生きたことばをモデル化する
自然言語処理と数学の接点

持橋大地
● NTT コミュニケーション科学基礎研究所

ことばを扱う学問は古典的には言語学であり，そこでは言語学者の経験と主観によって生み出された仮説を積み重ね，また反例を挙げて新説を生み出すことで研究が蓄積されてきた. これに対し, ことばを統計的に考える分野は計算言語学，または工学的な立場からは自然言語処理とよばれており，最近の電子テキストの増大とその処理の必要性によって，急速に研究が進んでいる分野である．この分野は言語学の一部ともいえるが，純粋に客観的なデータから，統計的・数学的なモデル化と大規模な実験的検証を行う点が従来の言語学と異なっている.[1)] 言語を統計的にとらえることによって，複雑で厖大な言語現象を計算機で自動的にモデル化できるとともに，規則ではとらえきれない曖昧性や例外，文脈依存性を数学的に適切に扱うことが可能になる.

言語の統計モデル

客観的にみると, 言語とは記号列だと考えることができる. 細かくみるとそれは文字からなっているが, ここでは英語のように, 言語は単語からなっているとして話を進めよう.

すぐ気づくことは，単語の頻度には大きな偏りがあるということである．表 1 に, 宮沢賢治『銀河鉄道の夜』におけることばの出現頻度を数えたテーブルを, 図 1 に順位–頻度を両対数でプロットしたグラフを示す．このように, 順位と頻度が反比例関係にあることは Zipf の法則といわれ, 1930 年代に発見された基本的事実の一つであり, 近年は言語を超えて, 自然界の多くの離散的現象に共通する Power law として知られるようになってきている [1].

[1)] 以前は計算機や電子テキストは存在しないか，速度が充分でなく，以下で紹介するような大規模統計モデルの計算は不可能であった．個人的には筆者は，これは現代的な意味での理論言語学の王道 (の少なくとも一つ) だと考えている.

順位	単語 w	$n(w)$	$p(w)$
1	の	1266	0.055005
2	。	1120	0.048662
3	、	988	0.042927
4	た	951	0.041319
5	て	884	0.038408
18	ジョバンニ	189	0.008212
34	カムパネルラ	101	0.004388
104	風	26	0.001130
104	天の川	26	0.001130
482	橙	5	0.000217
482	ボート	5	0.000217
482	ステーション	5	0.000217
1307	燈火	1	0.000043
1307	天蚕	1	0.000043
1307	鶴嘴	1	0.000043

表 1: 『銀河鉄道の夜』における単語の頻度と順位.

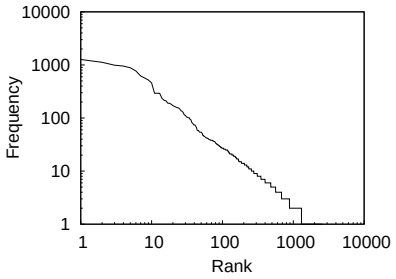
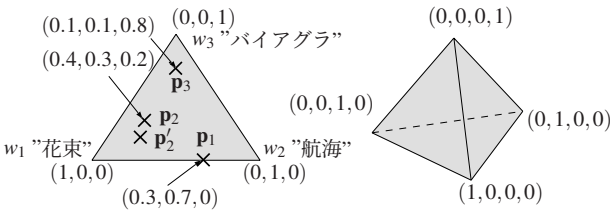


図 1: 『銀河鉄道の夜』での単語順位－頻度の両対数プロット．順位×頻度がほぼ一定となる Power law が，右下がりの直線として現れている.

さて, では表 1 のような単語の相対頻度は, どんな本やメールをもってきても常に同じなのだろうか. 明らかにそうではなく，上位は大まかに同じでも,「風」「ステーション」など中位〜下位の語は話題や内容によって大きく異なってくるはずである．実際，広告メール (スパム) の自動判別は，このような違いをもとに，一般に以下で述べるような確率モデルを使って行われている.

いま，表 1 の頻度を確率にかきかえ，全体で N 語の文章の中で，単語 i が n_i 回現れたとすると，その確率は単純には,

$$p_i = \frac{n_i}{N} \tag{1}$$



(a) 2 次元単体 (正三角形)　　(b) 3 次元単体 (正四面体)

図 2: 単体と確率分布.

と考えてよい．このとき，各単語の出現確率を V 次元 (V は語彙数) のベクトルで表した $\mathbf{p}=(p_1,p_2,\cdots,p_V)$ は $\forall i\ p_i\geq 0, \sum_i p_i=1$ をみたす確率分布であり，高校のベクトルの授業を思い出してみるとわかるように，これは単体 (Simplex) とよばれる，正三角形や正四面体を一般化した $(V-1)$ 次元の図形の内部に含まれている (図 2).

たとえば，語彙が $(w_1,w_2,w_3)=$ ("花束","航海", "バイアグラ″) の 3 個しかない ($V=3$) としよう．単語の生起分布には $\mathbf{p_1}=(0.3,0.7,0)$, $\mathbf{p_2}=(0.4,0.3,0.2)$, $\mathbf{p_3}=(0.1,0.1,0.8),\cdots$ のように無限の可能性があるが，これらはすべて単体の内部に，↑で示したように含まれている．このとき，広告メールは $\mathbf{p}_3$ のような確率分布から，通常のメールは $\mathbf{p}_1,\mathbf{p}_2$ のような確率分布から生成されたと考えられる．

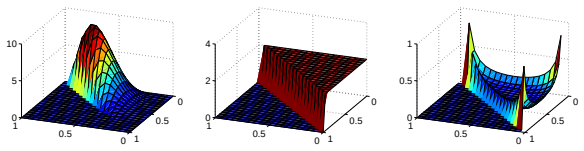
ただし，高々数百語のテキストを使った式 (1) による $\mathbf{p}$ の推定が唯一の分布であるかには疑問がある．メールは $\mathbf{p}_2$ から生成されたのかもしれないし，それから微妙にずれた $\mathbf{p}_2'$ から生成されたのかもしれない．

このような不確定性を表現するには，$\mathbf{p}$ 自体の場所についての確率分布が必要になる．$\mathbf{p}$ は確率分布であったから，これは確率分布の確率分布となり，その最も簡単なものとして，ディリクレ分布

$$p(\mathbf{p})=\mathrm{Dir}(\mathbf{p}|\alpha)=\frac{\Gamma(\sum_i\alpha_i)}{\prod_i\Gamma(\alpha_i)}\prod_{i=1}^{V}p_i^{\alpha_i-1} \tag{2}$$

を考えることができる．ガンマ関数 $\Gamma(x)$ の現れる分数の部分は正規化定数なので無視してよいが，この分布はパラメータ $\alpha=(\alpha_1,\cdots,\alpha_V)$ の値によって，図 3 のようにさまざまな形をとる．

このように，単語の確率分布自体を (たとえば) ディリクレ分布から生まれたものと考えると，まず図 4(a) のように様々な $\mathbf{p}$ を生成するなだらかなディリクレ分布があり，ある文章 $\mathbf{w}=w_1w_2\cdots w_N$ は最も単純には，次のようにして生まれたと想像できる．



(a) Dir([5 2 2])　　(b) Dir([1 1 1])　　(c) Dir([0.1 0.1 0.1])

図 3: 様々なディリクレ分布とパラメータ $\boldsymbol{\alpha}$. (a) ある平均値をもつ分布，(b) 一様分布，(c) 特定の単語への偏り，などを表現できる.

(i) $\mathbf{p}\sim\mathrm{Dir}(\mathbf{p}|\alpha)$ を生成.

(ii) 単語 $w_i\sim\mathbf{p}$ $(i=1\cdots N)$ を生成.

ここで ∼ とは,「∼の確率分布に従って」という意味である．このとき $\mathbf{w}$ の確率は，様々な $\mathbf{p}$ の可能性について積分を行って期待値を計算すると，

$$p(\mathbf{w})=\int p(\mathbf{w}|\mathbf{p})p(\mathbf{p}|\alpha)d\mathbf{p} \tag{3}$$

$$=\int\prod_{i=1}^{V}p_i^{n_i}\cdot\frac{\Gamma(\sum_i\alpha_i)}{\prod_i\Gamma(\alpha_i)}\prod_{i=1}^{V}p_i^{\alpha_i-1}d\mathbf{p} \tag{4}$$

$$=\frac{\Gamma(\sum_i\alpha_i)}{\Gamma(N+\sum_i\alpha_i)}\prod_{i=1}^{V}\frac{\Gamma(n_i+\alpha_i)}{\Gamma(\alpha_i)} \tag{5}$$

と表すことができる [2)]．ここで n_i は単語 i が $\mathbf{w}$ の中に現れた回数であり，$\int d\mathbf{p}$ は $\int_0^1\cdots\int_0^1 dp_1\cdots dp_V$ を意味する．

多数の文章 $\mathbf{w}_1\cdots\mathbf{w}_D$ に関するこの確率の積 $\prod_{d=1}^{D}p(\mathbf{w}_d)$ は α に関して凸であり，ニュートン法を用いて，データの確率を最大にする事前分布のパラメータ α を求めることができる．

逆に $\mathbf{w}$ が与えられれば，それを生んだ $\mathbf{p}$ の確率分布は，ベイズの定理から式 (2) を用いて，図 4(b) のようなディリクレ事後分布

$$p(\mathbf{p}|\mathbf{w})\propto p(\mathbf{w}|\mathbf{p})p(\mathbf{p}) \tag{6}$$

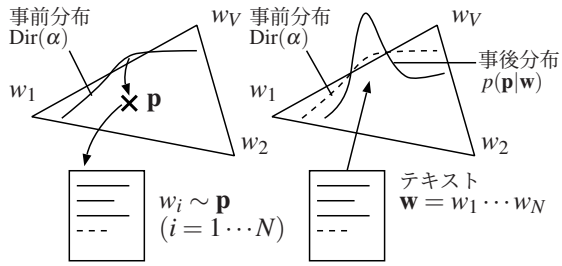
$$=\prod_{i=1}^{V}p_i^{n_i}\cdot\frac{\Gamma(\sum_i\alpha_i)}{\prod_i\Gamma(\alpha_i)}\prod_{i=1}^{V}p_i^{\alpha_i-1} \tag{7}$$

と推定でき，その期待値は

$$E[p_i|\mathbf{w}]=\frac{n_i+\alpha_i}{N+\alpha}\quad(\alpha=\sum_i\alpha_i) \tag{8}$$

となる．これは式 (1) と似ているが，文章にたまたま出現しなかった語 ($n_i=0$) にも,「事前確率」$\alpha_i/(N+\alpha)$ が割り当てられていることに注意されたい．つまり，これは最初の式 (1) と違い，どんな語も，式 (8) に従った確率で文章に含まれる可能性があるという，より自然なモデルになっている．

[2)] この分布は Polya 分布とよばれている.『数学セミナー』1993 年 10 月号 (p.32) の特集で示されている実際の適合度の良さは，このような生成モデルから説明することができる.



(a) **p** とテキスト **w** の生成　(b) **p** の事後分布の推定

図 4: Polya 分布の生成モデルとベイズ推定.

無限語彙モデル

上の議論では語彙の数 V は固定だとしていたが，落ち着いて考えてみると，言語の語彙は決して有限ではない．通常の辞書に含まれる語彙は数万語〜数十万語程度だが，現実には日々新しい語が生まれ，また古い語が忘れ去られてゆく．[3)]

このような現象をモデル化できるのが，ディリクレ過程 $\mathrm{DP}(\alpha, G_0)$ [2] と呼ばれる確率過程である．これは言語の場合はディリクレ分布を無限次元化したものと考えてよく，図 5 のように，もととなる確率分布 G_0 (基底測度とよばれる) に似た，離散確率分布 G を生成する確率過程である．G_0 が連続の場合，G は無限次元の分布となる．

G は図 5 の下のような姿をしているが，実際はこれ自体も **p** と同様，一つに決めることはできない．したがって，文章 $w_1 \cdots w_N$ が得られたとき，次の単語の分布は G の可能性をすべて考えて積分消去すると，

$$p(w|w_1 \cdots w_N) = \int p(w|G)p(G|w_1 \cdots w_N)dG \tag{9}$$

$$= \begin{cases} n_i/(N+\alpha) & (w = w_i) \\ \alpha G_0(w)/(N+\alpha) & (w : \text{新しい語}) \end{cases} \tag{10}$$

という確率をもつことが知られている．すなわち，$\mathrm{DP}(\alpha, G_0)$ では既存の単語 i が頻度 n_i に比例した確率で，未知の単語が αG_0 に比例した確率で，それぞれ現れることを意味する．

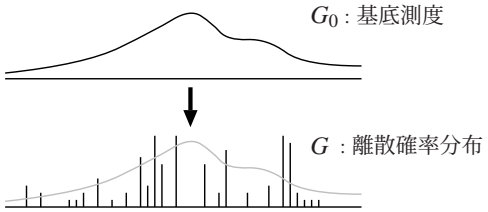


図 5: ディリクレ過程による，無限離散確率分布 G の生成．横軸が可能な単語の種類を表す.

[3)] ただし，以下のディリクレ過程で扱えるのは単語の新生のみであり，消滅過程のモデル化は現在研究の対象となっている.

α は新しい語が生まれる割合を制御する，推定可能なパラメータである．この式は Polya の壷，または集団遺伝学の分野では species sampling model として知られているモデルの特別な場合である.

もし G_0 が離散で，固定語彙 w_i についてのみ α_i/α を返すならば，これは式 (8) と等しいが，G_0 が連続ならば，どんな未知の語にも確率が与えられることに注意してほしい．このディリクレ過程から式 (10) にしたがって次々と単語を生成すると，その分布は図 6 のように，はじめに述べた Power law を再現することがわかる.

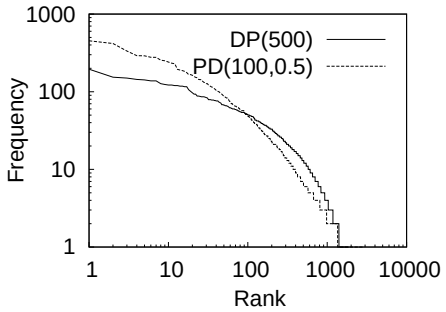


図 6: ディリクレ過程 $\mathrm{DP}(\alpha)$ とポアソン=ディリクレ過程 $\mathrm{PD}(\alpha, d)$ (次ページ) からランダムに生成した系列の順位–頻度のプロット．図 1 と同様の Power law が現れている.

nグラムモデルと無限nグラムモデル

さて，ここまでは言葉が互いに独立に生起するとしていたが，これはもちろん正しくない．たとえば，「公園 に」の次は「行く」「来る」「隣接」などが続きやすく,「言語」の次は「が」「処理」「学習」などの確率がずっと高くなるだろう．このような関係を単純化して，言葉がその前の$(n-1)$語の言葉に依存する (隣接した n 語の間の関係をとらえる) モデルは，nグラムモデルとよばれている．つまり，今までは 1 グラムのモデルを考えていたことになる.

nグラムモデルでは，文 $\mathbf{w} = w_1 w_2 \cdots w_T$ の確率は，条件つき確率の積として以下のように表される.

$$p(\mathbf{w}) = \prod_{t=1}^{T} p(w_t|w_{t-1}w_{t-2}\cdots w_1) \tag{11}$$

$$\simeq \prod_{t=1}^{T} p(w_t|\underbrace{w_{t-1}\cdots w_{t-(n-1)}}_{(n-1)\text{ 語}}). \tag{12}$$

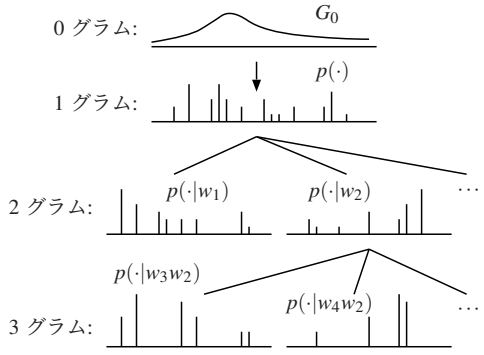


図 7: 階層ディリクレ過程による n グラム分布の生成.

ここで式 (11) では，公式 $p(x_1,x_2) = p(x_2|x_1)p(x_1)$ を繰り返し用いた．具体的には，下で説明するデータスパースネスの問題から伝統的に多く使われてきた 3 グラムの場合，文の確率は直前の 2 単語を条件に

$$p(\text{彼女 が 見る 夢}) = p(\text{彼女}) \times p(\text{が} \mid \text{彼女}) \times p(\text{見る} \mid \text{彼女 が}) \times p(\text{夢} \mid \text{が 見る}) \quad (13)$$

のように計算される．

連続する語の規則性をとらえるこのモデルは，言葉の $(n-1)$ 次のマルコフ過程であり，非常に簡単なモデルであるが，音声認識や統計的機械翻訳などで言語的に不適格な文の確率を小さくするためにきわめて有効であり，その中核の一部を構成している．

n グラムモデルでは n を増やすほど，言葉の間の関係をより精密にとらえることができる．しかし，ここでの問題は，語彙数が大きいために，単純な推定では式 (12) の条件つき確率が，n が大きくなるとほとんど 0 になってしまうことである．たとえば，有名な歌のフレーズに似ている “お魚くわえた三毛猫” は文法的には正しいが，現在この Google のカウントは 0 であり，$n()$ をカウントを表す関数として

$$p(\text{三毛猫} \mid \text{お魚 くわえた}) = \frac{n(\text{お魚 くわえた 三毛猫})}{n(\text{お魚 くわえた})} = 0 \quad (14)$$

となって，式 (12) からこのフレーズの確率が 0 と推定されてしまう．

ただし容易にわかるように，3 グラム確率 p(三毛猫 | お魚 くわえた) は 2 グラム確率 p(三毛猫 | くわえた) に似ており，それはさらに 1 グラム確率 p(三毛猫) を反映していると考えられる．このような再帰的な関係は，上のディリクレ過程を階層化した，階層ディリクレ過程 [3] によってとらえることができる．

> 「レンタ・カーは空のグラスを手にとり、蛇腹はすっかり暗くなっていた。それはまるで獲物を咀嚼しているようだった。彼は僕と同じようなものですね」と私は言った。「でもあなたはよく女の子に爪切りを買った。そしてその何かを振り払おうとしたが、今では誰にもできやしないのよ。私は長靴を棚の上を乗り越えるようにした。...

図 8: 村上春樹『世界の終りとハードボイルド・ワンダーランド』から学習した可変長 n グラムモデルによる，ランダムウォーク生成文．単語の区切りは省略している．

すなわち，図 7 のように，まず図 5 と同様に生成された 1 グラム分布 $p(\cdot)$ があり，これを基底測度 G_0 としたディリクレ過程によって 2 グラム分布 $p(\cdot|w_1)$ が生成され，さらにこれを基底測度 G_0 として 3 グラム分布 $p(\cdot|w_2w_1)$ が生成され…と漸層的に n グラム分布が生成されたと考えるわけである．

このとき，$p(w|w_2w_1)$ を求める際に $n(w_2w_1w) = 0$ であった場合，この分布が生成された親の確率分布 $p(w|w_1)$ を G_0 として式 (10) から確率を計算し，もしそこにもない場合，さらに親の確率分布 $p(w)$ を用いて…と階層をさかのぼって計算していくことで，すべての n グラム確率を計算することができる．学習テキストの各単語が実際にこの階層をどれだけ辿って生成されたのかは未知であるため，この確率計算には大規模な MCMC 法などを必要とする．

実際には言語の場合，ディリクレ過程の当てはまりは完全ではなく，その拡張である 2 パラメータ・ポアソン＝ディリクレ過程 [4] (Pitman-Yor 過程，図 6) を階層化した，階層 Pitman-Yor 過程 [5] が高い予測精度をもつことが確かめられている．

上の n グラムモデルでは文脈長は $(n-1)$ で固定されていたが，最近筆者は，これを無限可変長に拡張した [6]．式 (13) において,「夢」を予測するには実際には 1 語前の「見る」だけの文脈があればよく，一方,「日 米 首脳 会談」の最後の語「会談」の予測は，3 語前の「日」からの文脈が非常に有効である．

このように予測に適切な文脈長を隠れ変数と考えて確率モデル化すると，n グラムの n，すなわち図 7 での木の階層の深さ自体も可変とすることができ，より柔軟なモデルが得られる．図 8 に，村上春樹『世界の終りとハードボイルド・ワンダーランド』を使って学習したこのモデルから，ランダムに生成した文の一例を示す．

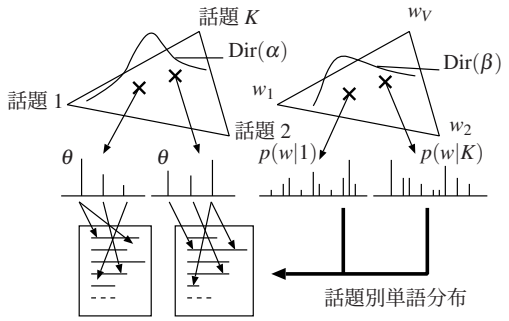


図 9: LDA による文章の生成モデル．文章ごとに隠れた話題分布 θ があり，そこからランダムに選ばれた話題をもとに，話題別単語分布からことばが生成される．

ことばの意味の統計モデル

ここまでは言語の比較的規則的な性質に着目してきたが，それでは，言葉の持つ“意味”を統計的に扱うことはできるのだろうか．

本稿の最初の例では，文章の各単語がみな同じ分布 $\mathbf{p}$ から生成されたとしていたが，これは少々単純化のしすぎだと思われる．実際のテキストには「が」「と」のような機能語,「市場価格」「ヴァイオリン」のようにその文章の話題を表す語など，意味の違う語が混ざり合っているからである．経済，芸術，スポーツ，... などの「話題」は文章によって異なり，しかも一つではなく同時に混ざり合っていると考えられるから[4]，これは図 9 のように，文章によって異なる確率分布 $\theta = (\theta_1, \cdots, \theta_K)$ で表され[5]，同様にディリクレ分布から各文章ごとに生成されていると考えるのが自然である．ここで K は潜在的な話題の総数であり，通常数 100 程度を考える[6]．

そこで，このモデル (LDA [7] とよばれる) では，各文章ごとにまず (1) 話題分布 θ を選び，次に (2) θ にしたがって話題をランダムに選び，最後に (3) その話題からことばが生成されたと考える．つまり，式で書くと，それぞれの文章 $\mathbf{w} = w_1 w_2 \cdots w_N$ は

(i) 文章のもつ話題分布 $\theta \sim \mathrm{Dir}(\alpha)$ を生成.

(ii) For $n = 1 \cdots N$,

(a) ある話題 $k_n \sim \theta$ を選択.

雪国[58]
国境の長い[58]トンネル[6]を[58]抜ける[58]と[36]雪国で[58]あった[29]。夜[36]の底が白く[58]なっ[36]た[29]。信号[6]所に[58]汽車[36]が[58]止まっ[6]た[29]。
向[36]側の[58]座席[6]から[58]娘[36]が[58]立って来て[36]、島村[54]の[58]前[29]の[58]ガラス窓[6]を[58]落し[66]た[29]。雪の冷気が流れ[58]こんだ[36]。…

図 10: 「雪国」冒頭の潜在的な話題の LDA による推定．ここでは最大事後確率の話題のみを示した．

(b) 話題 k_n から，単語 $w_n \sim p(w|k_n)$ を生成.

のように生成されたと考える．このとき，$\mathbf{w}$ の確率は

$$p(\mathbf{w}) = \int \prod_{n=1}^{N} \sum_{k=1}^{K} p(w_n|k_n) p(k_n|\theta) \cdot p(\theta|\alpha) d\theta \qquad (15)$$

$$= \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \left(\prod_k \theta_k^{\alpha_k - 1} \right) \prod_{n=1}^{N} \sum_{k=1}^{K} p(w_n|k) \theta_k d\theta \qquad (16)$$

となる．文章のもつ潜在的な話題分布 θ，各単語のもつ話題 k_n はすべて未知の確率変数であり，われわれが知っているのは単語の出現 $\mathbf{w}$ のみであることに注意されたい．さらに，話題ごとの単語生起分布 $p(w|k)$ も $\mathrm{Dir}(\beta)$ にしたがう未知の確率分布であり，どんな「話題」があるのかもすべてデータから自動的に推定することを考える．

2001 年に提案されたこのモデルは複雑で，一見解けないように見える．しかし，MCMC 法を使い巧妙な積分を行うと，各文章 d の各単語 w_{dn} のもつ話題 k は，ベイズの定理による確率

$$p(k|w_{dn}) \propto p(w_{dn}|k) p(k|d) \qquad (17)$$

$$= \frac{n^{w_{dn}}_{-dn,k} + \beta}{n^{\cdot}_{-dn,k} + V\beta} \cdot \frac{n^{d}_{-dn,k} + \alpha_k}{n^{d}_{-dn,\cdot} + \sum_k \alpha_k} \quad (k = 1 \cdots K) \qquad (18)$$

に従ってサンプリングすることができる [8]．ここで $n^{w_{dn}}_{-dn,k}$ は注目している単語 w_{dn} がデータ全体の中で話題 k に割り当てられた総数 (dn を除く) を表し，$n^{d}_{-dn,k}$ は文書 d の中で話題 k に割り当てられた単語の総数 (dn を除く) を表す[7]．

数百万語～数億語の学習テキストに対してこのサンプリングを繰り返し行うことで，各単語の生成された正しい話題と，文書の話題分布をすべて推定することができる．

[4] たとえば,「ヴァイオリンの市場価格」についての文書のような場合．本稿も，数学の話と言語の話の混合である．

[5] つまり言葉の上に，抽象的な「話題空間」を仮定している．

[6] θ が階層ディリクレ過程にしたがっていると考えると，隠れた話題の総数も可算無限個とすることができるが，ここでは簡単のため，K 個に固定して考える．

[7] $\cdot$ は，その変数について和をとることを表す.

$l(w\|58)$	単語 w
4.60517	ライラック
4.60517	雪上
4.60517	登山
4.60516	多年草
4.60515	山開き
4.60515	冬山
4.60515	岩場
4.60514	ソメイヨシノ
4.60513	岩肌
4.60511	咲き乱れ
4.60507	挿し
4.60482	雨期
4.60475	急流
4.60306	エコツーリズム
4.60279	水草
4.60260	花木
4.60249	トレッキング
4.60223	ドングリ
4.60093	湿原
4.60026	海鳥
:	:

(a) 話題 58: “自然”

$l(w\|6)$	単語 w
4.60517	北千住
4.60517	近畿運輸局
4.60517	橋げた
4.60517	山陽新幹線
4.60517	車両
4.60517	都営地下鉄
4.60517	総武線
4.60517	住之江公園
4.60517	車線
4.60516	関越
4.60512	ジープニー
4.60499	ローカル線
4.60446	奥屋
4.60430	運行
4.60316	停車
4.60269	番線
4.60205	横風
4.60204	脱線
4.60042	運賃
4.59881	支線
:	:

(b) 話題 6: “鉄道”

図 11: 毎日新聞テキストから学習した，LDA による話題別単語分布の特徴語．“” は筆者がつけたラベル．

図 10 に，毎日新聞 2000 年度の全文 (2887 万語) で学習したモデルをもとに，川端康成「雪国」の冒頭について推定した話題の例を示す．[8] 説明のため，ここではやや少なく $K = 100$ とした．「国境」–「雪国」–「雪」–「冷気」のような語が話題 58 に，「トンネル」–「ガラス」–「信号」–「窓」のような語が話題 6 に，共通して高い事後確率を持つことがわかる．[9] 図 11 に話題ごとの単語生起分布 $p(w|k)$ を，平均 $p(w) = 1/K \sum_{k=1}^{K} p(w|k)$ との対数比

$$l(w|k) = \log \frac{p(w|k)}{p(w)} \tag{19}$$

の上位順にソートして示す．話題分布は自動的に推定したものだが，ほぼ話題 58 が “自然” に，話題 6 が “鉄道” に関連していることがわかる．

式 (17) において，単語 w_{dn} のもつ潜在的な話題 k は，その文章のもつ文脈 (右辺第 2 項) にも依存していることに注意してほしい．すなわち，このモデルではことばの持つ多義性も，確率モデルの中で自然に表現されている[10]．

このようにして，言葉や文章の持つ “意味” を，大まかではあるが，テキストから完全に統計的に推定することができる．

[8] 連続して同じ話題となった場合は，まとめて表記した．
[9] 実際には，各単語は複数の話題に確率的に所属しているため，より微妙な意味が表現されていることに注意されたい．
[10] 「体」や「群」のように複数の話題に違った意味で現れる語の場合，文脈に応じて，どれかの話題に高い事後確率をもつ．

おわりに

自然言語の確率モデルについて，最新の研究の一端を紹介した．複雑かつ曖昧な自然言語を確率的にとらえることによって，ことばの持つ規則性を残しながら，その個性や文脈依存性を適切に扱うことが可能になる．本文ではふれなかったが，言語には構文構造があり，これを線形な単語列から (正解なしで) 獲得することは，現在研究が進められている非常に興味深いテーマの一つである．

いっぽう，本稿の立場とは異なるが，構文解析や形態素解析などにおいて，人間の与えた「正解」データが仮定できる場合は，これらは対数線形回帰などの手法により正解を予測する工学的問題として高精度で可能となっており，自然言語処理のもう一つの大きな一翼をなしていることを最後にふれておきたい．

[1] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[2] Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[3] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *JASA*, 101(476):1566–1581, 2006.

[4] Jim Pitman and Mark Yor. The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator. *Annals of Probability*, 25(2):855–900, 1997.

[5] Yee Whye Teh. A Hierarchical Bayesian Language Model based on Pitman-Yor Processes. In *Proc. of ACL/COLING 2006*, pages 985–992, 2006.

[6] Daichi Mochihashi and Eiichiro Sumita. The Infinite Markov Model. In *NIPS 2007*, 2007.

[7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *PNAS*, 101:5228–5235, 2004.

[もちはし だいち]