

統計的テキストモデル

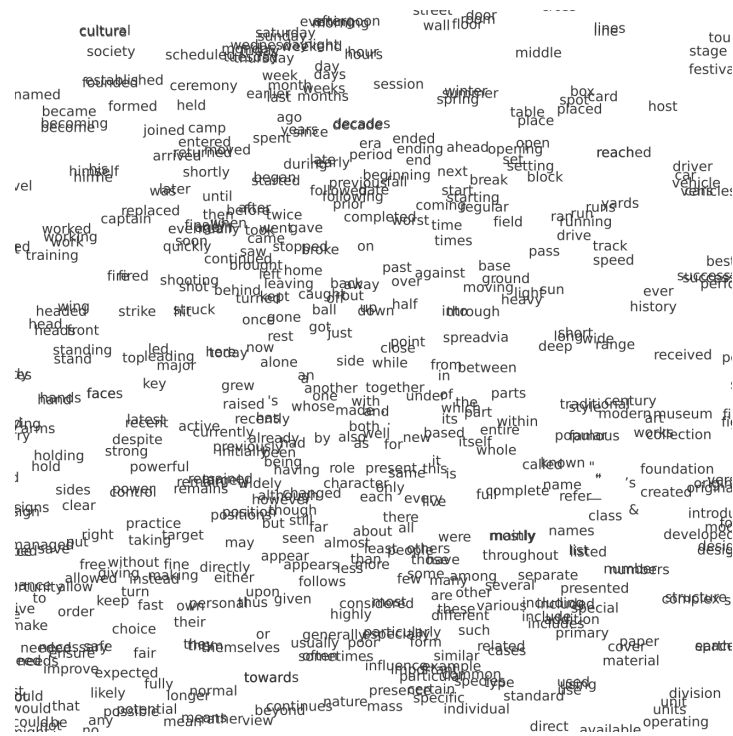
持橋 大地

統計数理研究所 数理・推論研究系

daichi@ism.ac.jp

2024年4月25日

\$Id: textmodel.tex,v 1.18 2024/04/04 11:20:40 daichi Exp \$



はじめに

本書は、テキストの統計的なモデル化について解説した本です。ここでいうテキストとは、Web ページやメールのような文書だけでなく、小説や新聞、法案、アンケートへの自由回答など、さまざまな範囲を含んでいます。よって本書の読者としては、テキストを扱う必要のある理系のエンジニアの方だけでなく、人文科学および社会科学系の方々、および言語に興味のある一般の方々を想定しています。

Web が現れて以来、電子化されたテキストを扱う自然言語処理は驚異的に進展しました。しかし、自然言語処理の教科書や情報源は、文書のカテゴリや単語の品詞のように、テキストから人の与えた正解ラベルを予測する教師あり学習を扱っている場合がほとんどで、現実に現れる、そうした正解ラベルのないテキストを統計的にどうモデル化して扱うか、という**教師なし学習**について体系的にまとめられた成書は、ほとんど存在しないのが現状です。「テキストマイニング」はこれに近い分野ですが、多くの場合、表面的なパッケージの使い方に終わっていたり、離散的な言葉の頻度に、従来の多変量解析を無理にあてはめていたりといった問題がありました。テキストのような離散データを取り扱うには、より適切な統計モデルが存在します。またその際に、内部でどのような数学的なモデル化と計算が行われているのかがわからなければ、目の前の問題に適用するためには、どこにどう手を入れたらよいのかを知ることはできません。

そこで本書では、テキストの統計的なモデル化について一から説明し、**ブラックボックスに頼らなくても、様々な分析を自分で自由に行えるようになること**を目的としています。本書で説明するようなテキストの高度な統計モデルにはパッケージがないことも多く、また、パッケージでブラックボックス化して扱うことは適切とはいえません。あえて「背景知識」のような無味乾燥な章は作ら

ず、必要になる数学的知識はすべてその場で、実例を交じえつつ説明していますので、本書を読むには高校レベルの数学を理解していれば充分で、あらかじめ機械学習の教科書を読んでおく必要はありません。むしろ、本書でテキストの統計モデルの基礎について一通り理解して道を作った上で、あらためて機械学習や統計学の本を読まれることで、それらがより読みやすくなり、理解が深まることを期待しています。

人文科学や社会科学においても、計量的な方法の必要性は増加の一途をたどっており、その多くの場合でテキストの分析が必要になっています。これまでは、統計が必要になるのは主に経済学を中心とした社会科学で、そこでは価格のような連続量が中心となっており、従来の統計学をそのまま適用することができました。これに対して、テキストのような**離散データ**が人文科学および社会科学の両方で必要になってきたのは、Webの発達によって電子テキストが容易に入手できるようになった比較的最近のことといえます。筆者ももとは文科系ですので、そうした分野の重要性はよく理解しているつもりです。本書をきっかけに、自分の手でテキストの統計的なモデル化と分析ができるようになっていただければと考えています。

また、本書はテキストを対象にしていますが、こうした離散データに対する方法論は言語だけでなく、他の種類の離散データについても同様に適用することができます。たとえば、文書に様々な種類の単語が含まれている状況は、コンビニやオンラインストアで客が様々な商品を購入する状況^{*1}と同じで、実際にこのための協調フィルタリングとよばれる分野は、本書で文書をモデル化するために導入するものと、まったく同じモデルを用いています。また、音楽の楽譜は離散的な記号で、言語のように構造を持っていますし、細胞内のDNAやゲノムはATGCあるいは20種類のアミノ酸を文字とした、一種の「言語」です。他にも、本書で紹介する離散データのための確率モデルは、多くの分野で適用できるのではないかと考えています。^{*2}

*1 これは、データマイニングの分野ではバスケット分析と呼ばれています。

*2 筆者はこれまでに、言語学、脳科学、音声認識、音楽情報処理、ロボティクス、バイオインフォマティクス、政治学といった分野と共同研究を行っています。

本書の対象読者

上に述べたように、本書はテキストの統計的なモデルに関心のある、広く人文科学・社会科学およびエンジニアの方、および言語の数学的なモデル化に興味を持つ一般の方を対象にしています。確率の基礎から始め、 n グラムモデルとは何か、Word2Vec は数学的には何をしているのか、テキストをどうクラスタリングすればよいのかなど、原理的な内容を一から解説します。

本書の特徴は、数式による定義を天下一に与えるのではなく、できる限りその導出や意味について解説していること、またブラックボックスのパッケージに頼らない、ということです。その代わりに、テキストをモデル化する自然言語処理の数理について丁寧に説明し、読者が自分の手で統計的な分析を使いこなせることを目標としています。たとえばパープレキシティ一つをとっても、定義の数式を頭ごなしに与えるのではなく、基礎となる自己情報量の説明から始め、どうしてパープレキシティを考えるのか、どういう意味を持っているのかを丁寧に説明するようにしました。テキストを扱う既存のパッケージでは、本書で紹介するような数学的な分析はほとんど行うことができませんが、それでも良いという方は多数の本がありますので、そちらをご参照ください。ただしもちろん、普段はパッケージを使った分析を行っている方にとっても、本書でテキスト分析の背景や理論について理解しておくことは、適切な分析を行う上で大きな助けになるでしょう。

必要となる数学について

本書はテキストの統計的なモデル化に関する本ですので、数式を用いることはどうしても必要になります。むしろ、言語のようなアナログな対象をどう数学的にモデル化していくかが、本書の主題といってもいいでしょう。ただし、必要になるのは数 III までの高校数学と、大学教養の数学 (解析と線形代数) の一部だけです。本書に限らず、統計学の理解には、数 III まではどうしても必要になります (e^x の微分・積分や対数の微分など)。現在はこのような場合のために、多くの優れた参考書が出版されていますので、必要な方は自習しておくようにしてください。筆者自身の数学も、数 III 以降はこうして自習したものです。と

いっても、難しい問題が解ける必要はなく、教科書レベルの内容が理解できていれば充分です。

逆に、本書では測度論のような高度な数学は仮定していません。確率変数のことを「役割を表す変数」(18 ページ)とするような直感的な記述は、意図的なものです。

深層学習との関係について

本書では、深層学習の手法としては最も基本的な単語埋め込み、文埋め込み、文書埋め込みまでの方法と、その背後にある理論について解説しました。現在の自然言語処理は、そうした埋め込みを背景にした LSTM や Transformer(BERT) といった深層学習の手法なしには語ることはできません。しかし、それらは実装して動かせても、理論的背景についてはほとんどわかっておらず、ほぼブラックボックスの状態です^{*3}。そこで本書では、単なるレシピ集となることを避け、それらについてはあえて紹介しないことにしました。ただし、本書で説明している単語埋め込み、文埋め込み、文書埋め込みの範囲でも(あるいは、それ以外の確率的手法を用いても)、非常に多くの自然言語処理ができることに注意してください。Transformer のような手法は非常に多くのデータと莫大な計算時間を必要としますが、それですべての自然言語処理の問題が解決するわけではなく、より深い応用には、本書で説明したような基礎知識と組み合わせることが必要です。深層学習手法の使い方については多くの書籍が出版されていますので、本書の3章の文献案内も参照してください。

実装とサポートサイトについて

本書では、Python 言語を使って、実際のテキストでの計算例を示しました。統計の分野では R 言語がよく使われていますが、R は実行速度が遅く、標準的にサポートされているデータ構造も少ないため、大量のデータを高速に扱う必要の

^{*3} これらが内部で何を行っているのかについての研究も一部では進みつつあり、筆者も、LSTM の内部状態が文法構造の埋め込みの深さをほぼ離散的に数えている[1]といった研究を行っています。

あるテキストの処理にはあまり向いていないからです。^{*4} Python 言語自体の説明はしていませんので、標準的な使い方は理解していることを前提としています。Python を知らない方は、巻末で紹介している参考文献をご覧ください。^{*5} ただし、実装部分をすべて理解できなくても、本文の記述は追えるように配慮したつもりです。核となる部分では C 言語に自動的にコンパイルされる Cython を用いれば、C 言語で実行する場合とほぼ同等の時間で実行することが可能です (5 章)。本書で用いた例や実装はすべて、本書のサポートサイト

```
http://www.ism.ac.jp/~daichi/textmodel/
```

で公開しています。また、Github のレポジトリ

```
https://github.com/daiti-m/textmodel/
```

でも同じファイルを公開していますので、コマンドラインから git で

```
% git clone https://github.com/daiti-m/textmodel
```

を実行すれば、本書で用いたスクリプトやデータをすべてダウンロードすることができます。

実験で用いているテキストデータはすべて自由に入手できるもので、上記のサポートサイトおよびレポジトリの “data” フォルダ

```
http://www.ism.ac.jp/~daichi/textmodel/data/
```

```
https://github.com/daiti-m/textmodel/data/
```

から入手することができます。サポートサイトには更新情報やリンク集など、有用な情報を載せていく予定ですので、合わせてご覧いただければ幸いです。

^{*4} 不可能ではなく、実際に計量政治学におけるテキスト分析のパッケージである `quanteda` は R 言語上のパッケージです。ただし、大量のテキストに対して重い計算を高速に実行するには、Rcpp のような外部言語の助けが必要になります。

^{*5} 参考書を使わなくとも、オンライン上の情報で Python の基本は十分にマスターすることができます。個人的には、M.Hiroi さんの「お気楽 Python プログラミング入門」http://www.nct9.ne.jp/m_hiroi/light/index.html#python_abc は大変お薦めで、これを読めば、本書に必要な基本は容易に理解できるのではないかと思います (Python2 系のため、`print` 文などに少し違いがあります)。<https://utokyo-ipp.github.io/> では東京大学の、<https://repository.kulib.kyoto-u.ac.jp/dspace/handle/2433/265459> では京都大学の Python プログラミング入門の実習テキストが、それぞれフリーで公開されています。

本書の記法

x	英小文字：変数
N	英大文字：定数またはデータ
\mathbf{x}	英小文字の太字：ベクトル
\mathbf{X}	英大文字の太字：行列
\mathcal{L}	カリグラフィック体：集合 (データセットや語彙など)
\mathbf{a}	タイプライター体：文字 (アルファベットの場合)
\mathbb{E}	期待値
\mathbb{V}	分散
\mathbb{R}	実数
\mathbb{I}	指示関数 ($\mathbb{I}()$ の中が真なら 1, 偽なら 0 を返す関数)
\mathbf{I}	単位行列
$\mathbf{0}$	ゼロベクトル (要素がすべて 0 のベクトル)
\cdot	ベクトルの内積
T	ベクトルおよび行列の転置
$\langle \dots \rangle_p$	確率分布 p による期待値
$\exp(x)$	e^x の別記法

確率分布の略記

\mathcal{N}	ガウス分布 (正規分布)
Po	ポアソン分布
Be	ベータ分布
Ga	ガンマ分布
Unif	一様分布
Mult	多項分布
Bernoulli	ベルヌーイ分布

目次

はじめに	i
1 テキストと言語のモデル化	1
1.1 言語とテキストの特徴	1
1.2 テキストの階層構造	3
1.3 教師あり学習と教師なし学習	4
1.4 統計的な方法とアドホックな方法	4
1.5 本書の自然言語処理のギャラリー	8
1.6 本書の構成と読み方	8
1.7 本書の例と実装について	9
2 文字の統計モデル	11
2.1 文字の頻度と出現確率	11
2.2 文字の同時確率	15
2.3 同時確率の周辺化	18
2.4 文字の条件つき確率	22
2.4.1 確率の連鎖則	24
2.4.2 ベイズの定理	27
2.5 文字 n グラムモデル	33
2.5.1 文字列の確率的生成	33
2.5.2 ゼロ頻度問題	39

2.6	統計モデルの学習と評価	45
2.6.1	学習データとテストデータ	45
2.6.2	予測確率とパープレキシティ	51
2.6.3	情報理論の基礎	53
2.6.4	統計モデルと汎化性能	63
2	章の演習問題	69
2	章の文献案内	72
3	単語の統計モデル	74
3.1	文字から単語へ	74
3.2	単語の統計と巾乗則	76
3.2.1	Heaps の法則	80
3.2.2	Zipf の法則	82
3.3*	単語の統計的フレーズ化	86
3.4	単語 n グラム言語モデル	93
3.4.1	ディリクレ分布	97
3.4.2	ディリクレ分布と多項分布	104
3.4.3	ディリクレ分布とポリア分布	106
3.4.4	階層ディリクレ言語モデル	112
3.4.5	Kneser–Ney 言語モデル	117
3.5	単語ベクトルとその原理	124
3.5.1	ニューラル n グラム言語モデル	125
3.5.2	スキップグラムと word2vec	130
3.5.3	単語ベクトルの学習	137
3.5.4	Word2vec と行列分解	142
3.5.5*	GloVe と意味方向の数理	148
3.5.6	単語ベクトルの分布とノルム	154
3	章の演習問題	162
3	章の文献案内	163
4	文の統計モデル	164

4.1	テキストの文分割	164
4.2	文ベクトルと意味的ランダムウォーク	167
4.2.1	RAND-walk モデル	168
4.2.2	文ベクトルの計算	169
4.3	構文解析と係り受け解析	175
4.4	隠れマルコフモデル (HMM)	180
4.4.1	HMM の状態推定	188
4.4.2	HMM のパラメータ推定	192
4.4.3	周辺化 Gibbs サンプリング	205
4.4.4	HMM による品詞の教師なし学習	212
4	章の演習問題	214
4	章の文献案内	215
5	文書の統計モデル	217
5.1	ナイーブベイズ法と単語集合表現	217
5.1.1	文書の分類確率	223
5.2	ユニグラム混合モデル (UM)	230
5.2.1	トピックの解釈と自己相互情報量	237
5.2.2	EM アルゴリズムによる学習	241
5.2.3*	UM のベイズ学習	246
5.3	ディリクレ混合モデル (DM)	250
5.3.1	単語単体と幾何的解釈	251
5.3.2	ポリア分布と単語のバースト性	257
5.4	潜在ディリクレ配分法 (LDA)	258
5.4.1	Gibbs サンプリングによる学習	260
5.4.2	LDA の幾何的解釈	270
5.4.3	トピックモデルの評価	274
5.5	ニューラル文書モデルと独立成分分析	279
5.5.1	文書ベクトルと Doc2Vec	282
5.5.2	単語ベクトル/文書ベクトルの解釈	290

5.6 確率的潜在意味スケーリング (PLSS)	293
5.6.1 Wordfish	295
5.6.2 確率的潜在意味スケーリング (PLSS)	297
5.6.3* PLSS の半教師あり学習	304
5章の演習問題	312
5章の文献案内	314
付録	316
A WSL の導入とコマンドラインの初歩	316
B MeCab のインストール	316
C デイリクレ分布の積分と期待値	316
D ポリア分布の α の最適化	316
E Kneser-Ney 平滑化の導出	316
F デイリクレ分布の α のベイズ推定	316
アルゴリズムの一覧	320
索引	323
参考文献	330

アルゴリズムの一覧

1 HMM の Gibbs サンプリング	204
2 HMM の周辺化 Gibbs サンプリング	210
3 UM の EM アルゴリズム	234
4 UM の周辺化 Gibbs サンプリング	248
5 DM の EM-Newton アルゴリズム	254
6 LDA の周辺化 Gibbs サンプリング	265
7 文書ベクトル (DocVec) の計算アルゴリズム	287

公式の一覧

(2.9) 同時確率の周辺化の公式	19
(2.11) 同時確率の周辺化の公式 (連続値の場合)	21
(2.18) 条件つき確率の公式	23
(2.19) 確率の連鎖則	25
(2.20) 独立な事象の同時確率	25
(2.34) ベイズの定理 (同時確率版)	32
(2.35) ベイズの定理 (比例版)	32

公式の一覧

321

(2.62) 自己情報量	55
(2.68) パープレキシティ	58
(2.69) KL ダイバージェンスの非負性	59
(2.75) クロスエントロピー	61
(2.76) クロスエントロピーとエントロピー	61
(3.0) Heaps の法則	80
(3.10) Zipf の法則	84
(3.30) デリクレ分布 (簡易版)	99
(3.31) デリクレ分布の期待値	100
(3.34) デリクレ分布 (正式版)	101
(3.36) ガンマ関数	101
(3.46) デリクレ事後分布	105
(3.48) デリクレ平滑化	106
(3.58) ポリア分布	109
(3.60) ポリア分布の最適化の公式	110
(3.70) 絶対平滑化	120
(3.72) Kneser–Ney 平滑化	121
(3.84) シグモイド関数	130
(3.85) \tanh とシグモイド関数の関係	131
(3.87) シグモイド関数の補関数	131
(4.18) HMM の同時確率	185
(4.67) HMM の周辺化 Gibbs サンプリングの公式	210
(5.4) ナイーブベイズ法の文書確率	224
(5.11) logsumexp	231
(5.23) 正規化自己相互情報量	240
(5.33) UM の文書確率	244
(5.51) DM の文書確率	255
(5.55) 指数分布族 DCM 分布	258
(5.58) LDA の文書同時確率	261
(5.75) LDA の周辺化 Gibbs サンプリングの公式	266

(0.138) ディリクレ分布の α のサンプリングの公式	318
(0.139) ディリクレ分布の η のサンプリングの公式	319

索引

Symbols

$\mathbb{I}()$, 199

∞ グラム言語モデル, 39

$p()$, 14

n グラム, 39

n グラム言語モデル, 39

0 グラム, 35

1 グラム, 15

2 グラム, 15

3 グラム, 39

4 グラム, 123

A

All-but-the-top, 172

awk, 36, 44, 63, 67

B

bag of words, 219

Baum–Welch アルゴリズム, 193, 196

BERT, 226

BFRY 過程, 85

bit, 55

C

CBOW, → 連続的単語集合, 137

CCG, 175

CKY アルゴリズム, 180

CRF, 9, 75, 181

Cython, 267

D

DCM 分布, → デイリクレ複合多項分布

DM, → デイリクレ混合モデル

DNA, ii, 124, 218

Doc2Vec, 282

DocVec, 283

double power-law, 85

E

EDCM 分布, 258

EM アルゴリズム, 182, 196, 215, 241, 261, 296, 298

G

Gibbs サンプルング, 193, 196, 247, 261, 262

H

hapax legomenon, → 孤語

Heaps の法則, 80

HMM, 183

I

ICA, 291

J

Jensen-Shannon ダイバージェンス, 60

JUMAN, 76, 180, 188

K

KL ダイバージェンス, → Kullback-Leibler ダイバージェンス, 161

Kneser-Ney 平滑化, 117, 121

Kullback-Leibler ダイバージェンス,
59

K 平均法, 235

L

LDA, 253, → 潜在ディリクレ配分法

LSI, 281, 282

M

MAP 推定, 306

MCMC 法, 179, 298

MeCab, 75, 180, 182, 221, 227, 309

mecab-ipadic-NEologd, 87

Moore-Penrose の擬似逆行列, 286

N

nat, 55

Normalized PMI, 239

NPMI, → 正規化自己相互情報量, 275

NPYCRF, 183

NPYLM, 41, 77

P

Pólya 分布, → ポリア分布

PCFG, 175

Perl, 36, 67

Pitman-Yor 過程, 253

PLSI, 260, 281

PLSS, 299

PMI, → 自己相互情報量

Pochhammer 関数, 111

Pointwise Mutual Information, 239

Q

quanteda, v, 6, 86, 303

Q 関数, 244

R

Rao-Blackwell 化, 211, 270

RNN, → 再帰的ニューラルネットワーク

S

sed, 67, 86

SIF, 171

Softmax 関数, 127

spaCy, 86

SVD, → 特異値分解

SVM, 295
SVMlight 形式, 114, 219

T

tf.idf, 281, 288, 289

U

Unigram Mixtures, 235
uSIF, 172

V

VAE, 241
Viterbi アルゴリズム, 189, 202, 211

W

word2vec, 132

Z

Zipf の法則, 81, 84, 140

あ

圧縮, 117
アノテーション, 184
アンダーフィット, 65
意味, 177
ウェーバー＝フェヒナーの法則, 289
埋め込み文, 175
エントロピー, 56
オーバーフィット, 64
音楽, ii

か

階層構造, 4

階層ディリクレ過程, 117, 278
階層ディリクレ言語モデル, 113
階層 Pitman-Yor 過程, 117
階層ベイズ, 256
開発データ, 47
ガウス-エルミート求積, 306
ガウス過程, 8
過学習, 64, 212
係り受け解析, 175, 177
学習データ, 47
確率の連鎖則, 25
確率分布, 15, 97, 259
確率変数, 18
確率モデル, 97
隠れマルコフモデル, 183
加算平滑化, 41, 95
過少適合, 65
カルマンフィルタ, 8
感情分析, 227
幾何平均, 53, 54
期待値, 205
期待値伝搬法, 261
機能語, 169
逆温度, 129
逆文書頻度, 289
教師あり学習, i, 9, 166, 176, 181, 188
教師なし学習, i, 166, 176, 178, 180,
182, 193
教師なし形態素解析, 75, 77, 183
教師なし構文解析, 180

- 教師なし談話構造解析, 184
 協調フィルタリング, ii
 共役, 105
 行列分解, 280, 281
 局所解, 196, 247
 極性, 227
 均衡コーパス, 78
 グラフィカルモデル, 108
 クロスエントロピー, 61
 クロスバリデーション, 48, 51, 65
 訓練データ, 47
 経験ベイズ法, 110
 形態素解析, 75
 形態論, 3
 結合確率, 18
 言語学, 4, 67, 169, 180, 181, 217
 言語資源, 178
 検証データ, 47
 交差検証, 48
 構文, 175
 項目反応理論, 297
 コーパス, 78
 コーパス言語学, 82
 孤語, 82
 誤差逆伝搬法, 127
 固有表現認識, 87
 混合モデル, 96, 170
- さ**
- 再帰的ニューラルネットワーク, 130
 最尤推定, 15, 22, 212, 241
 算術符号, 39
 算術平均, 54
 次元の呪い, 125, 138, 180, 189
 事後確率, 32
 自己情報量, 55
 自己相互情報量, 8, 90, 144, → Point-wise Mutual Information
 指示関数, 199, 261
 地震, 313
 指数分布族, 258
 事前確率, 32
 尺度化, 293
 自由エネルギー, 243
 周辺化, 19
 周辺化 Gibbs サンプルング, 210, 248
 周辺確率, 17
 主成分分析, 5, 271
 出力確率, 185
 条件つき確率, 22
 状態遷移確率, 185
 情報量, 7, 53, 192
 情報理論, 39, 45, 53, 72, 75, 90, 117, 189
 深層学習, 93, 124, 138, 144
 心理学, 290, 293
 心理統計学, 297
 スキップグラム, 132, 139
 スコア, 192
 ストップワード, 237

スパース, 219
スムージング, → 平滑化
正規化自己相互情報量, 90, → Normalized PMI
政治学方法論, 295, 301
生成モデル, 107, 183
絶対平滑化, 117, 120
絶対割引, 118
ゼロ頻度問題, 40, 95, 125
潜在ディリクレ配分法, 212, 260, 271
潜在変数, 232, 242

た

ダイガンマ関数, 110, 255
対数尤度比, 147
タガー, 181
多項分布, 98, 185, 235, 297
単語集合, 137, → bag of words, 224
単語単体, 252, 271
単語分割, 75
単語-文書行列, 282
単語ベクトル, 125
単体, 98, 252
談話構造, 184
中心化, 156
中心極限定理, 292
チューリング完全, 68
長大語, 76
チョムスキー標準形, 179
低資源, 178
ディリクレ混合モデル, 252

ディリクレ複合多項分布, → ポリア分布

ディリクレ分布, 99, 252, 259

ディリクレ平滑化, 106, 117

データスパースネス, 125

テキストマイニング, i

テストデータ, 51

統計力学, 129, 139

同時確率, 18

動的計画法, 189

特異値分解, 145

独立, 26

トピック, 237

トピック単体, 271

トピックモデル, 237

トライグラム, → 3 グラム, 42, 116

トレリス, 188

な

ナイーブベイズ法, 225

内容語, 169

二重分節, 4

ニューラル n グラム言語モデル, 125

は

パーザー, 176

バースト性, 109, 250

パープレキシティ, 57

バイオインフォマティクス, 253, 264

バイグラム, → 2 グラム, 35, 95, 112, 120, 198

白色化, 157, 159, 291
 パス, 188
 バスケット分析, ii
 バックオフ, 115
 バリマックス回転, 290
 汎化性能, 64
 半教師あり学習, 9, 178, 183
 非負値行列因子分解, 280
 品詞, 75, 180
 符号長, 56, 192
 負担率, 232, 247
 物理学, 149, 196, 211, 215, 243
 負例サンプリング, 139
 分岐数, 54, 57
 文境界認識, 165
 分散, 205
 分散表現, 125
 文書, 217
 文書-単語行列, 219
 文書頻度, 276, 288
 分配関数, 139
 文分割, 165
 文ベクトル, 167
 文法, 175
 文脈, 181
 分類器, 295
 平滑化, 8, 41, 140
 平均分岐数, 57
 ベイジアン, 29
 ベイズ学習, 247

ベイズ推定, 241, 306
 ベイズの定理, 29
 ベータ関数, 316
 巾乗則, 84
 ベン図, 26
 変分ベイズ法, 261
 ポアソン過程, 313
 補助変数法, 316
 ポリア分布, 109, 254, 257, 267, 273,
 316

ま

マスク化言語モデル, 138
 マルコフ確率場, 138
 マルコフ性, 45, 183, 203
 マルコフブランケット, 202
 マルコフモデル, 45
 無限隠れマルコフモデル, 212
 文字列, 11, 33
 モンテカルロ EM アルゴリズム, 247,
 273
 モンテカルロ積分, 62

や

尤度関数, 32
 ユニグラム, → 1 グラム, 33, 157, 219
 ユニグラム混合モデル, → Unigram
 Mixtures
 用語頻度, 289
 予測, 50

ら

ラグランジュの未定乗数法, 15, 245

ラベルスイッチング, 211

離散データ, ii

理想点, 296

レジスター, 217

連続的単語集合, 132, 137

わ

歪度, 292

参考文献

- [1] Chihiro Shibata, Kei Uchiumi, and Daichi Mochihashi. How LSTM Encodes Syntax: Exploring Context Vectors and Semi-Quantization on Natural Text. In *COLING 2020*, pages 4033–4043, 2020.
- [2] André Martinet. *Éléments de linguistique générale, 5th ed.* Armand Colin: Paris, 2008.
- [3] Hinrich Schütze. Dimensions of Meaning. In *Proceedings of Supercomputing'92*, pages 787–796, 1992.
- [4] Christos H.Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempalad. Latent Semantic Indexing: A Probabilistic Analysis. *Journal of Computer and System Sciences*, 61:217–235, 2000.
- [5] 石井公成. 仏教学における N-gram の活用. In **東京大学東洋文化研究所附属東洋学情報センター報「明日の東洋学」**, number 8, pages 2–4, 2002.
- [6] 持橋大地, 大羽成征. **ガウス過程と機械学習**. 機械学習プロフェッショナルシリーズ. 講談社, 2019.
- [7] 高村大也. **言語処理のための機械学習入門**. 自然言語処理シリーズ. コロナ社, 2010.
- [8] 佐藤坦. **はじめての確率論 測度から確率へ**. 共立出版, 1994.
- [9] 清水泰隆. **統計学への確率論, その先へ: ゼロからの測度論的理解と漸近理論への架け橋**. 内田老鶴圃, 2021.
- [10] Pierre-Simon Laplace. *Essai Philosophique sur les Probabilités*. 1814. 『確率の哲学的試論』内井惣七 (訳), 岩波文庫, 1997.
- [11] Alastair J. Walker. An Efficient Method for Generating Discrete Random Variables with General Distributions. *ACM Transactions on Mathematical Software*, 31(3):253–256, 1977.
- [12] Feras Saad, Cameron Freer, Martin Rinard, and Vikash Mansinghka. The Fast Loaded Dice Roller: A Near-Optimal Exact Sampler for Discrete Probability Distributions. In *AISTATS 2020*, pages 1036–1046, 2020.
- [13] David J Ward, Alan F Blackwell, and David J. C. MacKay. Dasher - a Data Entry Interface Using Continuous Gestures and Language Models. In *UIST 2000*, 2000.
- [14] Li Du, Lucas Torroba Hennigen, Tiago Pimentel, Clara Meister, Jason Eisner, and Ryan Cotterell. A Measure-Theoretic Characterization of Tight Language

- Models. In *ACL 2023*, pages 9744–9770, 2023.
- [15] Daichi Mochihashi and Eiichiro Sumita. The Infinite Markov Model. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 1017–1024, 2008.
- [16] 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*, 48(12):4023–4032, 2007.
- [17] Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. Bayesian Unsupervised Word Segmentation with Nested Pitman-Yor Language Modeling. In *Proceedings of ACL-IJCNLP 2009*, pages 100–108, 2009.
- [18] Andrei A. Markov. An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains. In *Proceedings of the Academy of Sciences, St. Petersburg*, volume 7, pages 153–162, 1913.
- [19] Roman Jakobson. *一般言語学*. みすず書房, 1973.
- [20] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2007.
- [21] 金子弘昌. *Python で学ぶ実験計画法入門 ベイズ最適化によるデータ解析*. KS 情報科学専門書. 講談社, 2021.
- [22] 韓太舜, 小林欣吾. *情報と符号化の数理*. 培風館, 1999.
- [23] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- [24] Jihyeon Roh, Sang-Hoon Oh, and Soo-Young Lee. Unigram-Normalized Perplexity as a Language Model Performance Measure with Different Vocabulary Sizes. In *arXiv preprint*, 2020. <https://arxiv.org/abs/2011.13220>.
- [25] S. Kullback and R. A. Leibler. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [26] Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50:5–43, 2003.
- [27] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory, 2nd Edition*. Wiley Series in Telecommunications. Wiley-Interscience, 2013.
- [28] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [29] Tom Brown, Benjamin Mann, Nick Ryder, et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.
- [30] Larry Wall, Tom Christiansen, and Jon Orwant. *プログラミング Perl 第 3 版*. オライリー・ジャパン, 2002.
- [31] 山口和紀, 古瀬一隆. *新 The UNIX Super Text (上)(下) 改訂増補版*. 技術評論社, 2003.
- [32] Dale Dougherty and Arnold Robbins. *sed & awk プログラミング 改訂版*. A nutshell handbook. O’Reilly Japan, 1997.

- [33] A.V. エイホ, B.W. カーニハン, P.J. ワインバーガー. **プログラミング言語 AWK**. アジソン ウェスレイ・トッパン 情報科学シリーズ. トッパン, 1989.
- [34] Andrew Marc Greene. BASIX: An interpreter written in TEX. *TUGboat*, 11(3):381–392, 1990. <https://www.ctan.org/tex-archive/macros/generic/basix>.
- [35] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.
- [36] Christopher D. Manning and Hinrich Schütze. **統計的自然言語処理の基礎**. 共立出版, 2017.
- [37] 北研二. **確率的言語モデル**. 言語と計算 4. 東京大学出版会, 1999.
- [38] C.M. Bishop, 元田, 栗田, 樋口, 松本, and 村田 (監訳). **パターン認識と機械学習: ベイズ理論による統計的予測** (上)(下) (*Pattern Recognition and Machine Learning*). Springer/丸善出版, 2007, 2008.
- [39] Thomas M. Cover and Joy A. Thomas. **情報理論 —基礎と広がり—**. 共立出版, 2012.
- [40] 伊庭幸人. 「情報」に関する 13 章 – 私家版・情報学入門 –. *物性研究*, 78(2):172–193, 2002. <https://www.ism.ac.jp/~iba/a19.pdf>.
- [41] 近藤泰弘, 近藤みゆき. 平安時代古典語古典文学研究のための N-gram を用いた解析手法. In **言語情報処理学会第 7 回年次大会 発表論文集**, 2001.
- [42] Folger Karsdorp, Mike Kestemont, and Allen Riddell. *Humanities Data Analysis: Case Studies with Python*. Princeton University Press, 2021.
- [43] 持橋大地, 山田武士, 上田修功. ベイズ階層言語モデルによる教師なし形態素解析. **情報処理学会研究報告 2009-NL-190**, 2009.
- [44] Linguistic Data Consortium. Web 1T 5-gram Version 1, 2006. <https://catalog.ldc.upenn.edu/LDC2006T13>.
- [45] 言語資源協会. Web 日本語 N グラム 第 1 版, 2007. <https://www.gsk.or.jp/catalog/gsk2007-c>.
- [46] Henry Kučera and W. Nelson Francis. *Computational Analysis of Present-Day American English*. Brown University Press, 1967.
- [47] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *LREC 2012*, pages 759–765, 2012.
- [48] Harold Stanley Heaps. *Information Retrieval, Computational and Theoretical Aspects*. Academic Press, 1978.
- [49] Gustav Herdan. *Type-Token Mathematics: A Textbook of Mathematical Linguistics*. Mouton en Company, 1960.
- [50] George Kingsley Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. The MIT Press Classics Series. The MIT Press, 1935.
- [51] Micheline Petruszewycz. L’histoire de la loi d’Estoup-Zipf : documents. *Mathématiques et Sciences humaines, Tome*, 44:41–56, 1973.
- [52] 水谷静夫. **数理言語学**. 現代数学レクチャーズ D-3. 培風館, 1982.
- [53] Alain Lelu. Jean-Baptiste Estoup and the origins of Zipf’s law: a stenogra-

- pher with a scientific mind (1868-1950). *Boletín de Estadística e Investigación Operativa*, 30(1):66–77, 2014.
- [54] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.
- [55] 田中久美子. **言語とフラクタル 使用の集積の中にある偶然と必然**. 東京大学出版会, 2021.
- [56] S. Naranan and V.K. Balasubrahmanyam. Models for Power Law Relations in Linguistics and Information Science. *Journal of Quantitative Linguistics*, 5(1-2):35–61, 1998.
- [57] Martin Gerlach and Eduardo G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, 3:021006, 2013.
- [58] Fadhel Ayed, Juho Lee, and Francois Caron. Beyond the Chinese Restaurant and Pitman-Yor processes: Statistical Models with double power-law behavior. In *ICML 2019*, pages 395–404, 2019.
- [59] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き用辞書生成システム NEologd の運用 — 一文書分類を例にして —. In **情報処理学会 自然言語処理研究会研究報告**, pages NL-229–15, 2016.
- [60] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119, 2013.
- [61] 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*, 18(4):467–479, 1992.
- [62] Kenneth Ward Church and Patrick Hanks. Word Association Norms, Mutual Information, and Lexicography. In *ACL 1989*, pages 76–83, 1989.
- [63] Gerlof Bouma. Normalized (Pointwise) Mutual Information in Collocation Extraction. In *Proceedings of GSCL*, pages 31–40, 2009.
- [64] Rameshwar D. Gupta and Donald St. P. Richards. The History of the Dirichlet and Liouville Distributions. *International Statistical Review*, 69(3):433–446, 2001.
- [65] Luc Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986. <http://www.nrbook.com/devroye/>.
- [66] David J. C. MacKay and L. Peto. A Hierarchical Dirichlet Language Model. *Natural Language Engineering*, 1(3):1–19, 1994.
- [67] Thomas P. Minka. Estimating a Dirichlet distribution, 2000. <http://research.microsoft.com/~minka/papers/dirichlet/>.
- [68] Yee Whye Teh. A Bayesian Interpretation of Interpolated Kneser-Ney. Technical Report TRA2/06, School of Computing, National University of Singapore, 2006.
- [69] Phil Cowans. *Probabilistic Document Modelling*. PhD thesis, University of Cambridge, 2006. <http://www.inference.phy.cam.ac.uk/pjc51/thesis/index.html>.
- [70] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language

- modeling. In *Proceedings of ICASSP*, volume 1, pages 181–184, 1995.
- [71] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A Neural Probabilistic Language Model. In *Advances in Neural Information Processing Systems*, volume 13, 2000.
- [72] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.
- [73] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. *Distributed Representations*, pages 77–109. MIT Press, Cambridge, MA., 1986.
- [74] John Bridle. Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters. In *NIPS 1989*, pages 211–217, 1989.
- [75] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *NAACL 2013*, pages 746–751, 2013.
- [76] Jun Suzuki and Masaaki Nagata. Right-truncatable Neural Word Embeddings. In *NAACL 2016*, pages 1145–1151, 2016.
- [77] Omer Levy and Yoav Goldberg. Dependency-Based Word Embeddings. In *ACL 2014 (short paper)*, pages 302–308, 2014.
- [78] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS 2010*, pages 297–304, 2010.
- [79] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.
- [80] Omer Levy and Yoav Goldberg. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems 27*, pages 2177–2185, 2014.
- [81] John A. Bullinaria and Joseph P. Levy. Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, 39:510–526, 2007.
- [82] John A. Bullinaria and Joseph P. Levy. Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists, Stemming and SVD. *Behavior Research Methods*, 44:890–907, 2012.
- [83] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.
- [84] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In *EMNLP 2014*, pages 1532–1543, 2014.
- [85] 横井 祥, 下平 英寿. 単語埋め込みの確率的等方化. In **言語処理学会第 27 回年次大会**, pages A7–1, 2021. https://www.anlp.jp/proceedings/annual_meeting/2021/pdf_dir/A7-1.pdf.
- [86] Adriaan M. J. Schakel and Benjamin J. Wilson. Measuring Word Significance using Distributed Representations of Words. In *arXiv preprint*, 2015. <https://>

- arxiv.org/abs/1508.02297.
- [87] 大山百々勢, 横井祥, 下平英寿. 単語ベクトルの長さは意味の強さを表す. In **言語処理学会第 28 回年次大会**, pages A5–1, 2022.
- [88] H. S. Terrace, L. A. Petitto, R. J. Sanders, and T. G. Bever. Can an Ape Create a Sentence? *Science*, 206(4421):891–902, 1979.
- [89] Toshitaka N. Suzuki. Animal linguistics: Exploring referentiality and compositionality in bird calls. *Ecological Research*, 36:221–231, 2021.
- [90] Nipun Sadvilkar and Mark Neumann. PySBD: Pragmatic Sentence Boundary Disambiguation. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 110–114, 2020.
- [91] Rachel Wicks and Matt Post. A unified approach to sentence segmentation of punctuated text in many languages. In *ACL 2021*, pages 3995–4007, 2021.
- [92] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-Thought Vectors. In *NIPS 2015*, 2015.
- [93] Prakhar Gupta, Matteo Pagliardini, and Martin Jaggi. Better Word Embeddings by Disentangling Contextual n-Gram Information. In *NAACL-HLT 2019*, pages 933–939, 2019.
- [94] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR 2017*, 2017.
- [95] Sanjeev Arora, Yuezhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A Latent Variable Model Approach to PMI-based Word Embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.
- [96] 伊庭幸人. **ベイズ統計と統計物理**. 岩波講座 物理の世界. 岩波書店, 2003.
- [97] Jiaqi Mu and Pramod Viswanath. All-but-the-Top: Simple and Effective Post-processing for Word Representations. In *ICLR 2018*, 2018.
- [98] Kawin Ethayarajh. Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 91–100, 2018.
- [99] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL System Demonstrations*, pages 55–60, 2014. <http://www.aclweb.org/anthology/P/P14/P14-5010>.
- [100] Nikita Kitaev and Dan Klein. Constituency Parsing with a Self-Attentive Encoder. In *ACL 2018*, pages 2676–2686, 2018.
- [101] Mai Omura and Masayuki Asahara. UD-Japanese BCCWJ: Universal Dependencies Annotation for the Balanced Corpus of Contemporary Written Japanese. In *Second Workshop on Universal Dependencies (UDW 2018)*, pages 117–125, 2018.
- [102] 松田寛. GiNZA — Universal Dependencies による実用的日本語解析 —. **自然言語処理**, 27(3):695–701, 2020.

- [103] Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. Bayesian Inference for PCFGs via Markov Chain Monte Carlo. In *Proceedings of HLT/NAACL 2007*, pages 139–146, 2007.
- [104] J. ホップクロフト, R. モトワニ, J. ウルマン. **オートマトン 言語理論 計算論 I**. サイエンス社, 2003.
- [105] 能地宏. 文に隠れた構文構造を発見する統計モデル (特集「統計的言語研究の現在」). **統計数理**, 64(2):145–160, 2016.
- [106] Sadao Kurohashi and Makoto Nagao. Building a Japanese Parsed Corpus while Improving the Parsing System. In *Proceedings of LREC 1998*, pages 719–724, 1998. <http://nlp.kuee.kyoto-u.ac.jp/nl-resource/corpus.html>.
- [107] 持橋大地, 能地宏. 無限木構造隠れ Markov モデルによる階層的品詞の教師なし学習. **情報処理学会研究報告 2016-NL-226**, 12:1–11, 2016.
- [108] Julian Kupiec. Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language*, 6(3):225–242, 1992.
- [109] Bernard Merialdo. Tagging English Text with a Probabilistic Model. *Computational linguistics*, 20(2):155–171, 1994.
- [110] Sharon Goldwater and Tom Griffiths. A Fully Bayesian Approach to Unsupervised Part-of-Speech Tagging. In *Proceedings of ACL 2007*, pages 744–751, 2007.
- [111] Jurgen Van Gael, Andreas Vlachos, and Zoubin Ghahramani. The infinite HMM for unsupervised PoS tagging. In *EMNLP 2009*, pages 678–687, 2009.
- [112] Phil Blunsom and Trevor Cohn. A Hierarchical Pitman-Yor Process HMM for Unsupervised Part of Speech Induction. In *ACL 2011*, pages 865–874, 2011.
- [113] Jun Suzuki and Hideki Isozaki. Semi-Supervised Sequential Labeling and Segmentation Using Giga-Word Scale Unlabeled Data. In *ACL:HLT 2008*, pages 665–673, 2008.
- [114] Ryo Fujii, Ryo Domoto, and Daichi Mochihashi. Nonparametric Bayesian Semi-supervised Word Segmentation. *Transactions of ACL*, 5:179–189, 2017.
- [115] 竹内孔一, 松本裕治. 隠れマルコフモデルによる日本語形態素解析のパラメータ推定. **情報処理学会論文誌**, 38(3):500–509, 1997.
- [116] Andrew J. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.
- [117] G. David Forney Jr. The Viterbi Algorithm: A Personal History. In *arXiv preprint*, 2005. arXiv:cs/0504020 [cs.IT].
- [118] Laurence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.
- [119] Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.
- [120] Julian Besag. On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society, Series B*, 48(3):259–302, 1986.

- [121] Daichi Mochihashi. Unbounded Slice Sampling. Technical Report Research Memorandum No.1209, The Institute of Statistical Mathematics, 2020. arXiv:2010.01760.
- [122] Christophe Andrieu. On random- and systematic-scan samplers. *Biometrika*, 103(3):719–726, 2016.
- [123] Jun S. Liu. The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.
- [124] George Casella and Christian P. Robert. Rao-Blackwellisation of Sampling Schemes. *Biometrika*, 83(1):81–94, 1996.
- [125] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The Infinite Hidden Markov Model. In *NIPS 2001*, pages 577–585, 2001.
- [126] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet Processes. *JASA*, 101(476):1566–1581, 2006.
- [127] David J.C. MacKay. Ensemble Learning for Hidden Markov Models. Technical report, Cavendish Laboratory, University of Cambridge, 1997.
- [128] Matthew J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003. <https://cse.buffalo.edu/faculty/mbeal/thesis/>.
- [129] Steven L. Scott. Bayesian Methods for Hidden Markov Models. *Journal of the American Statistical Association*, 97:337–351, 2002.
- [130] Kevin Knight. Bayesian Inference with Tears, 2009. <https://kevincrawfordknight.github.io/papers/bayes-with-tears.pdf>.
- [131] Reina Akama, Kento Watanabe, Sho Yokoi, Sosuke Kobayashi, and Kentaro Inui. Unsupervised Learning of Style-sensitive Word Vectors. In *ACL 2018*, pages 572–578, 2018.
- [132] Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The Multilingual Amazon Reviews Corpus. In *EMNLP 2020*, pages 4563–4568, 2020.
- [133] 株式会社ロンウィット. Livedoor ニュースコーパス, 2014. <http://www.rondhuit.com/download.html#ldcc>.
- [134] 宮内裕人, 鈴木陽也, 秋山和輝, 梶原智之, 二宮崇, 武村紀子, 中島悠太, 長原一. 主観と客観の感情極性分類のための日本語データセット. In **言語処理学会第28回年次大会**, pages 1495–1499, 2022.
- [135] Haruya Suzuki, Yuto Miyauchi, Kazuki Akiyama, Tomoyuki Kajiwara, Takashi Ninomiya, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. A Japanese Dataset for Subjective and Objective Sentiment Polarity Classification in Micro Blog Domain. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*, pages 7022–7028, 2022.
- [136] Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.
- [137] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Alloca-

- tion. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [138] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.
- [139] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *arXiv preprint*, 2013. arXiv:1312.6114.
- [140] 篠崎寿夫, 松森徳衛, 吉田正広. **現代工学のための変分学入門**. 現代工学社, 1991.
- [141] Greg C.G. Wei and Martin A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man’s Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.
- [142] Kenneth W. Church. Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to $p/2$ than p^2 . In *COLING 2000*, pages 173–179, 2000.
- [143] K. Sjölander, K. Karplus, M.P. Brown, R. Hughey, R. Krogh, I.S. Mian, and D. Haussler. Dirichlet Mixtures: A Method for Improved Detection of Weak but Significant Protein Sequence Homology. *Computing Applications in the Biosciences*, 12(4):327–345, 1996.
- [144] 山本 幹雄, 貞光 九月, 三品 拓也. 混合ディリクレ分布を用いた文脈のモデル化と言語モデルへの応用. **情報処理学会研究報告 2003-SLP-48**, pages 29–34, 2003.
- [145] Issei Sato and Hiroshi Nakagawa. Topic models with power-law using Pitman-Yor process. In *KDD 2010*, page 673–682, 2010.
- [146] 山本幹雄, 持橋大地. Topic に基づく統計的言語モデルの最前線 –PLSI から HDP まで-. In **言語処理学会年次大会 2006 チュートリアル**, pages 11–28, 2006.
- [147] 貞光九月, 待鳥裕介, 山本幹雄. 混合ディリクレ分布パラメータの階層ベイズモデルを用いたスムージング法. **情報処理学会研究報告 2004-SLP-53**, pages 1–6, 2004.
- [148] Charles Elkan. Clustering Documents with an Exponential-Family Approximation of the Dirichlet Compound Multinomial Distribution. In *ICML 2006*, pages 289–296, 2006.
- [149] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. In *Neural Information Processing Systems 14*, 2001.
- [150] Yee Whye Teh, David Newman, and Max Welling. A Collapsed Variational Bayesian Inference Algorithm for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, 2006.
- [151] Tom Griffiths. Gibbs sampling in the generative model of Latent Dirichlet Allocation. Technical Report 11, Stanford University, 2002.
- [152] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *PNAS*, 101:5228–5235, 2004.
- [153] Thomas Minka. Expectation-Propagation for the Generative Aspect Model. In *UAI 2002*, page 352–359, 2002.
- [154] 佐藤一誠. **トピックモデルによる統計的潜在意味解析**. 自然言語処理シリーズ 8. コロナ社, 2015.
- [155] J.K. Pritchard, M. Stephens, and P. J. Donnelly. Inference of Population Struc-

- ture Using Multilocus Genotype Data. *Genetics*, 155:945–959, 2000.
- [156] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation Methods for Topic Models. In *ICML 2009*, pages 1105–1112, 2009.
- [157] Wray Buntine and Aleks Jakulin. Applying Discrete PCA in Data Analysis. In *UAI 2004*, pages 59–66, 2004.
- [158] Hanna M. Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why Priors Matter. In *NIPS 2009*, pages 1973–1981, 2009.
- [159] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. Reading Tea Leaves: How Humans Interpret Topic Models. In *NIPS 2009*, 2009.
- [160] Jey Han Lau, David Newman, and Timothy Baldwin. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *EACL 2014*, pages 530–539, 2014.
- [161] Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. Topic Modeling with Wasserstein Autoencoders. In *ACL 2019*, pages 6345–6381, 2019.
- [162] 石井健一郎, 上田修功. 続・わかりやすいパターン認識—教師なし学習入門—. オーム社, 2014.
- [163] Daniel Lee and H. Sebastian Seung. Algorithms for Non-negative Matrix Factorization. In *Advances in Neural Information Processing Systems*, volume 13, 2000.
- [164] John Canny. GaP: A Factor Model for Discrete Data. In *SIGIR 2004*, pages 122–129, 2004.
- [165] 持橋大地. GaP, NMF, and more, 2006. <http://chasen.org/~daiti-m/paper/gap-nmf.pdf>.
- [166] S. Deerwester, Susan T. Dumais, and George W. Furnas. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
- [167] Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *SIGIR 1999*, pages 50–57, 1999.
- [168] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In *ICML 2014*, pages 1188–1196, 2014.
- [169] Daichi Mochihashi. Researcher2Vec: Neural Linear Model of Scholar Recommendation for Funding Agency. In *International Society for Scientometrics and Informatics (ISSI 2023)*, 2023.
- [170] 南風原朝和. 心理統計学の基礎. 有斐閣アルマ. 有斐閣, 2002.
- [171] Sungjoon Park, JinYeong Bak, and Alice Oh. Rotated Word Vector Representations and their Interpretability. In *EMNLP 2017*, pages 401–411, 2017.
- [172] Hiroaki Yamagiwa, Momose Oyama, and Hidetoshi Shimodaira. Discovering Universal Geometry in Embeddings with ICA. In *EMNLP 2023*, pages 4647–4675, 2023.
- [173] 東京大学教養学部統計学教室. 統計学入門. 基礎統計学 I. 東京大学出版会, 1991.

- [174] Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. Ultradense Word Embeddings by Orthogonal Transformation. In *NAACL 2016*, pages 767–777, 2016.
- [175] Bianca Zadrozny and Charles Elkan. Transforming Classifier Scores into Accurate Multiclass Probability Estimates. In *KDD 2002*, pages 694–699, 2002.
- [176] Jonathan B. Slapin and Sven-Oliver Proksch. A Scaling Model for Estimating Time-Series Party Positions from Texts. *American Journal of Political Science*, 52(3):705–722, 2008.
- [177] Kohei Watanabe. Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages. *Communication Methods and Measures*, pages 1–23, 2020.
- [178] 持橋大地. 確率的潜在意味スケールリング. In **情報処理学会研究報告 2021-NL-249**, number 9, pages 1–16, 2021.
- [179] Philipp Dufter and Hinrich Schütze. Analytical Methods for Interpretable Ultradense Word Embeddings. In *EMNLP-IJCNLP 2019*, pages 1185–1191, 2019.
- [180] Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. In *EMNLP 2020: System Demonstrations*, pages 23–30, 2020.
- [181] Malay Ghosh. Inconsistent maximum likelihood estimators for the Rasch model. *Statistics & Probability Letters*, 23(2):165–170, 1995.
- [182] Radford M. Neal. *MCMC Using Hamiltonian Dynamics*. Chapman and Hall/CRC, 2011.
- [183] Qing Liu and Donald A. Pierce. A Note on Gauss-Hermite Quadrature. *Biometrika*, 81(3):624–629, 1994.
- [184] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC, 2013. <http://www.stat.columbia.edu/~gelman/book/>.