スケール不変な木構造棒折り過程に基づく無限階層トピックモデル

江島舟星 [1]　　持橋大地 [2]
[1] ハーバード大学大学院 政治学科　[2] 統計数理研究所
shuseieshima@g.harvard.edu　　daichi@ism.ac.jp

概要

階層的トピックモデルは、多様なトピックを木構造で整理するために活用されてきた。しかし既存手法では、木が深くなるにつれて各トピックの平均確率が指数的に小さくなり、似通ったトピックが多数推定されてしまう。本研究ではこの課題に対し、スケール不変木構造棒折り過程を用いた階層的トピックモデル (ihLDA) を提案する。提案手法はトピック生成時にその親を考慮することで、トピック平均確率の減衰を抑え、深い木構造の推定を可能にする。また、提案手法は木構造棒折り過程の階層化でもある。提案手法が、ニューラルモデルを含む既存手法より 2 つの指標で優れることを実験的に確認した。

1　はじめに

トピックモデルは文章の要約、注釈、分類のために幅広く用いられているが [1, 2, 3, 4]、大規模コーパスから数千ものトピックを推定できるモデルが登場したことで [5, 6, 7, 8]、大量のトピックを解釈する必要性が高まった。

階層的トピックモデルは、トピックの階層的構造もコーパスから学習することで、トピックの整理と解釈を助けるものである [9, 10, 11, 12, 13, 14, 15]。しかし既存手法では、用いられる確率が小さく、似通ったトピック–単語分布を持つ多数のトピックが推定される。これは確率的モデルだけでなく、近年のニューラルモデルにおいても共通の現象である。

トピックの平均確率が減衰する原因は、階層モデルで用いられる棒折り過程 [16] が、各トピックの平均確率を木が深くなるにつれ指数的に小さくすることにある。既存手法では推定される木構造を一定の深さで打ち切るなどのヒューリスティックに基づき、トピックの減衰を防いできた。

これに対し本研究は、スケール不変な木構造棒折り過程と、それに基づく無限階層トピックモデル (ihLDA) を提案し、以下の 3 点を実現する。

1. トピック生成時にその親を考慮し、トピック平均確率の減衰を抑えた深い木構造の推定
2. 木構造棒折り過程 (TSSB)[11] の階層ベイズ化
3. 木構造に属する全てのトピックを数え上げる必要のない、効率的なサンプリング

棒折り過程を用いる既存手法に提案手法を取り入れることは容易であり、本研究は確率的モデル、ニューラルモデル双方の改善に資するものである。

2　木構造棒折り過程

木構造棒折り過程 (TSSB)[11] は、2 つの棒折り過程を [16] を組み合わせ、理論的には無限の深さと幅を持つトピック木を図 1 のように実現する。

階層的トピックモデルは潜在トピック $\boldsymbol{\epsilon} = \epsilon_1 \epsilon_2 \cdots$ を各文書の各単語に割り当てる。深さ $|\boldsymbol{\epsilon}|$ のトピック $\boldsymbol{\epsilon}$ は先祖と子を持ち、$\{\boldsymbol{\kappa} : \boldsymbol{\kappa} \prec \boldsymbol{\epsilon}\}$ が先祖集合を表す。また、$\boldsymbol{\epsilon}$ の親は $\boldsymbol{\epsilon}'$ で、$\boldsymbol{\epsilon}$ の子は $\{\boldsymbol{\epsilon} k : k \in 1, 2, 3, \ldots\}$ で示す。

TSSB におけるトピック $\boldsymbol{\epsilon}$ の確率 $\pi_{\boldsymbol{\epsilon}}$ は、

$$\pi_{\boldsymbol{\epsilon}} = \nu_{\boldsymbol{\epsilon}} \prod_{\boldsymbol{\kappa} \prec \boldsymbol{\epsilon}} (1 - \nu_{\boldsymbol{\kappa}}) \cdot \prod_{\boldsymbol{\kappa} \preceq \boldsymbol{\epsilon}} \phi_{\boldsymbol{\kappa}}\,,\ \phi_{\boldsymbol{\epsilon} k} = \psi_{\boldsymbol{\epsilon} k} \prod_{j=1}^{k-1} (1 - \psi_{\boldsymbol{\epsilon} j}) \quad (1)$$

で表される。式 (1) の第 1 項はトピック $\boldsymbol{\epsilon}$ が縦方向に止まる確率であり、続く積は $\boldsymbol{\epsilon}$ の先祖を縦方向には通過しつつ、横方向では $\boldsymbol{\epsilon}$ とその先祖に止まる確率である。縦横方向に止まる確率は、ベータ分布に従うとする。

$$\nu_{\boldsymbol{\epsilon}} \sim \mathrm{Be}(1, \alpha_0)\,,\ \psi_{\boldsymbol{\epsilon}} \sim \mathrm{Be}(1, \gamma_0)\,. \quad (2)$$

さらにここで λ を用いて、深さ固有の α_0 を $\alpha_{\boldsymbol{\epsilon}} = \alpha_{|\boldsymbol{\epsilon}|} \cdot \lambda^{|\boldsymbol{\epsilon}|-1}, 0 \leq \lambda \leq 1$ と定め、深いほど単語が止まりやすくする。以下、$\alpha_{\boldsymbol{\epsilon}}$ を α と記す。

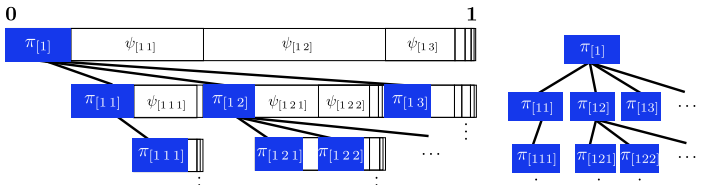

図 1　木構造棒折り過程 [11]。青の区間はその幅に比例した確率 π のトピックを、角括弧は各トピックの経路を、黒線は親子関係を、ψ は横方向の棒折り過程を示す。右図はトピックのみを抜き出したものである。

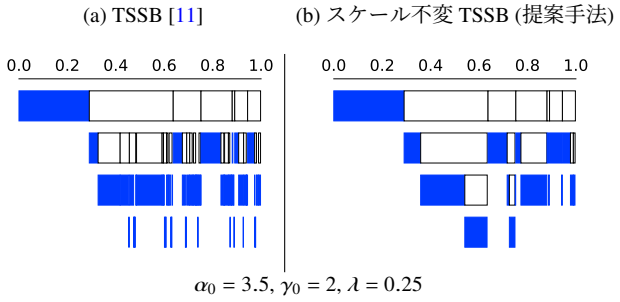

図 2　TSSB [11] (a) とスケール不変 TSSB (b)。青の区間はその幅に比例したトピックを示す。(a)(b) は同じハイパーパラメータから生成された木構造であるが、(b) の提案手法は確率 π_ϵ が小さすぎるトピックを生成しない。

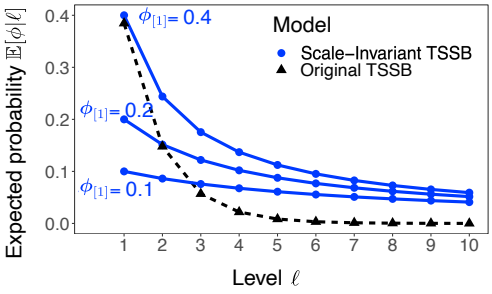

図 3　深さ ℓ における横方向の平均分割確率。3 つの異なる確率を持つ木の根を提案手法では示した。提案手法の減衰は、TSSB よりも遅い。γ_0 の値は 0.8 に固定した。

3　スケール不変木構造棒折り過程

TSSB は階層的トピックモデルに用いられるが [14, 15]、木構造が深くなるほどトピックの平均確率が指数的に小さくなるという問題を抱えている。図 2(a) に示されるように、3 段目、4 段目のトピックの確率はより上部のトピックと比べて小さい。

このような TSSB の性質は、横方向パラメータ ψ_ϵ が深さに関わらず同じ平均確率を持つために生じる。付録 A が示す通り、深さ ℓ における横方向の平均分割確率は $\mathbb{E}[\phi|\ell] \approx 1/(2\gamma+1)^\ell$ であり、深さ ℓ が分母の指数にある。このため、図 3 の点線が示すような指数的な減衰が生じてしまう。

そこで本研究では、式 (2) の γ_0 を親ノード ϵ' から

$$\psi_\epsilon \sim \mathrm{Be}(1, \phi_{\epsilon'}\gamma_0), \tag{3}$$

としてスケールを修正する。ここで根ノードにおいては $\phi_{\epsilon'}=1$ である。以下 $\gamma=\phi_{\epsilon'}\gamma_0$ と表記する。

式 (3) で鍵となるのは、親トピック ϵ' における横方向の分割確率 $\phi_{\epsilon'}$ を用いて、その子トピックの**相対的な**棒の長さ ψ_ϵ を求めている点である。これによって、付録 A に示した通り、深さ ℓ での平均的な棒の長さを $\mathbb{E}[\phi|\ell] \approx 1/(2\gamma + 1/\mathbb{E}[\phi|\ell-1])$ $(\ell \geq 2)$ とし、深さ ℓ に対してスケール不変となる分割を実現した。図 3 の実線が提案手法であるスケール不変木構造棒折り過程 (スケール不変 TSSB) を表している。

図 2(b) にこのスケール不変 TSSB からのサンプルを、(a) と同じハイパーパラメータを用いて示した。提案手法を用いると、木構造が深くなってもトピックの確率が小さくなりにくいことがわかる。

4　スケール不変木構造棒折り過程に基づく無限階層トピックモデル

トピックモデルは、文書–トピック分布とトピック–単語分布の組み合わせであり、本論文が提案する ihLDA は、スケール不変 TSSB を前者に、階層 Pitman-Yor 過程 [17] を後者に用いる。

文書–トピック分布　それぞれのトピックが用いられる確率は文書ごとに異なるが、トピック自体は全ての文書で共有されなければならない。このため本研究は、スケール不変 TSSB を階層木構造棒折り過程 (HTSSB) [18] に適用する。図 4 の通り、HTSSB は親 TSSB から子 TSSB を各文書に対して階層的に生成する。

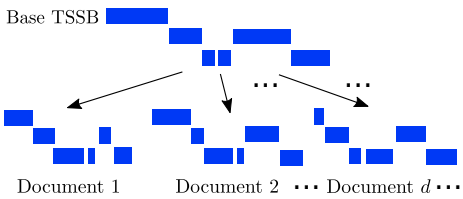

図 4　HTSSB は親 TSSB から子 TSSB を階層的に生成する。子 TSSB は各文書の文書–トピック分布である。

HTSSB は階層ディリクレ過程 (HDP)[19] を式 (2) と (3) の縦横方向のパラメータに適用する。HTSSB での親子関係を示すために、チルダ記号を用いて子 TSSB のトピック ϵ に対応する親 TSSB のト

ピックを $\tilde{\boldsymbol{\epsilon}}$ と記す。HDP に従うと、縦方向の確率は $\nu_{\boldsymbol{\epsilon}} \sim \mathrm{Be}(a\tau_{\tilde{\boldsymbol{\epsilon}}}, a(1-\sum_{\boldsymbol{\kappa}\le\tilde{\boldsymbol{\epsilon}}}\tau_{\boldsymbol{\kappa}}))$、$\tau_{\boldsymbol{\epsilon}} = \nu_{\boldsymbol{\epsilon}}\prod_{\boldsymbol{\kappa}<\boldsymbol{\epsilon}}(1-\nu_{\boldsymbol{\kappa}})$、横方向の確率は $\psi_{\boldsymbol{\epsilon}k} \sim \mathrm{Be}(b\phi_{\tilde{\boldsymbol{\epsilon}}k}, b(1-\sum_{j=1}^{k}\phi_{\tilde{\boldsymbol{\epsilon}}j}))$ となり、各文書の木構造は式 (1) から求まる。子 TSSB でのトピックの割り当てが、HTSSB によって親 TSSB にも影響することに留意しなければならない。

トピック–単語分布 階層 Pitman-Yor 過程 (HPY)[17] によって、深さに応じてトピックの専門性を高めつつ、親子トピックに意味的な関連性を持たせることができる。$H_{\boldsymbol{\epsilon}}$ をトピック $\boldsymbol{\epsilon}$ のトピック–単語分布とすると、この事前分布が Pitman-Yor 過程に従い $H_{\boldsymbol{\epsilon}} \sim \mathrm{PY}(d_{|\boldsymbol{\epsilon}|}, \theta_{|\boldsymbol{\epsilon}|}, H_{\boldsymbol{\epsilon}'})$ となる。これを根ノード $H_{[1]} \sim \mathrm{PY}(d_0, \theta_0, H_0)$ に至るまで繰り返し、$H_0 = 1/V$ と単語数 V を用いて一様分布を考える。トピック–単語分布の木構造は文書-トピック分布の親 TSSB と同じであり、各文書の各トピックで対応するトピック–単語分布が保証されている。

生成過程 ihLDA の生成過程は、次の通りである。

1. 親 TSSB $\widetilde{\boldsymbol{\pi}}$ の生成。$\boldsymbol{\pi}$ は TSSB とする。
2. トピック–単語分布 $H_{\boldsymbol{\epsilon}}$ を HPY を使って $\widetilde{\boldsymbol{\pi}}$ の各トピックに対して生成。
3. 各文書 d に対して文書分布 $\boldsymbol{\pi}^{(d)} \sim \mathrm{HTSSB}(\widetilde{\boldsymbol{\pi}})$ を生成。
4. 文書 d の i 番目の単語に対し、トピックを $z_{di} \sim \boldsymbol{\pi}^{(d)}$ と生成し、次に単語 $w_{di} \sim H_{z_{di}}$ を生成する。

5 モデルの推論

トピック木構造に関わる式 (2) と (3) の縦横パラメータを推定するために、ディリクレ過程表現の一つである中華料理街過程 [20] を用いる。トピック $\boldsymbol{\epsilon}$ において、$n_0(\boldsymbol{\epsilon})$ が縦方向の停止数を、$m_0(\boldsymbol{\epsilon})$ が横方向の停止数を、$n_1(\boldsymbol{\epsilon})$ が縦方向の通過数を、$m_1(\boldsymbol{\epsilon})$ が横方向の通過数を表すとする。また $n(\boldsymbol{\epsilon}) = n_0(\boldsymbol{\epsilon}) + n_1(\boldsymbol{\epsilon})$、$m(\boldsymbol{\epsilon}) = m_0(\boldsymbol{\epsilon}) + m_1(\boldsymbol{\epsilon})$ と定義する。するとデータと他のパラメータで条件づけたパラメータの期待値がそれぞれ $\widehat{\nu}_{\boldsymbol{\epsilon}} = \mathbb{E}[\nu_{\boldsymbol{\epsilon}}|\mathrm{rest}] = (1+n_0(\boldsymbol{\epsilon}))/(1+\alpha+n(\boldsymbol{\epsilon}))$ と $\widehat{\psi}_{\boldsymbol{\epsilon}} = \mathbb{E}[\psi_{\boldsymbol{\epsilon}}|\mathrm{rest}] = (1+m_0(\boldsymbol{\epsilon}))/(1+\gamma+m(\boldsymbol{\epsilon}))$ となる。ここで式 (1) を用いることで、$\pi_{\boldsymbol{\epsilon}}$ の事後確率の期待値が求まる。さらに HTSSB においても同様に、文書 d における縦横パラメータの事後確率の期待値が

$$\mathbb{E}\left[\nu_{\boldsymbol{\epsilon}}^{(d)}|\mathrm{rest}\right] = \frac{a\tau_{\tilde{\boldsymbol{\epsilon}}} + n_0(\boldsymbol{\epsilon})}{a(1-\sum_{\boldsymbol{\kappa}<\tilde{\boldsymbol{\epsilon}}}\tau_{\boldsymbol{\kappa}}) + n(\boldsymbol{\epsilon})},$$

$$\mathbb{E}\left[\psi_{\boldsymbol{\epsilon}k}^{(d)}|\mathrm{rest}\right] = \frac{b\phi_{\tilde{\boldsymbol{\epsilon}}k} + m_0(\boldsymbol{\epsilon}k)}{b(1-\sum_{j=1}^{k-1}\phi_{\tilde{\boldsymbol{\epsilon}}'j}) + m(\boldsymbol{\epsilon}k)}$$

であることがわかる。親 TSSB の停止・通過回数の更新については、付録 B に詳述した。

次に、トピックの割り当てについてはレトロスペクティブサンプリング [21] と二分探索を組み合わせたギブスサンプリングを用いた。スライスサンプリングは現在のトピックと、ランダムに選ばれたトピックを比較するため、任意の $u \sim \mathrm{Unif}[0,1)$ に対応するトピックを見つけることさえできれば、全てのトピックを数え上げる必要がなく効率的である。付録 C に具体的なアルゴリズムを示した。

HPY のパラメータ推定は原論文 [17] に従い、その他はスライスサンプリング [22, 23] を用いた。

6 実験

設定 実験には *BBC News* [24]、*20News* [25]、独自の *Wikipedia* コーパスを用いた。文書数はそれぞれ 2,225、18,828、50,513 であり、ランダムに選んだ 80%を訓練データとした。

比較対象は nCRP [9]、rCRP [12]、TSNTM [14]、nTSNTM [15] で、それぞれ既定値を用いたが、10,000 イテレーションに 2 週間以上かかったモデルは結果から除外した。全てのモデルを同一基準で比較するために、各トピックは少なくとも 100 語に割り当てられているものとした。

量的比較 既存研究で用いられているトピック固有性 (TU)[26, 27, 15]、平均重複 (AO)[15] に加えて、新しく木構造多様性 (TD) を提案する。TU は大きいほどトピックが平均的に固有の単語を上位語に持ち、AO は小さいほど親子の上位語の重複が少ない。ただし親子トピックは意味的な重なりを持つはずなので、上位語の重複が低いとき解釈性が高いかは慎重に検討せねばならない。TD は子トピックの固有性を親トピックの重要性で重み付けしたもので、以下のように定義する。

$$\mathrm{TD} = \sum_{\boldsymbol{\epsilon}\in\mathcal{T}} w_{\boldsymbol{\epsilon}} \frac{|\mathcal{V}_{\mathcal{C}(\boldsymbol{\epsilon})}|}{u|\mathcal{C}(\boldsymbol{\epsilon})|}\ ;\ w_{\boldsymbol{\epsilon}} = \frac{|\mathcal{D}(\boldsymbol{\epsilon})|}{\sum_{\boldsymbol{\kappa}\in\mathcal{T}}|\mathcal{D}(\boldsymbol{\kappa})|}.$$

ここで、$\mathcal{C}(\boldsymbol{\epsilon})$ は $\boldsymbol{\epsilon}$ の子トピックの集合、$\mathcal{D}(\boldsymbol{\epsilon})$ は $\boldsymbol{\epsilon}$ の子孫の集合、$\mathcal{V}_N$ はトピック集合 $\mathcal{N}$ の上位 u 語から重複を除いた単語集合である。TD が大きいほど、子トピックが固有の単語を含む。

表 1 に各指標を 5, 10, 15 の上位語数で計算した平均をまとめた。提案手法は TD と TU で既存手法を上回り、割り当てが 100 語未満のトピックを含めても高い性能を示している (括弧内の値)。既存の確率的手法 (nCRP と rCRP) は、木が深さ 3 で打ち切られ

モデル	最大深さ	木構造多様性 (TD) (↑)			トピック固有性 (TU) (↑)			平均重複 (AO) (↓)			トピック数		
		BBC	20N	Wiki	BBC	20N	Wiki	BBC	20N	Wiki	BBC	20N	Wiki
ihLDA	3	2.24	**2.88**	**2.63**	**0.60**	**0.82**	**0.66**	0.28	0.11	0.16	38	27	17
		(2.24)	(2.86)	(2.49)	(0.60)	(0.80)	(0.63)	(0.28)	(0.14)	(0.19)	(38)	(31)	(18)
	≥ 4	**2.53**	**2.88**	2.50	0.55	0.76	0.65	0.26	0.12	0.15	85	67	73
		(2.54)	(2.80)	(2.51)	(0.49)	(0.51)	(0.63)	(0.30)	(0.38)	(0.16)	(134)	(203)	(101)
nCRP	3	1.92	2.16	–	0.36	0.32	–	**0.03**	0.02	–	517	2108	–
rCRP		0.15	–	–	0.01	–	–	0.53	–	–	278	–	–
TSNTM		1.98	2.54	2.47	0.43	0.80	0.64	0.26	0.09	0.06	22	41	44
nTSNTM		2.11	2.57	2.34	0.46	0.68	0.60	0.09	**0.01**	**0.02**	68	81	111

表 1　量的比較結果。比較のため 100 語未満にしか割り当てられなかったトピックを除外したが、提案手法については除外前の値を括弧内に示した。

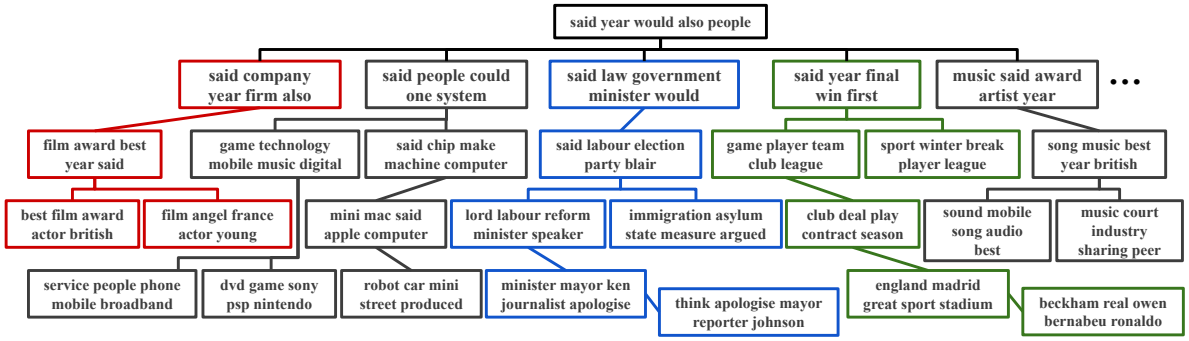

図 5　BBC コーパスを用いた深さ 6 の木の上位語。色付きの枝を図 6 でさらに説明している。

ているにも関わらず大量のトピックを推定した。

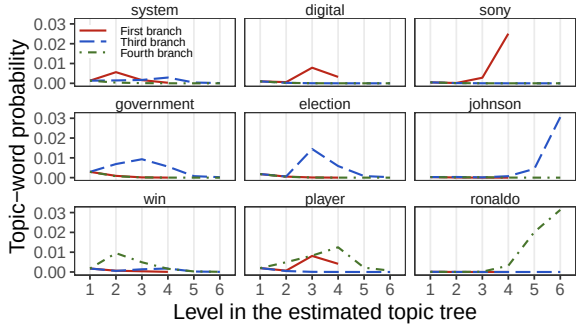

図 6　図 5 の左から 1 番目 (赤)、3 番目 (青)、4 番目 (緑) の枝のトピック–単語確率。固有名詞の確率はノードが深いほど高く、一般的な単語はノードが浅いほど高い。

木構造と上位語　図 5 に推定された深さ 6 の木と上位語を、図 6 にいくつかのトピック–単語確率を示した。固有名詞の確率はノードが深いほど高く、一般的な単語の確率はノードが浅いほど高い。またそれぞれの枝が固有のテーマを表していることが読みとれる。

7　関連研究

既存研究では確率の小さなトピックが作られることを防ぐために、深さ 3 の浅い木のみの推定や [9, 12, 13]、経験則に基づくトピック生成の制約 [14, 15] がなされてきた。これに対し提案手法は、深い木構造でも打ち切りなしに妥当なトピック数を推定できる (表 1 の括弧内)。また提案手法は TSSB の階層モデルであるから、単語を生成するトピックが既存のモデルのように文書ごとに制約されることがない [9, 10, 13]。

階層的トピックモデルは様々なモデルに拡張されており [28, 29, 30, 31, 32]、棒折り過程を使う既存手法に提案手法を容易に反映させることができる。

8　結語

既存の階層的トピックモデルには深くなるにつれてトピックの確率が指数的に減衰する問題があった。親トピックを考慮することでこの問題を解消する提案手法は、棒折り過程を用いる既存モデルにも広く応用可能である。実験は 2 つの指標で提案手法がニューラルを含む既存手法を上回ることを示し、定性的な検証でも $\ell \geq 4$ の深い木の推定が行えることがわかった。

一方、本研究で用いたギブスサンプリングは近似を伴う既存手法 [33] よりも遅いことがわかっている。将来的には分散アルゴリズムや変分推論を組み合わせ、より大規模なコーパスを処理できるようにしたい。また、本研究の教師なし学習で推定される木構造を利用者の考える分類と一致させるため、非階層的トピックモデルで提案されている半教師あり学習 [34, 35] を取り入れることも今後の課題である。

参考文献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[2] David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[3] Jonathan Chang and David M Blei. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, Vol. 4, pp. 124–150, 2010.

[4] Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, Vol. 111, pp. 988–1003, 2016.

[5] Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J Smola. Reducing the sampling complexity of topic models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 891–900, 2014.

[6] Hsiang-Fu Yu, Cho-Jui Hsieh, Hyokun Yun, S V N Vishwanathan, and Inderjit S Dhillon. A scalable asynchronous distributed algorithm for topic modeling. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1340–1350, 2015.

[7] Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric P. Xing, Tie Yan Liu, and Wei Ying Ma. LightLDA: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 1351–1361, may 2015.

[8] Jianfei Chen, Kaiwei Li, Jun Zhu, and Wenguang Chen. WarpLDA: A cache efficient O(1) algorithm for latent Dirichlet allocation. *Proceedings of the VLDB Endowment*, Vol. 9, , 2016.

[9] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pp. 17–24, 2003.

[10] David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested Chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, Vol. 57, No. 2, pp. 1–30, feb 2010.

[11] Ryan P. Adams, Zoubin Ghahramani, and Michael I. Jordan. Tree-structured stick breaking for hierarchical data. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, pp. 19–27, 2010.

[12] Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice Oh. Modeling topic hierarchies with the recursive Chinese restaurant process. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 783–792. ACM Press, 2012.

[13] John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. Nested hierarchical Dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 2, pp. 256–270, feb 2015.

[14] Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. Tree-structured neural topic model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 800–806, 2020.

[15] Ziye Chen, Cheng Ding, Zusheng Zhang, Yanghui Rao, and Haoran Xie. Tree-structured topic modeling with nonparametric neural variational inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 2343–2353, 2021.

[16] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, Vol. 4, No. 2, pp. 639–650, 1994.

[17] Yee Whye Teh. A Bayesian interpretation of interpolated Kneser-Ney. Technical report, National University of Singapore School of Computing, 2006.

[18] 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, 2016.

[19] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1566–1581, 2006.

[20] John Paisley and Lawrence Carin. Hidden markov models with stick-breaking priors. *IEEE Transactions on Signal Processing*, Vol. 57, No. 10, pp. 3905–3917, 2009.

[21] Omiros Papaspiliopoulos and Gareth O. Roberts. Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, Vol. 95, No. 1, pp. 169–186, 2008.

[22] Radford M Neal. Slice sampling. *The Annals of Statistics*, Vol. 31, pp. 705–767, 2003.

[23] Daichi Mochihashi. Unbounded slice sampling. *ISM Research Memorandum No. 1209*, 2020.

[24] Derek Greene and Pádraig Cunningham. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd International Conference on Machine learning*, pp. 377–384. ACM Press, 2006.

[25] Ken Lang. NewsWeeder: Learning to filter Netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 331–339, 1995.

[26] Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[27] Corentin Masson and Syrielle Montariol. Detecting omissions of risk factors in company annual reports. In *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*, pp. 15–21, 2020.

[28] Su Jin Shin and Il Chul Moon. Guided HTM: Hierarchical topic model with dirichlet forest priors. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 29, No. 2, pp. 330–343, feb 2017.

[29] Yueshen Xu, Jianwei Yin, Jianbin Huang, and Yuyu Yin. Hierarchical topic modeling with automatic knowledge mining. *Expert Systems with Applications*, Vol. 103, pp. 106–117, aug 2018.

[30] Ming Yang and William H. Hsu. HDPauthor: A new hybrid author-topic model using latent Dirichlet allocation and hierarchical Dirichlet processes. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 619–624, 2016.

[31] Xi Zou, Yuelong Zhu, Jun Feng, Jiamin Lu, and Xiaodong Li. A novel hierarchical topic model for horizontal topic expansion with observed label information. *IEEE Access*, Vol. 7, pp. 184242–184253, 2019.

[32] Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. Unsupervised abstractive opinion summarization by generating sentences with tree-structured topic guidance. *Transactions of the Association for Computational Linguistics*, Vol. 9, pp. 945–961, 2021.

[33] Diederik P Kingma and Max Welling. Stochastic gradient vb and the variational auto-encoder. In *Proceedings of the 2nd International Conference on Learning Representations*, Vol. 19, p. 121, 2014.

[34] Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 204–213, 2012.

[35] Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. Keyword assisted topic models. *arXiv 2004.05964*, 2020.

A　階層 TSSB でのトピック平均確率

棒降り過程 $\phi_k = v_k \prod_{j=1}^{k-1}(1-v_j)$, $v_k \sim \mathrm{Be}(1,\gamma)$ において、v_k の平均確率は $\mathbb{E}[v_k] = 1/(1+\gamma)$ であるから、k 番目に折られた棒の平均確率は以下となる、

$$\mathbb{E}[\phi_k] = \frac{1}{1+\gamma}\left(\frac{\gamma}{1+\gamma}\right)^{k-1} = \frac{1}{\gamma}\left(\frac{\gamma}{1+\gamma}\right)^{k}.$$

次に棒折り過程におけるトピックの平均確率は、

$$\begin{aligned}
\mathbb{E}[\phi] &= \sum_{k=1}^{\infty} \mathbb{E}[\phi_k]\cdot\phi_k \\
&\approx \sum_{k=1}^{\infty} \mathbb{E}[\phi_k]^2 \\
&= \sum_{k=1}^{\infty} \left(\frac{\gamma}{\gamma+1}\cdot\frac{1}{\gamma}\right)^2 \\
&= \frac{1}{\gamma^2}\sum_{k=1}^{\infty}\left(\frac{1}{\left(1+\frac{1}{\gamma}\right)^2}\right)^k \\
&= \frac{1}{2\gamma+1}
\end{aligned}$$

となる。ここでは ϕ_k の平均を近似として用いた。棒折り過程を用いると、深さ ℓ におけるトピックの平均確率は

$$\mathbb{E}[\phi\,|\,\ell] = \mathbb{E}[\phi\,|\,\ell-1]\cdot\mathbb{E}[\phi] \approx \frac{1}{(2\gamma+1)^{\ell}},$$

となる。最初の等号は深さ $\ell-1$ での棒が更に深さ ℓ において折られることを示している。3 節に示した提案手法では、平均確率が

$$\begin{aligned}
\mathbb{E}[\phi\,|\,\ell] &= \mathbb{E}[\phi\,|\,\ell-1]\cdot\mathbb{E}[\phi] \\
&\approx \mathbb{E}[\phi\,|\,\ell-1]\cdot\frac{1}{2(\gamma\cdot\mathbb{E}[\phi\,|\,\ell-1])+1} \\
&= \frac{1}{2\gamma+1/\mathbb{E}[\phi\,|\,\ell-1]} \quad \text{for } \ell\geq 2\,.
\end{aligned}$$

となる。木構造の根 ($\ell=1$) においては $\mathbb{E}[\phi\,|\,\ell=1] = 1/(2\gamma+1)$ である。木構造が深くなっても、トピックの平均確率が指数的に小さくなることはない。

B　HTSSB での停止・通過回数

単語 w が子 TSSB のトピック $\boldsymbol{\epsilon}$ で縦方向の停止をした場合、$a\nu_{\tilde{\boldsymbol{\epsilon}}}/(a+n(\boldsymbol{\epsilon}))$ に比例する確率で親 TSSB の対応ノード $\tilde{\boldsymbol{\epsilon}}$ を更新し、$n(\boldsymbol{\epsilon})/(a+n(\boldsymbol{\epsilon}))$ に比例する確率で子 TSSB の $\boldsymbol{\epsilon}$ のみを更新する。横方向についても同様に $b\psi_{\tilde{\boldsymbol{\epsilon}}}/(b+m(\boldsymbol{\epsilon}))$ に比例する確率で親 TSSB の対応ノード $\tilde{\boldsymbol{\epsilon}}$ を更新し、$m(\boldsymbol{\epsilon})/(b+m(\boldsymbol{\epsilon}))$ に比例する確率で子 TSSB の $\boldsymbol{\epsilon}$ のみを更新する。

C　推論アルゴリズム

トピックの探索と割り当ては以下の通りである。

Algorithm 1 トピックの探索

function `find_topic`$(u, \boldsymbol{\epsilon})$
if $u < \nu_{\boldsymbol{\epsilon}}$ **then**
　return $\boldsymbol{\epsilon}$
else
　$u = (u-\nu_{\boldsymbol{\epsilon}})/(1-\nu_{\boldsymbol{\epsilon}})$; $k = 1$
　while True **do**
　　if $u < 1 - \prod_{j=1}^{k}(1-\psi_{\boldsymbol{\epsilon}j})$ **then break**
　　else
　　　k += 1
　　　Create $\boldsymbol{\epsilon}k$ if necessary
　　end if
　end while
　$u = \dfrac{(u-1)(1-\psi_{\boldsymbol{\epsilon}k}) + \prod_{j=1}^{k}(1-\psi_{\boldsymbol{\epsilon}j})}{\psi_{\boldsymbol{\epsilon}k}\cdot\prod_{j=1}^{k}(1-\psi_{\boldsymbol{\epsilon}j})}$
　return find_topic$(u, \boldsymbol{\epsilon}k)$
end if
end function

Algorithm 2 トピックのギブスサンプリング

function `sample_assignment`$(\boldsymbol{\epsilon})$
$a = 0$; $b = 1$; $\rho = \mathrm{Unif}[0,1)\cdot p(\boldsymbol{\epsilon})$
while True **do**
　$u = \mathrm{Unif}[a, b)$
　$\boldsymbol{\epsilon}' =$ find_topic$(u, \boldsymbol{\epsilon}_{\mathrm{root}})$
　$q = p(\boldsymbol{\epsilon}')$
　if $q > \rho$ **then return** $\boldsymbol{\epsilon}'$
　else
　　if $\boldsymbol{\epsilon}' < \boldsymbol{\epsilon}$ **then** $b = u$ **else** $a = u$
　end if
end while
end function

D　評価指標の詳細

トピック固有性 (TU)　定義は $\mathrm{TU} = |\mathcal{T}|^{-1}\sum_{\boldsymbol{\epsilon}\in\mathcal{T}}\left(u^{-1}\sum_{u'=1}^{u} n(u',\boldsymbol{\epsilon})^{-1}\right)$ である。ここで $\mathcal{T}$ は推定されたトピック集合、$n(u',\boldsymbol{\epsilon})$ は $\boldsymbol{\epsilon}$ の u' 番目の上位単語が、全てのトピックでの上位 u 語に登場した回数である。TU が大きいほどトピックの固有性が高い。

平均重複 (AO)　AO は親子トピック間で上位 u 語の重複を平均し、$\mathrm{AO} = |\mathcal{T}|^{-1}\sum_{\boldsymbol{\epsilon}\in\mathcal{T}} |\mathcal{V}_{\boldsymbol{\epsilon}}\cap\mathcal{V}_{\boldsymbol{\epsilon}'}|\cdot u^{-1}$ と定義される。ここで $\mathcal{V}_{\boldsymbol{\epsilon}}$ は $\boldsymbol{\epsilon}$ の上位 u 語である。低い AO は親子関係の単語の重複が少ないことを示す。