社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[招待講演] 自然言語処理におけるベイズ統計

持橋　大地†

† ATR 音声言語コミュニケーション研究所 /
独立行政法人 情報通信研究機構
619-0288 「けいはんな学研都市」光台 2-2-2
E-mail: †daichi.mochihashi@atr.jp

あらまし　高次元の離散データを扱う自然言語処理において最近用いられるようになっている, ベイズ統計的手法について概観し, 最近の発展と展望について述べる. 自然言語処理の知識を特に仮定せず, 識別モデルのベイズ学習について概説し, ナイーブベイズとその教師なしベイズ学習である DM, および LDA について解説する. これらのベイズ的手法は画像, 音声のような連続データと容易に組み合わせることができ, 興味深い応用を持っている.

キーワード　離散データ, 自然言語処理, ディリクレ分布, LDA, DM, ナイーブベイズ

Bayesian approaches in Natural Language Processing

Daichi MOCHIHASHI†

† ATR Spoken Language Communication Research Laboratories /
National Institute of Information and Communications Technology
619-0288 2-2-2, Keihanna Science City, Kyoto, Japan
E-mail: †daichi.mochihashi@atr.jp

Abstract　This paper overviews Bayesian approaches in natural language processing that are becoming prominent. Without any knowledge of natural language processing, Bayesian approaches to both discriminative learning and generative modeling are described. Especially, naïve bayes and its full unsupervised Bayesian modeling, DM, and LDA are developed. These Bayesian approaches permit interesting joint modeling with continuous data, such as images and musics.

Key words　Discrete data, Natural language processing, Dirichlet distribution, LDA, DM, Naive Bayes

1. はじめに

英語, 日本語, 中国語, ‥などの自然言語を統計的機械学習の枠組みで扱う自然言語処理 [1] において近年, ベイズ的アプローチが注目を集めている. 下で述べるように, 統計的自然言語処理 (以下, 単に自然言語処理) は本来の対象である自然言語だけでなく, 高次元かつ離散的な時系列データとその相関を取り扱う一般的方法として, 広い応用範囲を持っており, 自然言語処理の発展が他分野に資するところは大きいと考えられる. また, 言語の確率的生成モデルに基づくベイズ的アプローチは, 識別学習などと比べ, 音声, 画像などとの結合モデルが構成しやすいという特徴を持ち, いくつかのモデルが提案されている.

そこで本稿ではまず, 自然言語処理の概要とその主な問題について述べ, 言語以外の通常の統計的機械学習と異なる特徴について説明する. 自然言語処理のアプローチには大きく分けて, 識別モデルと生成モデル (教師あり学習と教師なし学習) があり, どちらにも最近になってベイズ推定が適用されている. ただし, 識別モデルにおけるベイズ推定はごく最近になって適用され始めているもので, 階層ベイズ等のベイズ学習の特徴をまだ充分生かしたものにはなっていないため, 概容を紹介するにとどめ, 教師なし学習すなわち, 言語の確率的生成モデルに焦点を当てて, スパムのフィルタリングで知られるナイーブベイズの拡張を含む, いくつかの有名なモデルとその応用について概説する.

また, 自然言語のベイズモデルにおける最近の発展について触れ, 最後に識別モデルとの融合, 連続系との結合を含めた未来像について概観する.

2. 自然言語処理とは

2.1 自然言語処理の主な問題

自然言語処理とは, 自然言語を計算機によって扱う方法全般をさすが, Web などによるデータ量および計算資源の爆発的な増大に伴い, 近年では統計的機械学習による自然言語処理が主流となっており, ベイズ学習によるアプローチもその中に含ま

れている.

自然言語処理には多様な分野があるが, 主なものとして(注1)

- 統計的機械翻訳
- 構文解析, または係り受け解析
- 形態素解析
- 文書要約, 意見抽出, 質問応答
- 対話や独話, テキストのモデル化

などがあげられる. このうち, 統計的機械翻訳は翻訳モデル (翻訳確率を扱う) と言語モデル (言語としての自然さを扱う) に分けられ, それぞれが研究対象となっている. また, 上記のような狭義の自然言語処理だけでなく, Google などに代表される情報検索およびリンク解析は自然言語処理の問題に還元することができ, きわめて重要な研究対象である [2] [3].

他にも, 個人の行動履歴から, 集団全体のデータをもとに嗜好を予測し, 離散的な対象物 (商品など) の推薦を行う協調フィルタリング [4] は直接自然言語ではないものの, 下に述べるような自然言語と同じ特徴を備えており, 自然言語処理の枠組みの中で考えることができる.

2.2 自然言語処理の特徴

言語以外の通常の機械学習においては, データとして実数値のベクトルを考えることが多いが(注2), 自然言語を統計的機械学習の枠組みでとらえる場合, 以下のような特徴を持っており, 特別な取り扱いを必要とする.

- 超高次元, かつ疎な離散時系列データ
- 次元の間に高い相関がある
- データの階層的な構造
- 内観により, 豊富なタグを与えることができる.

言語データは通常, 単語列からなるものと考えてよいが, 可能な単語の種類は膨大であり, 原理的に可算無限個存在する. 音声認識や機械翻訳などでは普通, 高頻度の数万〜数 10 万語が使われている.(注3)

単語はシンボルの組み合わせからなっているが, アルファベットが高々数十文字しかない英語の場合と異なり, 漢字の場合はそれ自体も巨大な数があり(注4), 康熙字典での漢字の総収録数は 49,000 字に達している. これらの記号はきわめて高い相関を持っており, その記号列が, 文字→単語→構文木→文→文書→文書集合のような形に階層的に構成されることで言語データが成立している.

これ以外にも言語には多様な構造が存在するため, 内観によって豊富なタグ付けを行って構造を指定することができる. 典型的なものは動詞, 名詞, 形容詞, ··· などの形態素タグ, および構文解析情報であり, 最近では言葉や句の感情的な印象を正/負に分類することが注目されている [5]. 未知データに対してこれらを正しく予測することが, 研究の大きな目標の一つとなっている.

(注1): 自然言語処理の最も主要な国際会議である ACL の Call for Papers に, 取り扱う話題の主な例が挙げられている. 今年度 (2006 年) の場合は, http://www.acl2006.mq.edu.au/cfp/papers/ を参照されたい.

(注2): 音声や画像などはこのようなデータと考えることができる.

(注3): 最近公開された Google の Web ページクロールによる 5-gram データでは, 200 回以上出現した, 合計 1,365 万語の異なり語が使われている. (http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html)

(注4): 漢字自体も部首の組み合わせからなるが, 構造は単語のように一次元ではない.

3. 識別モデルのベイズ学習

上で述べたように, 人手によりタグの付いた学習データをもとに未知データのタグや構造を予測する識別モデルは, 自然言語処理の大きな方法の一つとなっている [6]. 以下では, 代表的な識別タスクである形態素などのタギング, PCFG による構文解析, および係り受け解析について, 最近提案されているベイズ的な推定法を概観する.

3.1 タギング

形態素や固有名詞など言語データのタグ付けは, タグを内部状態と考えた HMM によって捉えることもできるが [7], HMM ではタグの予測に使える素性が限られるため, 近年は最大エントロピー法の系列データへの拡張である Conditional Random Fields [8] およびその拡張が標準的に用いられるようになっている. 従来, この学習には最尤推定 [8] および MAP 推定が用いられてきたが, Qi ら (2005) で Power EP 法 [9] を用いたベイズ推定が提案され, 最尤推定および MAP 推定より高い性能を持つことが示されている [10].

3.2 構文解析

栗原ら [11] は, 確率文脈自由文法 (PCFG) による構文解析にベイズ推定を用いる方法を示した. PCFG のパラメータ推定では, 通常 EM アルゴリズムが用いられるが, 適用される例の少ない規則については過学習が起こる危険性がある. [11] では, 確率テーブルの事前分布をディリクレ分布とし, 変分ベイズ法による推定を動的計画法を用いて行い, 従来法より高い性能を報告している.

3.3 係り受け (依存構造) 解析

文法を必要としない係り受け解析は近年, 日本語だけでなく英語を始めとする多言語に適用され, 注目を集めている [12]. この推定には係る/係らないの分類器として, SVM などマージン最大化に基づく分類器や最大エントロピー法が用いられてきたが, Oliver ら (2006) は分類器として, オンライン推定が可能な Bayes Point Machines [13] を用い, 簡単なアルゴリズムでマージン最大化による分類とほぼ同等の性能を達成している. [14]

しかしながら, 識別モデルにおけるこれらの方法は基本的に, モデルの推定法にベイズ推定を用いることでよりロバストな推定値を得るものであり, 適切な事前分布の設定や, 隠れ変数を含む階層ベイズモデルのような形でベイズ的なモデリングの長所を生かしたものにはまだなっていない. このため, 上では概要のみを示した. これに対し, 言語の生成モデルにおいては近年, 豊富なベイズ的アプローチが展開されている.

4. 生成モデルのベイズ学習

生成モデルによる自然言語の教師なし学習では, 識別モデルにおけるタグにあたるものを隠れ変数として推定を行うことが基本となる. 形態素解析や構文解析のようにほぼ正解が一意に決定できる問題と異なり, 「言葉の意味」に当たるものをモデル化する場合には明示的な正解が存在しないため, このようなアプローチが有用である.

このようなモデルとして最初に提案されたものが, PLSI [15] をベイズ化した LDA [16] であり, その後別な観点から, 日本で DM [17] が提案されている. DM はよく知られたナイーブベイズ法 [18] の拡張と考えられるため, 以下ではまずナイーブベイズ, DM, LDA の順に導入し, それらの応用例について紹介する.

4.1 Naive Bayes 法

スパムメールの分類法として, ナイーブベイズ法が有名となった [19]. この方法では, 訓練データの各文書 d にカテゴリ c (スパム=1, 非スパム=0 など) を与え, 未知文書 $d = v_1 v_2 \cdots v_n$ に対して, c の事後確率を

$$p(c|d) \propto p(c)p(d|c) \propto p(c) \prod_{v \in d} p(v|c)^{n(d,v)} \tag{1}$$

のように推測する. ここで v は単語であり, $n(d,v)$ は文書 d 中での v の出現回数を表す. $p(c)$ および $p(v|c)$ がナイーブベイズ法のパラメータである. $p(v|c)$ は単純には, 最尤推定

$$p(v|c) \propto \textstyle\sum_{d \in c} n(d,v) \tag{2}$$

によって計算すればよいが, この値は極めてスパースであり, 一つでも訓練データ中でカテゴリ c の文書に現れなかった v があると $p(v|c) = 0$ となり, (1) 式全体が 0 になってしまう. このため, 実際には (2) 式のカウント $n(d,v)$ に小さな値 δ を加えること (ラプラススムージング) などが行われるが, 単語の頻度にかかわらず同じ値 δ を加えることには疑問がある上に, 最適な δ の値もモデルからは求めることができないという大きな問題がある.

4.2 Dirichlet Mixtures (DM)

これに対し, 確率分布 $\mathbf{p} = \{p(v|c)\}$ そのものを確率変数と考え, $\boldsymbol{\alpha}_c = (\alpha_{c1} \ldots \alpha_{cV})$ をパラメータとするディリクレ事前分布

$$\mathbf{p} \,|\, c \sim \mathrm{Dir}(\boldsymbol{\alpha}_c) \tag{3}$$

に従っているとすると, (1) 式は

$$p(c|d) \propto p(c)p(d|c) = p(c) \int p(d|\mathbf{p})p(\mathbf{p}|c)d\mathbf{p} \tag{4}$$

$$= p(c) \frac{\Gamma(\sum_v \alpha_{cv})}{\Gamma(\sum_v \alpha_{cv} + n)} \prod_v \frac{\Gamma(\alpha_{cv} + n(d,v))}{\Gamma(\alpha_{cv})} \tag{5}$$

と, $\mathbf{p}$ を積分消去した形で表すことができる. ここで

$$p(d|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_v \alpha_v)}{\Gamma(\sum_v \alpha_v + n)} \prod_v \frac{\Gamma(\alpha_v + n(d,v))}{\Gamma(\alpha_v)} \tag{6}$$

は Polya 分布とよばれ, 離散データの表現に適した特徴をもつ [20].

上では訓練データにおいてカテゴリ c が既知であるとしたが, これを未知とした場合が Dirichlet Mixtures (DM) である. したがって DM は Polya 分布を用いた通常の混合モデル (混合 Polya 分布) と考えることができ, EM アルゴリズムと Newton 法を併用することで, パラメータ α_{cv} および $p(c)$ $(c = 1 \cdots C, v = 1 \cdots V)$ を推定することができる [17]. 通常, 単語の総数 V は数万, カテゴリの総数 C は数 100 程度を考えるため, パラメータ α_{cv} の数は数 100 万個程度になる.

このとき, 式 (5) の形から, α_{cv} がナイーブベイズ法における δ に相当することに注意されたい. ナイーブベイズ法において δ の値は一様で, 適当に決めるしかなかったが, DM においてはこの値は各単語ごとに EM の中で自動的に最適化され, 適切なスムージングを与える. さらに α_{cv} 自体についても階層的に事前分布を与えることで, よりロバストなベイズ推定を行う方法も示されている [21].

さて, DM およびナイーブベイズ法においては, ある文書全体が 1 つの隠れたカテゴリ (トピック) に属しているとしたが, 文書には科学と政治, 音楽と地理情報など, 様々な話題が同時に含まれる可能性がある. これをモデル化するのが下の LDA である.

4.3 Latent Dirichlet Allocation (LDA)

LDA [16] では, 各文書には隠れた確率的なトピック分布 $\boldsymbol{\theta} = (p(c_1), \ldots, p(c_K))$ があり, 文書の各単語は,

（1） トピック $c \sim \mathrm{Discrete}(\boldsymbol{\theta})$ をサンプルする.

（2） 単語 $v \sim p(v|c)$ をサンプルする.

という手続きにしたがって生成され, 単語ごとに違ったトピックを持っていると仮定する. c だけでなく, $\boldsymbol{\theta}$ も隠れ変数であり, $\boldsymbol{\theta}$ がディリクレ事前分布

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha}) \tag{7}$$

に従うとすると, 文書 $d = v_1 \cdots v_N$ の生成モデルは

$$p(d|\boldsymbol{\alpha}, \boldsymbol{\beta})$$

$$= \int p(d|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta} \tag{8}$$

$$= \int \prod_{n=1}^{N} \sum_{c_n=1}^{K} p(v_n|c_n)p(c_n|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})d\boldsymbol{\theta} \tag{9}$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \left(\prod_k \theta_k^{\alpha_k - 1} \right) \prod_{n=1}^{N} \sum_{k=1}^{K} \theta_k \beta_{k v_n} d\boldsymbol{\theta} \tag{10}$$

となる. ただし, $\boldsymbol{\beta} = \{\beta_{kv}\} = \{p(v|k)\}$ とした.

この積分は intractable であるため, 推定には変分ベイズ法などを用いるが [16](注5), 最近では事後分布をよりよく近似するため, MCMC 法を使用することが多くなっている. LDA の場合は, $\boldsymbol{\theta}$ および $\boldsymbol{\beta}$ を積分消去することで, 効率のよい Rao-Blackwellized Gibbs を構成することができる. 具体的には, 文書 d の n 番目の単語 v_{dn} の持つトピック c を, 以下の式に従ってサンプリングして更新していく.

$$c \,|\, v_{dn} \sim \frac{n^{v_{dn}}_{-dn,c} + \beta}{n^{\cdot}_{-dn,c} + V\beta} \cdot \frac{n^{d}_{-dn,c} + \alpha}{n^{d}_{-dn,\cdot} + K\alpha} \,. \tag{11}$$

ここで K はトピックの総数, V は単語の総数, α, β はそれぞれ $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ のディリクレ事前分布のハイパーパラメータである. $n^{v_{dn}}_{-dn,c}$ は注目している単語 v_{dn} がデータ全体の中でトピック c に割り当てられた総数 (dn を除く) を表し, $n^{d}_{-dn,c}$ は文書 d の中でトピック c に割り当てられた単語の総数 (dn を除く) を表す. 導出については [22] などを参照されたい.

このように LDA を用いることで, 各文書についてその持つトピック分布の事後分布 $p(\boldsymbol{\theta}|d)$ が求められるだけでなく, 文書に含まれる各単語についてもトピックの事後分布が得られる.(注6)

4.4 Particle Filtering of Context

上のモデルにおいては, 「文書」が意味のまとまりを表す単位となっていた. これを応用することで, 文脈 $\mathbf{h} = v_1 v_2 \cdots v_h$ が与えられたとき, これを仮想的な文書とみなして, 以下のように次の語を確率的に予測することができる. ここでは v は単語としているが, 商品や Web ページ, ハイパーリンクなど離散的なアイテムの予測にも同様に使えることに注意されたい.

(注5) : http://chasen.org/~daiti-m/dist/lda/で筆者がパラメータ推定のツールを公開している. また, Gibbs を用いた MATLAB ツールキットも存在している. (http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm)

(注6) : 式 (11) からわかるように, 単語の持つトピック事後分布は文書の持つトピック分布に影響されるため, 同じ単語でも文脈によって, 違ったトピック分布を持つことができる (言葉の多義性).

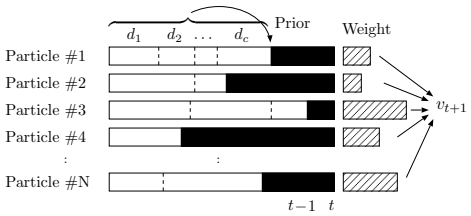

図 1　Particle Filter による文脈と変化点の推定.
Fig. 1　Particle Filter estimation of context and change points.

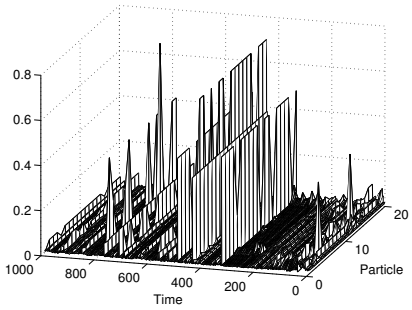

図 2　各 Particle の計算した, テキストの意味的変化点確率.
Fig. 2　Semantic change point probabilities of a text.

a)　DM の場合

$$p(v|\mathbf{h}) = \int_C p(v|\mathbf{p})p(\mathbf{p}|\mathbf{h})d\mathbf{p} \tag{12}$$

$$= \sum_{c=1} p(c|\mathbf{h}) \frac{n(\mathbf{h},v)+\alpha_{cv}}{\sum_v \big(n(\mathbf{h},v)+\alpha_{cv}\big)}, \tag{13}$$

$$p(c|\mathbf{h}) \propto p(c)\, \frac{\Gamma(\sum_v \alpha_{cv})}{\Gamma(\sum_v \alpha_{cv})+h} \prod_v \frac{\Gamma(\alpha_{cv}+n(\mathbf{h},v))}{\Gamma(\alpha_{cv})}. \tag{14}$$

b)　LDA の場合

$$p(v|\mathbf{h}) = \int \left(\sum_c p(v|c)p(c|\boldsymbol{\theta}) \right) p(\boldsymbol{\theta}|\mathbf{h})d\boldsymbol{\theta} \tag{15}$$

$$= \sum_c p(v|c) \left\langle \theta_c \right\rangle_{p(\boldsymbol{\theta}|\mathbf{h})}. \tag{16}$$

ここで $p(\boldsymbol{\theta}|\mathbf{h})$ は, 変分ベイズ EM 法などによって求める.

文脈をとらえるこの方法は, 機械翻訳や音声認識などにおいても精度向上をもたらすことが確かめられているが, $\mathbf{h}$ は単なる単語集合であり, 文脈が長くなると推定が悪化するという問題があった.

筆者はこれに対し, 隠れた多項分布 $\boldsymbol{\theta}$ (LDA はトピック分布, DM は単語分布) が確率的に変化する, 以下のような確率過程

$$\begin{cases} \boldsymbol{\theta_t} \sim \mathrm{Dir}(\boldsymbol{\alpha}) & \text{with probability } \rho \\ \quad = \boldsymbol{\theta}_{t-1} & \text{with probability } (1-\rho), \\ v_t \sim p(v|\boldsymbol{\theta}_t) \end{cases} \tag{17}$$

を考え, Particle Filter (逐次モンテカルロ法) を用いて $\boldsymbol{\theta}_t \neq \boldsymbol{\theta}_{t-1}$ となる変化点および変化の事前確率 ρ を推定し, 文脈を追跡する方法を提案した [23]. 時間 t で変化が起こったかどうかを表す二値の隠れ変数を I_t とおき, Particle Filter によって式 (17) の生成モデルを複数確率的にシミュレーションすることで, 最近の隠れた変化点以後の文脈を用いて最適な予測を得ることができる (図 1).

図 2 に, 実際のテキストにおける変化点確率を示した. このテキストでは, 話題が香港の政治問題→議会問題→中国内政→香港の経済問題と移り変わっており, 話題の変化を表す言葉 (「天安門事件」など) で変化点確率が高まっている. この方法

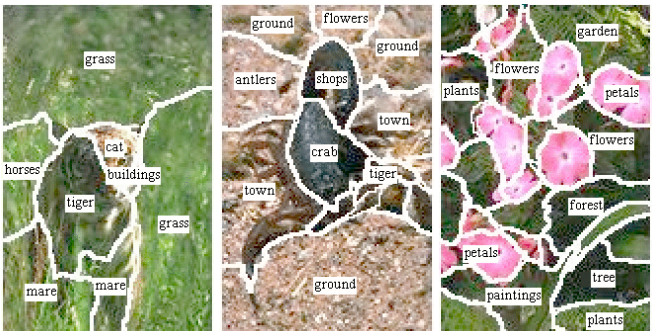

図 3　MoM-LDA [25] による画像と言葉のマッチング.
Fig. 3　Matching words and pictures using MoM-LDA [25].

は SMC によるオンライン推定だけでなく, ギブスサンプラを用いたバッチ推定によっても可能である [24].

4.5　Pictures, Songs, and Words

4.5.1　画像–言葉の結合モデル

4.3 の LDA のモデルにおいて, 観測データである単語はトピック $c \sim \boldsymbol{\theta}$ をサンプルした後, 多項分布 $v \sim p(v|c)$ に従ってサンプルされた. すなわち, これは各文書ごとに混合モデルを考えていることに相当するため, 必ずしも多項分布である必要はない. $p(\cdot|c)$ を多変量ガウス分布とし, 画素を生成するモデルを考えると, LDA を拡張することで画像とテキスト (キャプション)[注7] データの同時生成モデルを考えることができる.

[25] では, 画像・テキストの同時データ $d = \{W, B\}$ (W:画像キャプション, B:あらかじめ分割された画像領域セット) に対し, 隠れた階層クラスタリングを行って

$$p(d) = \sum_c p(c) \prod_{(w,b)\in d} \sum_l p(w,b|l,c)p(l|d) \tag{18}$$

という生成モデルを考え, LDA と同様にベイズ化して変分法によりパラメータを推定する. ここで (w,b) の対応は EM の各ステップの中で,

$$p(w,b) \simeq \sum_c p(c) \sum_l p(w,b|l,c)p(l|d) \tag{19}$$

を最大化するように対応づけを行う. l は階層クラスタリングの階層であり, 上位ノードほど一般的な画像 (空, 地面など) と単語を生成する. キャプション付き Corel 画像データベースから, この方法で画像領域と言葉の対応をとった例 [25] を図 3 に示す.

4.5.2　音楽–言葉の結合モデル

言葉と同時に使われるものに音楽 (歌) がある。[26] の興味深い研究では, 曲を音符・休符の間の 1 次マルコフ過程 (音符バイグラム) として近似し, 曲 k での音符 $i \to j$ の頻度テーブル M_{ijk} と, 曲 k の単語テーブル T_{wk} からなるデータ $\mathbf{X}_k = \{M_{ijk}, T_{wk}\}$ の確率を

$$p(\mathbf{X}_k|\boldsymbol{\theta}) = \sum_c p(c) \left[\prod_j p(j|c)^{I_j(M_{k0})} \prod_j \prod_i p(j|i,c)^{M_{ijk}} \prod_{v\in T_k} p(v|c)^{T_{wk}} \right] \tag{20}$$

のようにモデル化した.

ここで $I_j(M_{k0})$ は曲 k が音符 j で始まっているとき 1, それ

(注7)：Web サイトなどにおける多くの画像は, 関連するテキストと同時に使われている. また, 動画においては普通, 音声と時間的に対応がとれている.

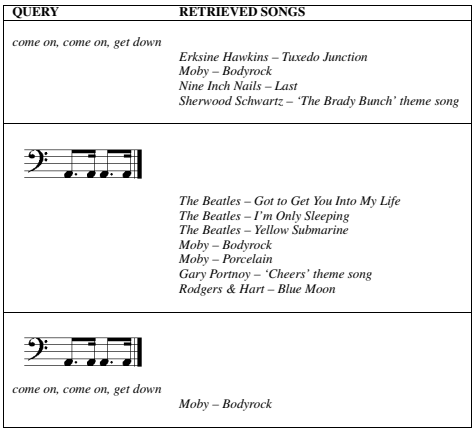

図 4　言葉と楽譜の一部を使った, 歌の確率モデルによる検索 [26].
Fig. 4　Probabilistic music retrieval from words and/or passages.

以外は 0 を返す関数である. 曲 k のクラスタリング $p(c|k)$ とモデルパラメータ $p(j|c), p(j|i,c), p(v|c)$ は EM アルゴリズムによって求めることができ [26], 図 4 のように言葉, 曲の一部, またはその両方から最も確率の高い曲を計算することができる (図 4 では両方用いた時, 正しい曲が選ばれているのに注意).

さらにこのモデルは (CD のジャケット) 画像を用いた音楽–言葉–画像の同時モデルに拡張されている [27].

5. 自然言語のベイズモデルの最近の発展

これまで自然言語処理における主なベイズ的なモデルについて見てきた. これらは精巧なモデルではあるものの, 実際にはまだいくつかの制約を持っている.

一つは, LDA や DM などのモデルが, いわゆる Bag of Words すなわちユニグラム [1] のモデルであり, 本来の時系列データの性質を充分表現していないことである. このため, これらを音声認識などで使われる, 単語の n-gram モデル(注8)に適用する際には, そのユニグラム分布のみを後付けで入れ替えるような方法が取られてきた. しかしながら, Teh(2006) [28] においてこのような n-gram モデルが階層 Poisson-Dirichlet 過程とよばれるノンパラメトリックな確率過程によって記述でき, 従来ヒューリスティックに行われてきた推定法とほぼ同等の性能を持つことが示され, n-gram モデルのような離散系列のベイズ的な取り扱いへの道が開かれている.

[28] のようなノンパラメトリックベイズ法は, (11) 式における隠れトピック数 K のようなモデル次元をデータから自動的に推論する方法として, 自然言語処理に限らず統計的機械学習全般において最近きわめて注目されている.

また, 自然言語処理におけるベイズ的な方法では, 離散分布 $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_K]$ の事前分布として単体上のディリクレ分布

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1} \tag{21}$$

が使われてきたが, この分布は各次元への分散がすべて等しいという問題が以前から指摘されてきた [16].(注9) 自然言語処理においては離散分布の各次元は単語やトピックであり, 高次元で, 互いにきわめて高い相関を持っていると考えられるため, これは大きな問題である. [29] では, LDA において $\boldsymbol{\theta}$ をロジスティック変換によって

$$\log \theta_i = \eta_i - \log \textstyle\sum_j \exp(\eta_j) \tag{22}$$

$$\boldsymbol{\eta} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{23}$$

と正規分布に対応づけ, Taylor 展開を使用して変分ベイズ法でその平均 $\boldsymbol{\mu}$ および分散–共分散行列 $\boldsymbol{\Sigma}$ を求める方法が提案されている.(注10)

ディリクレ分布の直接的な拡張として, Polya Trees [30] を用いる方法もあるが, この方法は次元を階層的にハードクラスタリングする必要があるため, 過学習を招きやすく [31], 自然言語処理のように高次元な問題にいかに適用するかは, まだ研究を要する [32].

6. 自然言語のベイズモデルの未来

これまでみてきたように, 自然言語処理においてベイズ統計的アプローチは多くの利点を持っている. その主な理由は,

- 隠れ変数をモデルに含めることができる.
- 連続的な対象を扱うことができる.
- パラメータの過学習を自然に防ぐことができる.

などであるように思われる.

観測される自然言語のデータは離散であるが, その裏にディリクレ分布のような連続的な事前分布を考えることで, 離散的な対象をより適切に扱うことが可能になる. また, 音声や画像のような他のモデルとの結合モデルを自然に考えることができる. 4.5 節で紹介したようなモデルをさらに深めることで, ロボティクスのような分野への適用も期待される.

最近になってブログの爆発的な流行から, ブログの分析が注目されているが, 通常のテキストと異なるブログの特徴の一つは, テキストに時間や場所が付加されたり, 含意されていることである. 与えられたテキストに対し, その描写している時間を識別モデルの枠組で分類するアプローチもあるが [33], 時間や場所は本来連続であり, またそれらを必ずしも含意しないテキストもあるため, このような推定問題にはベイズ的アプローチが有用であろうと思われる.(注11)

5 節で述べたように, 現在の自然言語のベイズモデルの問題は, 構造的データを充分モデル化し切れていない点にある. 例えば, 係り受け構造の生成モデルはまだ存在していない. しかし, 必ずしも全ての自然言語データに生成モデルを準備する必要があるとは限らず, 本稿で紹介したような, ベイズ的な識別モデルと生成モデルを融合する方法が今後模索されるとよいと考えている.

自然言語処理技術の総大成としての統計的機械翻訳(注12) はまだベイズ化されておらず, 決定的な最適化や探索が用いられているが, 言語の木構造を扱うことのできるベイズ統計的アプローチが開発されれば, 統計的機械翻訳も無数のサンプリングによって翻訳文をより適切に生成できる日が来るかもしれない.

(注8) : p(relinquish|he would) のような, 単語の連鎖確率を与える.
(注9) : 概念的には, これはベクトル空間において等分散の高次元ガウス分布を考えていることにほぼ等しい.
(注10) : トピックの場合には次元数が数 100 程度しかないため, この方法が適用できるが, 単語の分布では次元数が数万を超え, 共分散行列 $\boldsymbol{\Sigma}$ の直接的な推定は非現実的である.
(注11) : 時間は一周回るともとに戻ってくる性質があるため, von Mises-Fisher 分布のような分布の, テキストからの推定問題になると思われる.
(注12) : ここでいう統計的機械翻訳とはいわゆる翻訳だけでなく, 同言語内翻訳 (言い換え), 文書要約など, 言語の Noisy Channel モデルとしての基礎技術という意味を持っている.

文　献

[1] C. D. Manning and H. Schütze: “Foundations of Statistical Natural Language Processing”, MIT Press (1999).

[2] A. Berger and J. Lafferty: “Information Retrieval as Statistical Translation”, Proc. of SIGIR 1999, pp. 222–229 (1999).

[3] D. Cohn and T. Hofmann: “The Missing Link: a probabilistic model of document content and hypertext connectivity”, NIPS 2001 (2001).

[4] J. S. Breese, D. Heckerman and C. Kadie: “Empirical Analysis of Predictive Algorithms for Collaborative Filtering”, UAI 1998, pp. 43–52 (1998).

[5] H. Takamura, T. Inui and M. Okumura: “Extracting Semantic Orientations of Words using Spin Model”, Proc. of ACL 2005, pp. 133–140 (2005).

[6] 鹿島久嗣, 坪井祐太, 工藤拓：“言語処理における識別モデルの発展 – HMM から CRF まで –”, 言語処理学会第 12 回年次大会 (NLP2006) チュートリアル (2006).

[7] M. Asahara and Y. Matsumoto: “Extended Models and Tools for High-performance Part-of-Speech Tagger”, COLING 2000, pp. 21–27 (2000).

[8] J. Lafferty, A. McCallum and F. Pereira: “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, Proc. of ICML 2001, pp. 282–289 (2001).

[9] T. P. Minka: “Power EP”, Technical Report MSR–TR–2004–149, Microsoft Research Cambridge (2004). ftp://ftp.research.microsoft.com/pub/tr/TR-2004-149.pdf.

[10] Y. Qi, M. Szummer and T. P. Minka: “Bayesian Conditional Random Fields”, Proc. of AISTATS 2005 (2005).

[11] 栗原賢一, 亀谷由隆, 佐藤泰介：“動的計画法に基づく確率文脈自由文法の変分ベイズ法”, 情報処理学会研究報告 NL-159, pp. 209–214 (2004).

[12] “CoNLL-X Shared Task: Multi-lingual Dependency Parsing” (2006). http://nextens.uvt.nl/~conll/.

[13] R. Herbrich, T. Graepel and C. Campbell: “Bayes Point Machines”, Journal of Machine Learning Research, **1**, pp. 245–279 (2001).

[14] S. Corston-Oliver, A. Aue, K. Duh and E. Ringger: “Multilingual Dependency Parsing using Bayes Point Machines”, Proc. of HLT-NAACL 2006, pp. 160–167 (2006).

[15] T. Hofmann: “Probabilistic Latent Semantic Indexing”, Proc. of SIGIR ’99, pp. 50–57 (1999).

[16] D. M. Blei, A. Y. Ng and M. I. Jordan: “Latent Dirichlet Allocation”, Journal of Machine Learning Research, **3**, pp. 993–1022 (2003).

[17] 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, pp. 29–34 (2003).

[18] A. McCallum and K. Nigam: “A Comparison of Event Models for Naive Bayes Text Classification”, AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41–48 (1998).

[19] P. Graham: “A Plan for Spam” (2002). http://www.paulgraham.com/spam.html.

[20] T. P. Minka: “Estimating a Dirichlet distribution” (2000). http://www.stat.cmu.edu/~minka/papers/dirichlet/.

[21] 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, pp. 1–6 (2004).

[22] T. L. Griffiths and M. Steyvers: “Finding scientific topics”, PNAS, **101**, pp. 5228–5235 (2004).

[23] D. Mochihashi and Y. Matsumoto: “Context as Filtering”, NIPS 2005 (2005).

[24] 持橋大地, 菊井玄一郎：“Gibbs Sampling による確率的テキスト分割と複数観測への拡張”, 言語処理学会年次大会 2006, pp. 212–215 (2006).

[25] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei and M. I. Jordan: “Matching Words and Pictures”, Journal of Machine Learning Research, **3**, pp. 1107–1135 (2003).

[26] E. Brochu and N. de Freitas: ““Name That Song!”: A Probabilistic Approach to Querying on Music and Text”, NIPS 2002 (2002).

[27] E. Brochu, N. de Freitas and K. Bao: “The Sound of an Album Cover: Probabilistic Multimedia and IR”, AISTATS 2003 (2003).

[28] Y. W. Teh: “A Hierarchical Bayesian Language Model Based On Pitman-Yor Processes”, Proc. of COLING/ACL 2006, pp. 985–992 (2006).

[29] D. Blei and J. Lafferty: “Correlated Topic Models”, NIPS 2005 (2005).

[30] T. Minka: “The Dirichlet-tree distribution” (1999). http://research.microsoft.com/~minka/papers/dirichlet/minka-dirtree.pdf.

[31] 山本 (2005). personal communication.

[32] W. Li and A. McCallum: “Pachinko allocation: DAG-structured mixture models of topic correlations”, ICML 2006, pp. 577–584 (2006).

[33] 野呂太一, 乾孝司, 高村大也, 奥村学：“イベントの生起時間帯判定”, 情報処理学会研究報告 NL-117, pp. 7–14 (2005).