早稲田大学大学院 基幹理工学研究科

博 士 論 文 概 要

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

2013 年 5 月

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がある．この研究では，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の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，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．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を課題とするが，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．

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実用的な観点にて優位性が知られているニーザー・ナイ法と呼ばれる予測法に対して，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，類似データをもとに当該データを予測するための，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（SP M）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．これは，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において，学習データとして用いる当該データの増加に従い，双方で求められた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．

早稲田大学　博士（工学）　学位申請　研究業績書

氏 名　末永　高志　　印

（2014 年　1 月　現在）

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
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

早稲田大学　博士（工学）　学位申請　研究業績書

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）
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
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
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

早稲田大学　博士（工学）　学位申請　研究業績書

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）
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