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.

Infinite SCAN: 単語の意味変化と語義数の同時推定

井上　誠一† 　小町　　守† 　小木曽智信†† 　高村　大也††† 　持橋　大地††††

† 東京都立大学 システムデザイン研究科　〒191–0065 東京都日野市旭ヶ丘 6–6
†† 国立国語研究所　〒190–8561 東京都立川市緑町 10–2
††† 産業技術総合研究所　〒135–0064 東京都江東区青海 2–4–7
†††† 統計数理研究所　〒190–8562 東京都立川市緑町 10–3

あらまし　本研究では，単語の通時的な意味変化のモデル化において，ディリクレ過程をガウス確率場の上に考え，トピックモデルを組み合わせることで，単語によって異なる語義数を自動的に推定できる階層ベイズモデルを提案する．擬似データを用いた実験によって，意味変化と語義数を正しく推定できることを示した．また，実データを用いた実験においては，複数の語義を持つ解析対象の単語の意味変化の推定結果を示し，結果を分析した．
キーワード　意味変化，階層ベイズモデル，トピックモデル，ガウス確率場，ディリクレ過程

Infinite SCAN: Joint Estimation of Changes and the Number of Word Senses with Gaussian Markov Random Fields

Seiichi INOUE†, Mamoru KOMACHI†, Toshinobu OGISO††, Hiroya TAKAMURA†††, and Daichi MOCHIHASHI††††

† Graduate School of Systems Design, Tokyo Metropolitan University　6–6 Asahigaoka, Hino City, Tokyo, 191–0065 Japan
†† National Institute for Japanese Language and Linguistics　10–2 Midoricho, Tachikawa City, Tokyo 190–8561 Japan
††† National Institute of Advanced Industrial Science and Technology　2–3–26 Aomi, Koto Ward, Tokyo 135–0064 Japan
†††† National Institute of Statistical Mathematics　10–3 Midoricho, Tachikawa City, Tokyo 190–8562 Japan

Abstract　In this study, we propose a hierarchical Bayesian model that can automatically estimate the number of senses for each word by considering the Dirichlet process on the Gaussian Markov random field in the modeling of the diachronic meaning change of words. In the experiments using pseudo data, we show that proposed model can correctly estimates changes and the number of word senses. In the experiment using actual data, we show the estimated results and analiysis for multiple target words.
Key words　Meaning change, Hierarchical Bayesian model, Topic model, Gaussian Markov random field, Dirichlet process

1. はじめに

言語は動的なシステムであり，常に進化し，話者とその環境の需要に適応している [1]．全ての言語において，単語はさまざまな意味を持ち，その分布や広がりはジャンルや文脈によって異なっている．例えば，“cute” という単語は 18 世紀初頭に登場し，もともと “賢い” という意味で使われていたが，19 世紀後半に “狡猾な” という意味で使われ，現代においては，“魅力的な” という意味で使用されている [2]．英語における Corpus of Historical American English (COHA) [3] や，Google Books などのオンラインライブラリ，日本語においては，日本語歴史コーパス (CHJ) [4] といった，近年の通時コーパスの整備によって言語変化の統計的な調査が容易になった．通時的な単語の**意味変化**を自動的に捉えることは，辞書学や言語学といった分野に対する学術的な貢献のみならず，実際のアプリケーションに対しても応用が期待できる．例えば，時間が紐づいた情報を対象とした情報検索や質問応答などにおいて，文書や単語の意味表現がより正確になることにより，クエリの曖昧性解消や文書検索の精度を向上させることができる．

近年では，単語の分散表現を用いて意味変化を検出する手法が数多く提案されているが [5]〜[7]，これらは意味変化の検出はできるが，意味の趨勢や変化の様子を捉えることはできない．

年代	用例	スニペット
1853	The driver made room for the trunk on the top of the **coach**.	{driver, make, room, trunk}
1900	The chair passed the **coach**, which immediately fell in behind it, the horses proceeding at a walk.	{chair, pass, fell, horse, proceed, walk}
1949	Tell him if I start **coaching** , it'll be as a head **coach** at a top school.	{tell, start, coach, head, top, school}
2003	The head football **coach**'s absolute dictatorship of the football field was reproduced.	{head, football, absolute, dictatorship, football, field, reproduce}

図 1　SCAN の入力であるスニペット集合．対象単語を “coach” とした場合の例を示す．2 列目に示した各用例に対して，5.1 節で説明する前処理を行ったものがスニペットとなる．

それに対し，解釈性の高い確率的生成モデルを用いて意味変化を捉える試みもある．Emms ら [8] は，単語の新たな意味の出現を捉えるための動的な生成モデルを提案しており，Frermann ら [9] は，意味の出現だけではなく変化のパターンも捉えることができるモデル，dynamic Bayesian model of Sense ChANge (SCAN) を提案した．しかし，実際に単語の意味変化をモデル化したい場合は解析対象の単語の**語義数**が自明であることは多くないにも関わらず，これらのモデルは語義数を事前に設定しなければならないという大きな問題がある．

そこで，我々はディリクレ過程をガウス確率場の上に考え Frermann らのモデルを拡張することで，語義数をコーパスから自動で推定し，同時に単語の意味変化を捉えることのできる階層ベイズモデルを提案する [1]．実験では，擬似データを用いた検証を行った結果，提案モデルが解析対象の単語の意味変化と語義数を正しく推定できることを定量的および定性的に示した．また，実データを用いた実験では，推定された単語の意味変化と語義数の評価を行い，英語と日本語の単語に対し分析を行った．以下に本研究の貢献を示す．

- 単語の意味変化と語義数を同時に推定できる階層ベイズモデルを提案した．
- 擬似データを用いて，提案モデルが解析対象の単語の意味変化と語義数を正しく推定できることを定量的，定性的に示した．
- 英語と日本語のテキストデータを用いた実験において，提案モデルによって推定された意味変化と語義数の評価を行い，英語と日本語の単語に対して分析を行った．

2. 関 連 研 究

2.1 Dynamic Bayesian model of sense change (SCAN) [9]

Frermann らは，単語の通時的な意味の発展を捉える，動的なベイズモデル（SCAN）を提案した．SCAN では，対象単語 w 一つに対して一つのモデルが構築され，入力は，対象単語 w が含まれる文章の文脈単語集合 $\boldsymbol{c}$ で構成されるスニペット（短い文書）と，その文章が出現した年のラベルとなっている．スニペットの例を図 1 に示す．

SCAN において，時点 $t \in \{1 \dots T\}$ のスニペット集合は時点ごとのユニグラム混合:

- 意味上の K 次元多項分布 ϕ_t（意味分布）
- 各語義 k の語彙上の V 次元多項分布 $\psi_{t,k}$（語義–単語分布）

でモデル化される．また，それぞれの事前分布にはガウス分布が仮定され，次のように変換することで ϕ を得る:

- 多次元ガウス分布から K 次元ベクトル $\boldsymbol{\alpha}$ を生成
- softmax 変換 $\phi_k = \exp(\alpha_k)/\sum_{k=1}^{K}\exp(\alpha_k)$ によって $K-1$ 次元単体に射影.

ψ についても同様である．ここでは K は既知としており，これが実際には大きな問題となる．そして，意味分布と語義–単語分布のパラメータ ϕ, ψ が, 時間変化と共に変化をするように, 事前分布に 1 階の内的ガウス確率場 (intrinsic Gaussian Markov random field; iGMRF) [10] を定義する．iGMRF は，“近傍と似た値をとる” 事前分布であり [2]，実数ベクトル $\mathbf{x} = (x_1, x_2, \dots, x_T)$ について，ガウス分布 $\mathcal{N}(\mu, \sigma^2)$ を用いて，次のように定義される:

$$x_t \mid \boldsymbol{x}_{-t}, \kappa \sim \mathcal{N}\left(\frac{1}{2}(x_{t-1} + x_{t+1}), \frac{1}{2\kappa}\right). \tag{1}$$

ただし，$\boldsymbol{x}_{-t}$ は $\boldsymbol{x}$ から x_t を除いたものであり，κ は精度パラメータである．また，意味分布と語義–単語分布の事前分布であるガウス分布は，それぞれ変化の度合いをコントロールするパラメータとして κ_ϕ，κ_ψ を持つ．特に，対象単語 w によって意味の変化の “速度” は異なるため，意味分布の事前分布の精度パラメータ κ_ϕ はデータから推定する．

これらを踏まえると，SCAN の生成モデルは次のようになる．ただし，$\mathrm{Ga}(a, b)$ はガンマ分布， $\mathrm{Mult}(\theta)$ は多項分布を表す．

（1） Draw $\kappa_\phi \sim \mathrm{Ga}(a, b)$

（2） For time interval $t = 1 \dots T$

　（a） Draw sense distribution

　　i. $\alpha_t \mid \boldsymbol{\alpha}_{-t}, \kappa_\phi \sim \mathcal{N}\left(\frac{1}{2}(\alpha_{t-1} + \alpha_{t+1}\right), \kappa_\phi^{-1})$

　　ii. $\phi_t = \mathrm{Softmax}(\alpha_t)$

　（b） For sense $k = 1 \dots K$

　　i. Draw sense-word distribution

　　A. $\beta_{t,k} \mid \boldsymbol{\beta}_{-t}, \kappa_\psi \sim \mathcal{N}\left(\frac{1}{2}(\beta_{t-1,k} + \beta_{t+1,k}), \kappa_\psi^{-1}\right)$

　　B. $\psi_{t,k} = \mathrm{Softmax}(\beta_{t,k})$

　（c） For snippet $d = 1 \dots D$

(注1)：実装は https://github.com/seiichiinoue/iscan で公開している．

(注2)：カーネルが隣接する時間に限定されるガウス過程の特別な例と捉えることもできる．

i. Draw sense $z_d \sim \mathrm{Mult}(\phi_t)$
ii. For context position $i = 1 \ldots I$
   A. Draw word $w_{d,i} \sim \mathrm{Mult}(\psi_{t,z_d})$

2.2 Logistic stick-breaking process [11]

ディリクレ過程は有限次元の多項分布を生成するディリクレ分布の無限次元版であり，その実現例の一つとして，stick-breaking process (SBP) [12] が挙げられる．SBP では，確率の総和である長さ 1 の棒を折っていくことで，k 番目のクラスに対する確率 π_k を決定し，デルタ関数 $\delta(\theta_k)$ を基底測度 G_0 からサンプルした場所 $\theta_k \sim G_0$ に立てていくことで，ディリクレ過程 $\mathrm{DP}(\alpha, G_0)$ に従う確率分布 G を生成することができ，生成過程は次のようになる:

$$\pi_k = v_k \prod_{k'}^{k-1}(1 - v_{k'}),\ v_k \sim \mathrm{Be}(1, a),\ \theta_k \sim G_0, \tag{2}$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k). \tag{3}$$

ただし，$\mathrm{Be}(1, a)$ はベータ分布である．

Ren ら [11] は，各クラスが何らかの共変量と紐づいている[3]場合，それをロジスティック変換し各クラスの確率とすることで，同様にディリクレ過程を実現する logistic stick-breaking process (LSBP) を提案した．各クラスのガウス分布に従う確率変数を β_k としたとき，LSBP によって生成される確率分布 G_β は次のようになる:

$$\pi(\beta_k) = \sigma(\beta_k) \prod_{k'=1}^{k-1} \left(1 - \sigma(\beta_{k'})\right), \tag{4}$$

$$G_\beta = \sum_{k=1}^{\infty} \pi(\beta_k)\delta(\theta_k). \tag{5}$$

ただし，$\sigma(x)$ はシグモイド関数を表す．これは，各クラス k の確率を決める確率変数を，ベータ分布から生成するのではなく，元々紐づいている共変量 β_k をロジスティック変換し，stick-breaking 表現によって確率 π_k を生成しており，あとは SBP と同様に，それを $\theta_k \sim G_0$ に立てていくことで確率分布 G_β を生成する．

3. 提案手法

3.1 Infinite SCAN

我々は，2.1 節で紹介した SCAN において，対象単語によって異なる語義数をコーパスから自動的に推定できるように，2.2 節で紹介した LSBP を用いて拡張を行った階層ベイズモデルを提案する．

提案モデルでは，SCAN と同様に対象単語 w 一つに対して一つのモデルを考え，入力も SCAN と同様に，対象単語 w が含まれる文章の文脈単語集合 $\boldsymbol{c}$ で構成されるスニペットとその文章が出現した年のラベルとする．

(注3)：SCAN の場合はガウス分布に従う確率変数が各語義，語義における各単語と紐づいている．

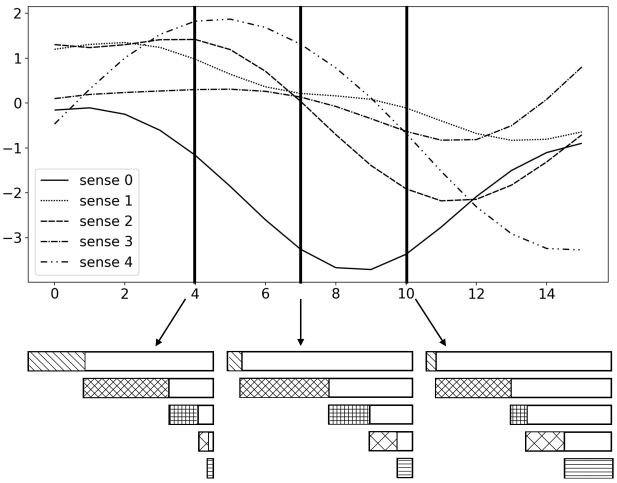

図 2　意味分布におけるガウス分布に従う確率変数の LSBP 変換.

時点 t に出現したスニペット集合は，ϕ_t と $\psi_{t,k}$ それぞれの上に定義される意味分布と語義–単語分布によって表されるが，それぞれはガウス分布に従っている．ここで，意味分布のパラメータ ϕ_t について，対象単語 w によって異なる語義数 S_w をコーパスから自動的に推定できるように，LSBP を用いて生成過程を次のように変更する:

$$\alpha_t \mid \boldsymbol{\alpha}_{-t}, \kappa_\phi \sim \mathcal{N}\left(\frac{1}{2}(\alpha_{t-1} + \alpha_{t+1}), \kappa_\phi^{-1}\right), \tag{6}$$

$$\phi_{t,k} = \sigma(\alpha_{t,k}) \prod_{k'=1}^{k-1} \left(1 - \sigma(\alpha_{t,k'})\right)\ (k = 1, \ldots, K). \tag{7}$$

ただし，K は考慮される語義の最大数である．ここで，LSBP は無限次元の多項分布を生成できるが，実際には十分な語義の次元があればそれ以上使われることはないため，本研究では SCAN と同様に $K = 8$ とした[4]．図 2 に意味分布の LSBP による変換のイメージを示す．横軸は時間を表し，縦軸はガウス分布に従う確率変数のスケールを表しており，時点 t で切り取った時の確率変数の集合 $\{\alpha_{t,1}, \ldots, \alpha_{t,K}\}$ に対し，LSBP 変換を行うことで各語義の確率を得ている．また，提案モデルでは，意味分布のパラメータ ϕ_t は，全体を見て正規化する softmax 変換ではなく，それぞれの意味における停止確率を語義ごとに確率化する LSBP によって構築される．そのため提案モデルでは，SCAN のように κ_ϕ を全ての語義 $k \in \{1 \ldots K\}$ で共有するのではなく，各語義の値 α ごとに異なる分散 $\kappa_\phi^{(k)}$ を仮定し推定する．

このように，提案モデルでは，意味分布において時間軸上に広がるガウス確率場の上に，ディリクレ過程を実現する LSBP を導入することで，意味分布を実質的に無限次元化し，対象単語 w に適した語義数 S_w をコーパスから自動で推定することができる．

3.2 MCMC 法による推定

提案モデルの推定には，ブロック化 Gibbs sampler を用いる．提案モデルにおける推定パラメータは，スニペットに割り当て

(注4)：事前実験より，8 個以上の語義を持つ単語はほとんど存在しなかった．

Algorithm 1: MCMC Procedure

Initialize $\kappa_\phi^{(k)} = 4.0$ (for all k)
Initialize $\kappa_\psi = 100.0$
Initialize $a = 7.0$, $b = 3.0$
for $t = 1 \dots T$ **do**
  Initialize $\alpha_t \sim \mathcal{N}\left(\frac{1}{2}(\alpha_{t-1} + \alpha_{t+1}), \kappa_\phi^{-1}\right)$
  Set $\phi_t = \mathrm{LSB}(\alpha_t)$
  for $k = 1 \dots K$ **do**
    Initialize $\beta_{t,k} \sim \mathcal{N}\left(\frac{1}{2}(\beta_{t-1,k} + \beta_{t+1,k}), \kappa_\psi^{-1}\right)$
    Set $\psi_{t,k} = \mathrm{Softmax}(\beta_{t,k})$
  end
end
for $j = 1 \dots J$ **do**
  Sample z according to Eq. (8)
  for $t = 1 \dots T$ **do**
    Sample ϕ according to posterior in Eq. (9)
    Sample ψ according to posterior in Eq. (10)
  end
  Sample κ_ϕ according to posterior in Eq. (11)
end

図 3　MCMC 法による Infinite SCAN の推定.

られる語義 z，意味分布と語義–単語分布の事前分布である α（ϕ の LSBP 変換前），β（ψ の softmax 変換前），ガンマ分布に従う意味分布の事前分布の精度パラメータ κ_ϕ である．サンプリングでは，順番に推定パラメータ以外のパラメータを所与として，事後分布からサンプリングする．サンプリングの擬似コードを図 3 に示した．各パラメータの初期値は基本的に Frermann らに従っている．ただし，本研究では，対象単語の意味の変化を “語義–単語分布の変化” ではなく，できるだけ “意味分布の変化” で捉えたいため [5]，Perrone ら [13] に従い，語義–単語分布の事前分布の精度パラメータ κ_ψ の値を比較的大きな 100.0 とした．

a） スニペットの語義

d 番目のスニペットの語義 z_d は，サンプリング時点でのパラメータ ϕ，ψ を用いて，次の事後分布に従ってサンプリングされる:

$$p(z_d \mid \boldsymbol{w}, t, \phi, \psi) \propto p(z_d \mid t) p(\boldsymbol{w} \mid t, z_d) = \phi_{z_d}^{(t)} \prod_{w \in \boldsymbol{w}} \psi_w^{(t,z_d)} \tag{8}$$

b） 意味分布のパラメータ

意味分布の事前分布はガウス分布となっているため，多項分布との間に共役性がないので，ディリクレ–多項分布のようなサンプリングはできない．Linderman ら [14] は，Pólya-gamma 補助変数を用いることで，LSBP を用いて表現される，ガウス分布を事前分布として持つ多項分布のパラメータに対するギブスサンプリングを提案しており，本研究ではそれに従って推定を行う[6]．α の事後分布は次のようになる:

$$p(\alpha_t \mid z, \boldsymbol{\alpha}_{-t}, \omega) \propto \mathcal{N}\left(\omega^{-1} f(c) \mid \alpha_t\right) \mathcal{N}\left(\alpha_t \mid \boldsymbol{\alpha}_{-t}, \kappa_\phi^{-1}\right) \propto \mathcal{N}\left(\alpha_t \mid \tilde{\mu}, \tilde{\kappa}_\phi^{-1}\right). \tag{9}$$

ただし，ω は補助変数であり，c_k を k 番目の語義に属しているスニペットの個数，$N(c_k) = \sum_k c_k - \sum_{j<k} c_j$ として，Pólya-gamma 分布 $\omega \mid z, \alpha_t \sim \mathrm{PG}(N(c_k), \alpha_t)$ からサンプリングされる．また，$f(c_k) = c_k - N(c_k)/2$ として，事後分布の平均と分散は，$\tilde{\mu} = (f(c_k) + \mu_k \kappa_\phi) \cdot \tilde{\kappa}_\phi$，$\tilde{\kappa}_\phi = (\omega_k + \kappa_\phi)^{-1}$ である．

c） 語義–単語分布のパラメータ

語義–単語分布も意味分布と同様に事前分布としてガウス分布を仮定しているため，ディリクレ–多項分布のようなサンプリングはできない．Mimno ら [15] は，softmax 変換を用いて表現される，ガウス分布を事前分布としてもつ多項分布のパラメータに対するギブスサンプリングを提案しており，本研究ではそれに従って推定を行う[7]．スニペット数を D，スニペット長を N_d とすると，β の事後分布は次のようになる:

$$p(\beta_t \mid z, \beta_{-t}, \kappa_\psi^{-1}) \propto \prod_{d=1}^{D} \left(\prod_{n=1}^{N_d} \frac{\exp(\beta_{w_n}^{(t,z_d)})}{1 + \sum_{w' \neq w} \exp(\beta_{w'}^{(t,z_d)})} \right) \mathcal{N}(\beta_t \mid \boldsymbol{\beta}_{-t}, \kappa_\psi^{-1}) \tag{10}$$

d） 精度パラメータ

平均が既知であるガウス分布の精度パラメータは，ガンマ事後分布に従う．ガンマ分布の形状パラメータを a，尺度パラメータを b とすると，κ_ϕ の事後分布は次のようになる:

$$p(\kappa_\phi^{(k)} \mid \alpha_k, a, b) = \mathrm{Ga}\left(a + \frac{T}{2}, b + \frac{1}{2} \sum_{t=1}^{T} (\alpha_{t,k} - \bar{\alpha}_k)\right) \tag{11}$$

ただし，$\bar{\alpha}_k = \frac{1}{T} \sum_t \alpha_{t,k}$ は語義 k における α の平均である．

4. 擬似データを用いた検証

4.1 擬似データの生成

実データでの検証に先立ち，提案モデルが，任意の意味変化と語義数をきちんと推定できるか検証するため，擬似データによる検証を行う．

擬似データの生成において，意味分布の生成にはガウス過程を用い，任意の語義数分の意味変化曲線をガウス過程からサンプリングし，その曲線を時点ごとに softmax 関数で変換するこ

表 1　各モデルによって推定された意味分布と正解の意味分布の Kullback-Leibler 距離.

語義数 S_w	2	3	4	5
SCAN ($K = 8$)	0.2612	0.2788	0.1031	0.0085
Infinite SCAN	**0.0009**	**0.0016**	**0.0044**	**0.0043**

表 2　実験に使用したコーパスの年代とサイズ.

コーパス	年代	単語数
COHA （英語）	1810–2009	142,587,656
CHJ（日本語）	1874–1997	36,701,284

(注5)：前者の場合，語義–単語分布のみで意味変化が説明されてしまい，語義を正しく捉えられないため，語義数の推定が難しくなる.
(注6)：詳細は参考文献 [14] を参照されたい.
(注7)：詳細は参考文献 [15] を参照されたい.

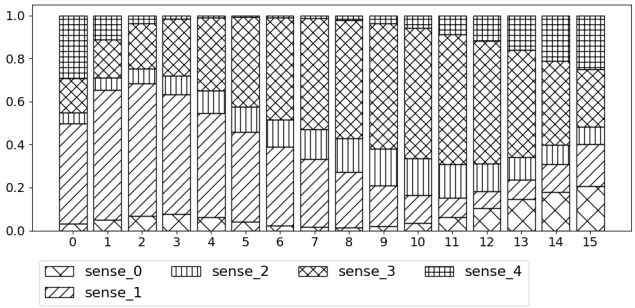

図 4　語義数が $S_w = 5$ の場合の意味変化の正解例.

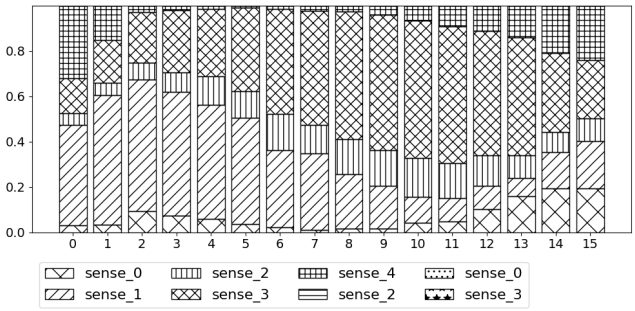

図 5　語義数が $S_w = 5$ の擬似データに対する提案モデルの推定結果. 正解例に合わせて語義の並べ替えを行った.

とで意味分布を得た．また，語義–単語分布の生成にはディリクレ分布を用い，これらの意味分布と語義–単語分布を用いて正解データをランダムに生成した．本実験では，擬似データの時点数は $t = 16$ [8]，語彙サイズを $V = 3,000$ [9]で固定し，語義数を $S_w = 2, \ldots, 5$ と動かしてスニペットをランダムに生成した．

4.2 結　　果

表 1 に，真の語義数を $S_w = 2, \ldots, 5$ と動かした場合における SCAN と提案モデルの擬似データに対する推定結果の比較を示した．ここでは，意味の分布を正しく表現できるか定量的に測るため，指標として推定された意味分布と正解の意味分布の $\mathbb{R}^{T \times K}$ 上での Kullback-Leibler 距離を用いた．結果から，語義数を自動的に推定できない SCAN と比較して提案モデルが優れていることがわかる．

また，図 4, 5 に，語義数が $S_w = 5$ の場合の正解例と，擬似データに対する提案モデルの推定結果を示した．横軸が時点，縦軸が時点における語義の確率を表し，色に対応する語義を凡例に示した．ただし，凡例では，各語義 k に対応する推定された語義–単語分布において最も支配的であった語義を示した．提案モデルの推定結果（図 5）においていくつかの語義が重複して現れているが，それらの語義の確率は微小であり，無視できることに注意されたい．図より，提案モデルは意味変化をほとんど正確に捉えており，また，真の語義数 $S_w = 5$ も正しく推定していることがわかる．

(注8)：実験で使用したコーパスのうち，最も多い時点数が 16 であったため.
(注9)：実験で使用したコーパスの平均語彙サイズが 3,000 程度であったため.

表 3　対象単語の一覧とスニペットの統計量.

単語	年代	スニペット数	語彙サイズ
image	1810–2009	19,499	19,104
nature	1810–2009	83,188	33,375
pass	1810–2009	36,605	24,555
record	1815–2009	33,992	23,886
coach	1811–2009	9,758	11,962
power	1810–2009	142,527	42,932
団塊	1895–1997	92	634
取り組む	1887–1997	647	2,083

表 4　SCAN と提案モデルの sense coherence と sense diversity．計算には表 3 上段に示した 4 つの英語コーパスの対象単語を用いた.

	Coherence	Diversity
SCAN ($K = 8$)	0.124	0.275
Infinite SCAN	**0.146**	**0.398**

5. 実　　験

5.1 データセット

本研究では，英語と日本語のテキストを用いて実験を行い，英語の通時コーパスである Corpus of Historical American English (COHA) [3] と，日本語の通時コーパスである，国立国語研究所の「日本語歴史コーパス (CHJ)」[4] の一部として公開されている「昭和・平成書き言葉コーパス」を用いた．コーパスの統計量を表 2 に示した．また，それぞれのコーパスに対して以下の前処理を行った．

a) 英　　語

トークナイズ，見出し語化を行ったのち，ストップワードの削除を行った．また品詞タグ付けを行い，名詞，動詞，形容詞のみを抽出し使用した．以上の全ての処理は Natural language toolkit (NLTK) [16] を用いた．

b) 日　本　語

トークナイズ，見出し語化と固有名詞等の正規化，高頻度語の削除を行った．日本語テキストは英語とは異なり，ストップワードのリストが明確に存在しないため，本研究では，Mikolov ら [17] や，Levy ら [18] による単語分散表現の推定におけるヒューリスティクスに従って，高頻度語を以下の確率:

$$p = 1 - \sqrt{\frac{g}{f(w)}} \tag{12}$$

に従って削除することで対処した．ただし，$f(w) = \#(v)/\sum \#(v)$ であり，g はコーパスのサイズに依存する閾値で，本研究では $g = 10^{-4}$ とした．

以上の前処理を行ったのち，文脈窓幅を 10 として対象単語が出現する箇所の前後の文脈単語集合をコーパスから抽出し，対象単語のコーパスを作成した．以下の実験で用いる対象単語の一覧とスニペットの統計量を表 3 に示す．上段には 5.2 節の実験設定で説明する，語義数の評価に用いた単語を，下段には意味変化の評価に用いた単語を示した．

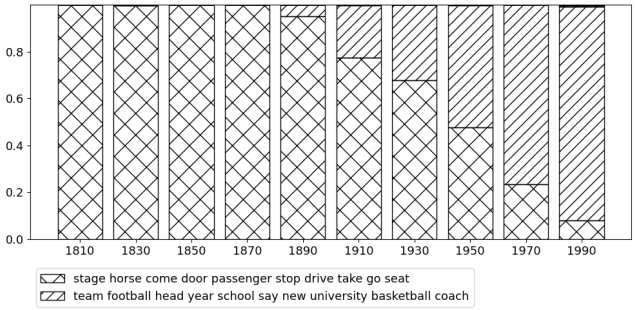

図 6　対象単語 “coach” に対する提案モデルの推定結果.

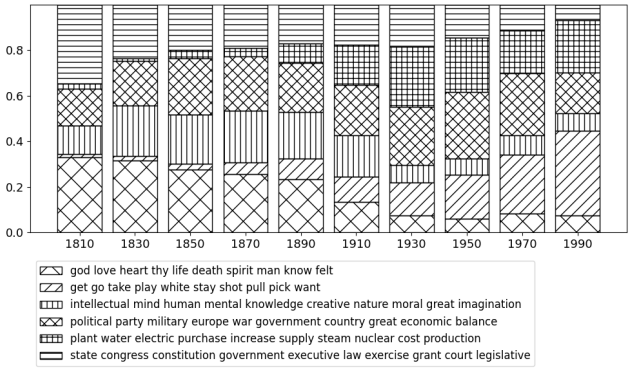

図 7　対象単語 “power” に対する提案モデルの推定結果.

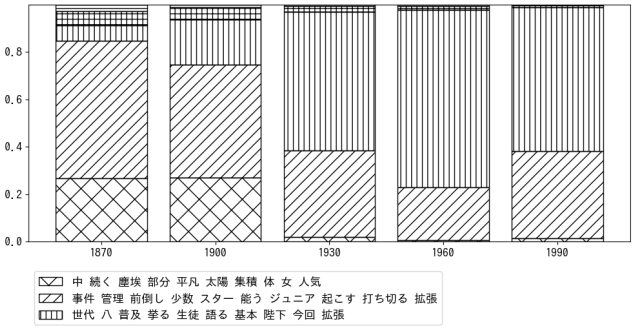

図 8　対象単語 “団塊” に対する提案モデルの推定結果.

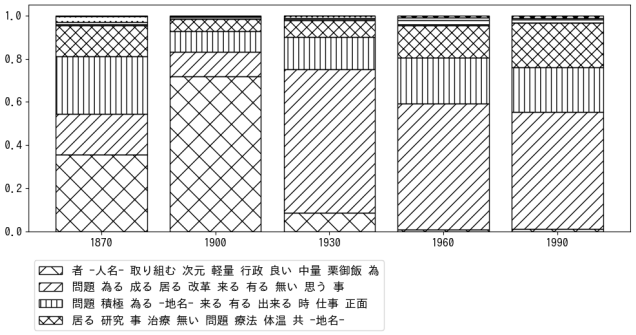

図 9　対象単語 “取り組む” に対する提案モデルの推定結果.

5.2 実験設定

提案モデルの特徴は，解析の対象単語の**語義数**と**意味変化**を自動で推定できることであるため，実験では，両者の評価を定量的，定性的に行う．**語義数**の自動推定に関する実験では，英語コーパスを用い，提案モデルの推定の性能を，トピックモデルの評価に用いられる指標である coherence [19] と diversity [20] を用い，推定された意味の一貫性と多様性（非冗長性）を定量的に評価する．**意味変化**の推定に関する実験では，英語コーパスに加えて，日本語コーパスを用いて評価を行う．それぞれの言語における意味変化が既知である複数の対象単語に対し，提案モデルの推定を行い，意味分布の可視化を行うことで定性的に評価を行う．

実験では，提案モデルの推定において，Gibbs サンプリングのイテレーション回数を 2000 回，時点 t に含まれる期間を決める時間間隔 Δt を英語コーパスを用いた実験では 20，日本語コーパスを用いた実験では 30 [10] とした．また，推定に使用する語彙は頻度の上位 3000 語とした．

5.3 語義数の評価

推定された語義数の評価では，意味の coherence と diversity を用いて，ベースラインモデルである SCAN と提案モデル Infinite SCAN の比較を行う．Sense coherence は Lau ら [19] に従い，次のように定義した:

$$\mathrm{SC} = \sum_{k=1}^{K} \frac{\eta}{45} \sum_{i=1}^{10} \sum_{j=i+1}^{10} f(w_i^k, w_j^k). \qquad (13)$$

ここで，η は語義における正規化定数であり，Lau らは単純に $\eta = \frac{1}{K}$ としているが，本研究では，提案モデルにおける確率が微小な語義の語義–単語分布を正当に評価するため，各語義 k の確率を用いて $\eta = p(k)$ とした．また，$f(w, w')$ は単語の意味空間上の類似度であり，本研究では word2vec [17] を用いて計算されるコサイン類似度とした．Sense diversity は，Dieng ら [20] を参考に，全ての語義における上位 10 単語の集合のうち，重複のない単語の割合とした．

表 4 に，対象単語に対する SCAN と提案モデルの推定結果を用いて計算したスコアを示した．ただし，SCAN は語義数が既知でない状態で推定を行う提案モデルの設定に合わせるため，意味分布の次元数 K は，Frermann らに従いデフォルト値の 8 としていることに注意されたい．どちらの指標においても提案モデルは SCAN を上回る結果となった．Coherence の改善については，推定された語義–単語分布の代表的な単語（図 6 ~9 の凡例に示された単語）の意味的な一貫性が高いということを示している．Diversity の改善については，推定された語義–単語分布の代表的な単語の意味的な重複が小さくなり，適切な粒度で語義数を推定できていることを示している．

5.4 意味変化の評価

意味変化の定性的な評価では，表 3 の下段に示した，英語と日本語の単語を用いる．図 6 ~9 に提案モデルによる対象単語に対する推定結果を示した．凡例には，語義 k における語義–単語分布 $\psi_k = \sum_t \psi_{t,k}$ の Normalized pointwise mutual information (NPMI) [21] の高い単語を上から 10 個示した．以下の分析では，簡単のため，それぞれの凡例において，行に対応する語義を上から順番に「1 番目の意味」，「2 番目の意味」，…，「K 番目の

(注10)：日本語コーパスは，英語コーパスと比べてコーパスの時間間隔が広く，スパース性を考慮して若干大きめのサイズとした．

意味」と記述する．

まず，英語を対象とした実験では，Frermann らを参考に選択した単語である，“coach” と “power” を用いて評価を行う．“coach” では，コーパスから推定された語義数は 2 つとなった．馬車や乗り物としての意味（1 番目の意味）から指導や指導者という意味（2 番目の意味）に変化していることが捉えられている．“power” については，コーパスから推定された語義数は 6 つとなった．宗教や教会の文脈で使用される力（1 番目の意味）は，相対的な支配力は下がりつつも，現在に至っても使用されていることがわかる．経済的な力や，国際的な政治力や戦争などの文脈における力（4 番目の意味）についても，過去から現在にわたって広く使われており，世界大戦前後の時期はコーパスが社会的背景に影響を受けることから支配的になっていることが確認できる．また，肉体上・精神上の自然の能力，体力，知力といった意味（3 番目の意味）と，法や政治的な力といった意味（6 番目の意味）は現在に近づくにつれて使われることが少なくなっている．一方で，強さとしての意味（2 番目の意味）や動力の意味（5 番目の意味）は現在になるにつれて使用されることが増えてきていることがわかる．

次に，日本語の “団塊” と “取り組む” についての評価を行う．昭和・平成書き言葉コーパスは，表 3 に示した通り英語とは異なりコーパスサイズが小さく，データスパースネスの問題がある．特に，時点によってスニペット数が大きく異なる場合に推定に大きな影響があり，以下で分析する例でも一部ノイズが入った結果となった．“団塊” については，捉えられた語義数は約 3 つとなった．過去は，地学関係での団塊（1 番目の意味）といった意味で使用されていたが，現在ではほとんどそういった意味では使われなくなり，ビジネス的な文脈での「団塊の世代」の用法（2 番目の意味），「団塊の世代」にまつわる意味（3 番目の意味）として使われることが多くなっていることがわかる．ここで，推定結果では，2 番目の意味が過去にも支配的であると捉えているが，これは対象単語 “団塊” のコーパスにおいて，1910 年から 1980 年の間のデータが非常に少ないため，GMRF によって隣接する意味分布 ϕ_k に近い値を推定値としてしまっているためであることに注意されたい．“取り組む” については，捉えられた語義数は約 4 つとなっており，相撲の取り組み（1 番目の意味），政治的な文脈における取り組み（2 番目の意味），仕事で起きた問題に対する取り組み（3 番目の意味），医療や研究における取り組み（4 番目の意味）という意味が捉えられた．過去は 1 番目の意味が支配的であったように，相撲の取り組みとして使用されることが多かったが，課題やプロジェクトに取り組むといった使用が増えてきていることがわかる．

これらの結果は歴史的背景や辞書からも妥当であり，提案モデルはコーパスからある程度正しく意味変化を捉えることができていることがわかる．

6. まとめと今後の展望

本研究では，単語の通時的な意味変化のモデル化において，単語によって異なる語義数を自動的に推定できる階層ベイズモデル Infinite SCAN を提案し，ベースラインモデルである SCAN より優れた性能を持つことを定量的，定性的に示した．擬似データを用いた実験では，意味変化と語義数を正しく推定できることを示した．また，実データを用いた実験においては，英語と日本語を対象として，複数の単語の推定結果を示し，分析した．今後は，データスパースネスの問題への対処を含めたモデル推定の安定化や，本稿では行えなかった，語義数の自動推定に対する外的な定量評価を，辞書や知識ベースを用いて行っていきたい．

謝　　辞

本研究は国立国語研究所の共同研究プロジェクト「現代語の意味の変化に対する計算的・統計力学的アプローチ」，「通時コーパスの設計と日本語史研究の新展開」および JSPS 科研費 19H00531，18K11456 の研究成果の一部を報告したものである．

文　　献

[1] Jean Aitchison. 2001. Language Change: Progress Or Decay?. Cambridge Approaches to Linguistics. Cambridge University Press.

[2] Angus Stevenson, editor. 2010. The Oxford English Dictionary. Oxford University Press, third edition.

[3] Mark Davies. 2010. The Corpus of Historical American English (COHA). Available online at https://www.english-corpora.org/coha/.

[4] 国立国語研究所. 2021. 日本語歴史コーパス (バージョン 2021.3, 中納言バージョン 2.5.2). https://ccd.ninjal.ac.jp/chj/ (2022 年 2 月 5 日確認)

[5] Robert Bamler, and Stephan Mandt. 2017. Dynamic Word Embeddings. In *Proceedings of the 34th International Conference on Machine Learning*, edited by Doina Precup and Yee Whye Teh, 70:380–89.

[6] Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW'15, p. 625–635.

[7] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 1489–1501.

[8] Martin Emms, and Arun Kumar Jayapal. 2016. Dynamic Generative Model for Diachronic Sense Emergence Detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1362–73.

[9] Lea Frermann, and Mirella Lapata. 2016. A Bayesian Model of Diachronic Meaning Change. *Transactions of the Association for Computational Linguistics* 4: 31–45.

[10] Havard Rue and Leonhard Held. 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

[11] Lu Ren, Lan Du, Lawrence Carin, and David Dunson. 2011. Logistic Stick-Breaking Process. *Journal of Machine Learning Research*: JMLR 12: 203–39.

[12] Jayaram Sethuraman. 1994. A Constructive Definition of Dirichlet Priors. Statistica Sinica 4 (2): 639–50.

[13] Valerio Perrone, Marco Palma, Simon Hengchen, Alessandro Vatri, Jim Q. Smith, and Barbara McGillivray. 2019. GASC: Genre-Aware Semantic Change for Ancient Greek. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 56–66.

[14] Scott Linderman, Matthew J. Johnson, and Ryan P. Adams. 2015. Dependent Multinomial Models Made Easy: Stick-Breaking with the Polya-Gamma Augmentation. In *Advances in Neural Information Processing Systems*, vol. 28, 3456–64.

[15] David Mimno, Hanna Wallach, and Andrew McCallum. 2008. Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Pri-

ors. In *Neural Information Processing Systems Workshop on Analyzing Graphs*. vol. 61.

[16] Steven Bird and Ewan Klein and Edward Loper. 2009. Natural language processing with Python: analyzing text with the natural language toolkit, OReilly Media, Inc.

[17] Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations*.

[18] Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the association for computational linguistics* 3: 211-225.

[19] Jey Han Lau and David Newman and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539.

[20] Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

[21] Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. 2009. *Proceedings of the German Society for Computational Linguistics and Language Technology*, 30:31–40.