方法では，低次のモデルから高次のモデルまでを，ある重みで加算する予測式が用いられ，重みのパラメータは EM アルゴリズム等を用いてチューニングされている．

この他に，実証的な観点から有効性が知られているニーザー・ナイ法と呼ばれる以下の予測式も提案されている．

$$D_2(x^m, \boldsymbol{\theta}) = \frac{\max\{c(y|x^m) - d_{\text{discount}}, 0\}}{\sum_{y \in Y} c(y|x^m)} + \frac{d_{\text{discount}}}{\sum_{y \in Y} c(y|x^m)} \left|\{y : c(y|x^m) > 0\}\right| D_2(x^{m-1}, \boldsymbol{\theta}) \quad (1)$$

ここで $c(y|x^m)$ は学習データに含まれる x^m に後続する y の出現頻度である．しかしなが

ら，この式に含まれるパラメータの算出法等の理論的な意味は明らかでない．

その他，N グラムモデルを基盤に自然言語の特性を制約として組み込んだ，クラス N グラムモデル，可変長 N グラムモデルなど様々な形式による方法論が提案され，実証的な観点で優れた性能をもつことが知られており，自然言語のデータに適した様々な関数モデルが蓄積されている．

本研究では，これらとは違うアプローチである，統計的決定理論のアプローチを単語予測問題に適用している．統計的決定理論のアプローチでは，データが発生する分布に一般的にはパラメトリックな確率モデルを仮定し，一部のパラメータを未知とする．また，これとは別に意思決定を行うための決定関数を定義する．そのもとで，損失関数や危険関数を設定する．さらに，設定した危険関数に対して，ミニマックス基準やベイズ基準などのもとで最適な決定関数を導出することを主要な課題としている．

第 3 章では，統計的決定理論のアプローチの単語予測問題への適用について述べている．まず，N グラムモデルを確率モデルとし，その次数とパラメータの双方を未知と仮定する．次に，決定関数は，単語列 x_p^{N-1} に後続する単語 y_p を予測する関数で，n 対の学習データ $(x^{N-1}, y)^n$ と x_p^{N-1} を引数とする関数 $D(x_p^{N-1}, (x^{N-1}, y)^n)$ として定義される．

予測誤り率による損失として，損失関数

$$L(D(x_p^{N-1}, (x^{N-1}, y)^n), Y|\boldsymbol{\theta}_m) = \sum_{y_p \in Y} d(D, y_p) p(y_p|x_p^{N-1}, \boldsymbol{\theta}_m) \tag{2}$$

を設定する．ただし，$\boldsymbol{\theta}_m$ は次数 m のモデルにおける条件付き多項分布のパラメータ，$d(D, y_p)$ は $D = y_p$ なら 0，$D \neq y_p$ なら 1 を返す関数である．ベイズ基準のもとで最適な予測法は，以下のように導出される．

$$D(x_p^{N-1}, (x^{N-1}, y)^n) = \\ \arg\max_y \sum_{m=1}^{N} p(m|(x^{N-1}, y)^n) \int_{\boldsymbol{\theta}_m} p(y|x_p^{N-1}, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|(x^{N-1}, y)^n) d\boldsymbol{\theta}_m \tag{3}$$

これは，各々の次数のモデルの事後確率でその次数のモデルの予測分布を重み付け加算した値を求め，それを最大とする y が，N の次数決定の課題を解決する理論的に最適な方法であることを示している．

また，式 (3) は積分計算が含まれているが，そのパラメータ $\boldsymbol{\theta}_m$ の事前分布 $f(\boldsymbol{\theta}_m)$ に自然共役分布であるディレクレ分布を仮定することで，

$$\int_{\boldsymbol{\theta}_m} p(y|x_p^{N-1}, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|(x^{N-1}, y)^n) d\boldsymbol{\theta}_m = \frac{c(y|x_p^{m-1}) + \alpha(y|x_p^{m-1})}{\sum_{y \in Y} c(y|x_p^{m-1}) + \sum_{y \in Y} \alpha(y|x_p^{m-1})} \tag{4}$$

により算術的に容易に求められる．ただし，$\alpha(y|x^{m-1})$ は，ディレクレ分布のパラメータ，$c(y|x^{m-1})$ は学習データに含まれる x^{m-1} に後続する y の出現頻度である．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ため，オンライン処理にも適用可能で，計算量的な優位性を有した予測法である．

また，式 (3) を漸化式の形で表現すると，従来法の式 (1) と類似の式が得られる．この両式を比較することで，理論的な意味が不明であったニーザー・ナイ法の解釈が可能となった．例えば，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．

また，実データによる提案法の単語予測実験を行い，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が示されている．

第 4 章では，本アプローチの単語予測における分野適応問題への適用について述べている．予測を行いたい文章と同じ母集団のデータ（以下，当該データと呼ぶ）が少量しか得られない場合，類似する分野で大量に入手可能なデータ（以下，類似データと呼ぶ）を活用し，

予測精度を向上することが考えられる．ここでの課題は，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から発生するとは限らないことである．分野適応問題とは，n_t 個の当該データの対 $(x^{N-1},y)^{n_t}$ と n_f 個の類似データの対 $(x^{N-1},y)^{n_f}$，および単語列 x_p^{N-1} が与えられたもとで，両学習データを有効に用いた予測の決定関数 $D_{\mathrm{da}}(x_p^{N-1},(x^{N-1},y)^{n_t},(x^{N-1},y)^{n_f})$ を求める問題と定義することができる．なお，以下では，$c_t(y|x^{N-1})$ と $c_f(y|x^{N-1})$ はそれぞれ当該データと類似データに含まれる x^{N-1} に後続する y の出現頻度とする．

この問題に対して，統計的決定理論のアプローチに基づき，類似データと当該データの双方が条件付き多項分布で発生し，そのパラメータが確率的に変化する数理モデルを仮定し，予測誤差を評価基準として設定したもとで，最適な予測法を導出している．まず，類似データと当該データの多項分布のパラメータを $\theta_{f,y}$ と $\theta_{t,y}$ とし，それぞれは $\theta_{f,y}=\frac{A_{f,y}}{\sum_{y\in Y}A_{f,y}},\theta_{t,y}=\frac{A_{t,y}}{\sum_{y\in Y}A_{t,y}}$ と表現されるとする．ここで，$A_{f,y}$，$A_{t,y}$ は $A_{t,y}=A_{f,y}U_{f,y}$ により確率的に変化すると仮定する．ただし，$U_{f,y}$ は，$f(U_{f,y})=\frac{\Gamma(\rho\alpha'_{f,y}+(1-\rho)\alpha'_{f,y})}{\Gamma(\rho\alpha'_{f,y})\Gamma((1-\rho)\alpha'_{f,y})}U_{f,y}^{\rho\alpha'_{f,y}-1}(1-U_{f,y})^{(1-\rho\alpha'_{f,y})-1}$ で表現されるベータ分布に従う確率変数で，$\rho\in(0,1)$，$\alpha'_{f,y}=c_f(y|x^{N-1})+\alpha_{f,y}$ である．また，$A_{f,y}$ の事前分布としては，$f(A_{f,y})=\frac{1}{\Gamma(\alpha_{f,y})}A_{f,y}^{\alpha_{f,y}-1}\mathrm{e}^{-A_{f,y}}$ で表現されるガンマ分布を仮定している．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 と呼ばれる確率モデルに対応している．

このような確率モデルを仮定し，損失関数を式 (2) とした場合の，ベイズ基準のもとで導出される最適予測法は以下となる．

$$
\begin{aligned}
&D_{\mathrm{da}}(x_p^{N-1},(x^{N-1},y)^{n_t},(x^{N-1},y)^{n_f})=\\
&\arg\max_y\sum_{m=1}^{N}p(m|(x^{N-1},y)^{n_t})\\
&\qquad\frac{c_t(y|x_p^{m-1})+\rho\{c_f(y|x_p^{m-1})+\alpha_f(y|x_p^{m-1})\}}{\sum_{y\in Y}c_t(y|x_p^{m-1})+\rho\{\sum_{y\in Y}c_f(y|x_p^{m-1})+\sum_{y\in Y}\alpha_f(y|x_p^{m-1})\}}
\end{aligned}
\tag{5}
$$

さらに，この予測法は単語予測問題と同様のアルゴリズムで実装可能であること，N グラムモデルを基盤とした同様の形式をもった従来法に対する理論的な解釈が与えられること，実データによる実験を行い従来法と同等の予測精度であることなどが示されている．

第 5 章では，結論と今後の発展について述べている．今後の発展としては，本論文では N グラムモデルを確率モデルと仮定したが，本アプローチ法は従来研究で蓄積された様々な言語モデルにも適用可能であり，さらに単語予測問題以外への応用も考えられることが述べられている．

以上を総括すると，本論文は，自然言語処理の代表的な問題である単語予測問題とその応用である分野適応問題に対して，統計的決定理論のアプローチを適用したことが特長で，最適な予測法の導出や従来法の理論的な新しい解釈を与える等，自然言語処理研究に新たなアプローチを付け加えその有効性を示していることは高く評価できる．よって, 本論文は博士 (工学) の学位論文として価値あるものと認める.

2014 年 1 月

審査員（主査）	早稲田大学教授	博士（工学）	（早稲田大学）	松嶋 敏泰
	早稲田大学教授	工学博士	（早稲田大学）	大石 進一
	早稲田大学教授	工学博士	（早稲田大学）	匂坂 芳典
	早稲田大学名誉教授	工学博士	（大阪大学）	平澤 茂一