時系列構造化ニューラルトピックモデル

宮本 望 [1]　磯沼 大 [1]　*高瀬 翔 [2]　森 純一郎 [1,3]　坂田 一郎 [1]
[1] 東京大学　[2] 東京工業大学　[3] 理研 AIP
{nmiyamoto, isonuma, isakata}@ipr-ctr.t.u-tokyo.ac.jp
sho.takase@linecorp.com　mori@mi.u-tokyo.ac.jp

概要

本研究ではトピック間の依存関係を捉えつつ、その時系列的発展を扱うことができる時系列構造化ニューラルトピックモデルを提案する。本モデルは、トピックの依存関係を self-attention 機構に基づいてモデル化することで、トピックの分化・統合過程を捉える。さらに、アテンションの重みが文書間の引用関係を反映するように、引用正則化項を導入する。本モデルは Perplexity や Coherence において、既存の時系列トピックモデル [1] を上回ることを確認した。また、実際の論文群を用いて、本モデルがトピックの遷移過程を捉えられることを検証した。

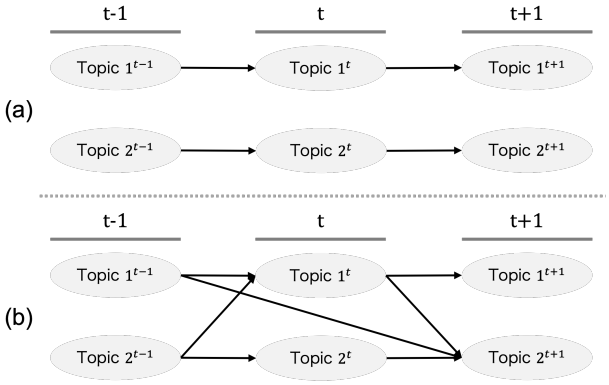

図 1　(a) 時系列トピックモデルと (b) 提案モデルの比較。

1　はじめに

文書のトピックとその比率を推定するトピックモデルは、Latent Dirichlet Allocation (**LDA** [2]) をはじめ、自然言語処理において広く利用されている。その一種である時系列トピックモデル [1, 3] は、トピックの時系列上の変遷を追うことで、時間経過に伴うトピック中の単語変化の可視化を可能にした。

しかし、既存の時系列トピックモデルは、トピックの時系列変化を各トピックで独立に扱っており、新規トピックの形成に対し過去のトピックがどう寄与したのか、トピック間の依存関係を捉えることはできない（図 1(a)）。特に論文など、引用関係により依存関係にある文書のモデル化において、各トピックが独立して変化し、前の時刻のトピックにのみ依存しているという前提は適切ではない。

そこで本研究では、トピック間の依存関係を捉えながらその時系列発展を扱うことができるトピックモデルとして、時系列構造化ニューラルトピックモデルを提案する（図 1(b)）。具体的には、トピックの依存関係を self-attention 機構 [4] に基づきモデル化することで、新規トピックがどの既存トピックに基づき生まれたかを明らかにする。さらに、文書の引用関係を反映するようにアテンションの重みを正則化する引用正則化項を導入し、アテンションの重みをトピックの分化・統合過程として解釈できるようにした。これにより、複数時刻に跨った学術トピックの分化・統合など複雑な過程を定量的に捉えることが可能になる。

評価実験では、従来モデルの ETM [5] や D-ETM [1] と比較して、Perplexity と Coherence において、提案モデルがその性能を上回ることを確認した。さらに、提案モデルが学術トピックの分化・統合過程を捉えられていることを定性的に確認した。

2　事前準備

まず、本研究の基礎となる時系列トピックモデルとして、Dynamic Embedded Topic Model (**D-ETM** [1]) を解説する。D-ETM はトピックの時系列変化をモデル化することで、時系列文書の解析を行う。具体的には、単語の意味空間におけるトピックごとの埋め込み表現（トピック埋め込み）とトピック比率の平均を変化させることで、トピックの時間変化を捉

* 現在の所属は LINE 株式会社.

える。D-ETM の生成プロセスを以下に示す。

1. 時刻 $t \in \{1, \ldots, T\}$ について:

k 番目のトピック埋め込みのサンプル:

$$\boldsymbol{\alpha}_k^{(t)} \sim \mathcal{N}(\boldsymbol{\alpha}_k^{(t-1)}, \sigma^2 \boldsymbol{I}) \tag{1}$$

トピック比率の平均のサンプル:

$$\boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{\eta}_{t-1}, \delta^2 I) \tag{2}$$

2. 各文書 $d \in \{1, \ldots, D\}$ について:

$$\text{トピック比率のサンプル: } \boldsymbol{\theta}_d \sim \mathcal{LN}(\boldsymbol{\eta}_{t_d}, \gamma^2 \boldsymbol{I}) \tag{3}$$

3. 文書 d 中の各単語 $n \in \{1, \ldots, N_d\}$ について:

$$\text{トピックの割り当て: } z_{d,n} \sim \text{Cat}(\boldsymbol{\theta}_d) \tag{4}$$

$$\text{単語のサンプル: } w_{d,n} \sim \text{Cat}(\boldsymbol{\beta}_{z_{d,n}}^{(t_d)}) \tag{5}$$

ただし、$\text{Cat}(\cdot)$ はカテゴリカル分布、$\mathcal{LN}(\cdot, \cdot)$ はロジット正規分布を指す。σ と δ と γ はハイパーパラメータである。単語分布 $\boldsymbol{\beta}_k^{(t)}$ は以下の式で計算される。

$$\boldsymbol{\beta}_k^{(t)} = \text{softmax}(\boldsymbol{\rho}^\top \boldsymbol{\alpha}_k^{(t)}) \tag{6}$$

$\boldsymbol{\rho} \in \mathbb{R}^{L \times V}$ は L 次元の単語埋め込み行列であり、$\boldsymbol{\rho}_v \in \mathbb{R}^L$ は v 番目の単語埋め込みに対応する。

D-ETM では、変分推論 [6] で $\boldsymbol{\alpha}$ と $\boldsymbol{\eta}$ と $\boldsymbol{\theta}$ の事後分布を推定する。特に、トピック埋め込み $\boldsymbol{\alpha}$ については、以下の平均場近似を用いる。

$$q(\boldsymbol{\alpha}_k^{(t)}) = \mathcal{N}(\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\sigma}_k^{(t)}) \tag{7}$$

$$q(\boldsymbol{\alpha}) = \prod_k \prod_t q(\boldsymbol{\alpha}_k^{(t)}) \tag{8}$$

ただし、$\boldsymbol{\alpha}_k^{(t)}$ は平均 $\boldsymbol{\mu}_k^{(t)}$ と分散 $\boldsymbol{\sigma}_k^{(t)}$ のガウス分布に従いサンプルされる。

この平均場近似は、事後分布において各トピックが独立であることを仮定しており、トピック間の依存関係を推定することができない。一方、本研究では、self-attention 機構によりトピック埋め込みを推定する構造化変分推論を導入し、トピック間の依存関係をモデル化する。

3 提案モデル

本章では、提案モデルである時系列構造化ニューラルトピックモデルについて述べる。生成プロセスは D-ETM と同一である。

3.1 トピック埋め込みの推論

D-ETM とは異なり、提案モデルは構造化変分推論によりトピック埋め込みを推論する。self-attention

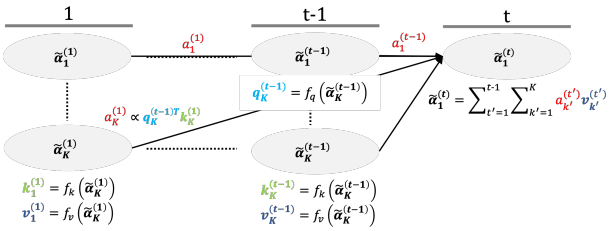

図 2 self-attention によるトピック埋め込みの素の作成。

機構を利用して、過去の全てのトピック埋め込みからトピック埋め込みを計算する。

$$\tilde{\boldsymbol{\alpha}}_k^{(t)} = \text{self} - \text{attention}(\tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)}) \tag{9}$$

$$q(\boldsymbol{\alpha}_k^{(t)} | \tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)}) = \mathcal{N}(f_\mu(\tilde{\boldsymbol{\alpha}}_k^{(t)}), f_\sigma(\tilde{\boldsymbol{\alpha}}_k^{(t)})) \tag{10}$$

変数 $\tilde{\boldsymbol{\alpha}}_k^{(t)}$ はトピック埋め込みの素であり、f_μ と f_σ は $\tilde{\boldsymbol{\alpha}}_k^{(t+1)}$ を変分ガウス分布に変換する関数である。

self-attention の計算 式 (9) を計算するため、時刻 $t-1$ 以前の全てのトピック埋め込みの素 $\tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)}$ からキー $\boldsymbol{K}_{1:K}^{(1:t-1)} \in \mathbb{R}^{K(t-1) \times L}$ 及びバリュー $\boldsymbol{V}_{1:K}^{(1:t-1)} \in \mathbb{R}^{K(t-1) \times L}$ を計算する。また、時刻 $t-1$ の k 番目のトピック埋め込み $\tilde{\boldsymbol{\alpha}}_k^{(t)}$ からクエリ $\boldsymbol{q}_k^{(t-1)} \in \mathbb{R}^L$ を計算する。

$$\boldsymbol{K}_{1:K}^{(1:t-1)} = f_k(\tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)}) \tag{11}$$

$$\boldsymbol{V}_{1:K}^{(1:t-1)} = f_v(\tilde{\boldsymbol{\alpha}}_{1:K}^{(1:t-1)}) \tag{12}$$

$$\boldsymbol{q}_k^{(t-1)} = f_q(\tilde{\boldsymbol{\alpha}}_k^{(t-1)}) \tag{13}$$

ただし、f_q, f_k, f_v は多層パーセプトロンである。このとき、時刻 t のトピック k に対するアテンションの重み $\boldsymbol{a}_k^{(t)} \in \mathbb{R}^{K(t-1)}$ は、以下の式で表される。

$$\boldsymbol{a}_k^{(t)} = \text{softmax}(\boldsymbol{q}_k^{(t-1)} \boldsymbol{K}_{1:K}^{(1:t-1)\top}) \tag{14}$$

$\boldsymbol{a}_k^{(t)}$ は時刻 t 以前の各トピックに依存している確率を表しているとみなせる。これを重みとして利用し、$\boldsymbol{V}^{(1:t-1)}$ の重み和を計算することで、$\tilde{\boldsymbol{\alpha}}_k^{(t)}$ を求める。一連の過程を図 2 に示す。

$$\tilde{\boldsymbol{\alpha}}_k^{(t)} = \boldsymbol{a}_{1:K}^{(1:t-1)} \boldsymbol{V}_{1:K}^{(1:t-1)} \tag{15}$$

$\tilde{\boldsymbol{\alpha}}_k^{(t)}$ は、残差接続 [7] を経て正規化される。残差接続を行うことで、勾配爆発や勾配消失を防ぐことが可能になる。また、事前分布では $\boldsymbol{\alpha}_k^{(t)}$ は $\boldsymbol{\alpha}_k^{(t-1)}$ を平均とする正規分布により生成されると仮定しているため、時刻 t のトピック k は時刻 $t-1$ のトピック k に比較的類似したものが推定される。

self-attention の導入の経緯 self-attention 機構の代わりに、トピック間の依存関係を表す学習可能な重み $\boldsymbol{w} \in \mathbb{R}^K$ を用いて、$\tilde{\boldsymbol{\alpha}}_k^{(t)} = \boldsymbol{w}^\top \tilde{\boldsymbol{\alpha}}_{1:K}^{(t-1)}$ としてパラメータ化する方法も本研究では検討した。しかし、

この方法では 1 時刻間のトピック間の依存関係しかモデル化できず、学術論文で多く見られる複数時刻に跨るトピック間の依存関係のモデル化には不適当である。一方、self-attention 機構は任意の数のトピックを入力として用いることが可能であり、かつ各時刻を通じて f_q, f_k, f_v のパラメータが共有されるため、パラメータ数はトピック数に依らず一定である。よって、self-attention 機構は複数時刻に跨るトピック間の依存関係のモデル化により適当であることから、本研究では self-attention 機構を採用した。

3.2 全体の推論と ELBO

提案モデルによる文書の尤度は以下のように与えられる。

$$
\begin{aligned}
&p(\boldsymbol{w}_{1:D}|\sigma,\delta,\gamma)\\
&=\int\Big\{\prod_d\prod_n(\boldsymbol{\beta}^{(t_d)}\cdot\boldsymbol{\theta}_d)_{w_{d,n}}p(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d})\Big\}\\
&\quad\Big\{\prod_t\prod_k p(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1})p(\boldsymbol{\alpha}_k^t|\boldsymbol{\alpha}_k^{t-1})\Big\}d\boldsymbol{\theta}d\boldsymbol{\eta}d\boldsymbol{\alpha} \qquad (16)
\end{aligned}
$$

ここで、事後分布 $p(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\alpha}|\boldsymbol{w}_{1:D})$ の近似事後分布として $q(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\alpha})$ を導入する。D-ETM と同様に、$q(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\alpha})$ は以下のように計算される。

$$
\begin{aligned}
q(\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\alpha}) &= \prod_d q(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d},\boldsymbol{w}_d)\times\prod_t q(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1},\tilde{\boldsymbol{w}}_t)\\
&\quad\times\prod_t\prod_k q(\boldsymbol{\alpha}_k^{(t)}|\boldsymbol{\alpha}_{1:K}^{(1:t-1)}) \qquad (17)
\end{aligned}
$$

ただし、$\boldsymbol{w}_d$ は文書 d の Bag-of-Words 表現を、$\tilde{\boldsymbol{w}}_t$ は正規化された時刻 t における全文書の Bag-of-Words 表現を表す。文書の対数尤度に対する変分下限は以下の式で求められる。

$$
\begin{aligned}
L_{doc} &= \sum_d \mathrm{E}_{q(\boldsymbol{\theta}_d)q(\boldsymbol{\eta}_t)q(\boldsymbol{\alpha}_k^{(t)})}\Big[\boldsymbol{w}_d^\top\log(\boldsymbol{\beta}^{(t_d)}\cdot\boldsymbol{\theta}_d)\Big]\\
&-\sum_d \mathrm{D}_{\mathrm{KL}}\Big[q(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d},\boldsymbol{w}_d)||p(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d})\Big]\\
&-\sum_t \mathrm{D}_{\mathrm{KL}}\Big[q(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1},\tilde{\boldsymbol{w}}_t)||p(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1})\Big]\\
&-\sum_t\sum_k \mathrm{D}_{\mathrm{KL}}\Big[q(\boldsymbol{\alpha}_k^{(t)}|\boldsymbol{\alpha}_{1:K}^{(1:t-1)})||p(\boldsymbol{\alpha}_k^{(t)}|\boldsymbol{\alpha}_k^{(t-1)})\Big] \qquad (18)
\end{aligned}
$$

4 引用正則化項ありの提案モデル

3 章で述べた提案モデルは、トピック間のアテンションの重みを解釈することが困難である。理想的には、アテンションはトピック間の潜在的な依存関係を表し、それは文書間の引用関係として顕在化されるべきである。そこで、アテンションの重みを引用関係に対応するように正則化を行うことで、アテンションの解釈可能性を担保する。また、引用正則化項によりテキストと引用を同時にモデル化することで、トピックの質的な向上が期待される。

アテンションの重みの正則化にあたり、トピック比率 $\boldsymbol{\theta}$、アテンションの重み $\boldsymbol{a}$、文書比率 $\boldsymbol{\phi}$ に基づいて文書間の引用をモデル化する。文書のペア $(i,j)\in\{1,\dots,D\}\times\{1,\dots,D\}$ 間の引用は以下のようにモデル化される。

1. 引用トピックの割り当て: $z_i \sim \mathrm{Cat}(\boldsymbol{\theta}_i)$ (19)
2. 被引用トピックの割り当て: $z_j \sim \mathrm{Cat}(\boldsymbol{a}_{z_i}^{(t_i)})$ (20)
3. 被引用文書のサンプル: $d_j \sim \mathrm{Cat}(\boldsymbol{\phi}_{z_j})$ (21)

ただし、$\boldsymbol{a}_k^{(t_i)}\in\mathbb{R}^{K(t_i-1)}$ はアテンションの重みであり、過去の全トピック $z_j\in\{1,\dots,K\}\times\{1,\dots,t_i-1\}$ に依存する確率の分布を指す。文書比率 $\boldsymbol{\phi}_k\in\mathbb{R}^D$ は、あるトピックが割り当てられた被引用文書の確率分布を示しており、以下の式で表される。

$$
\phi_{z_j}^{(d_j)} = \frac{\theta_{d_j}^{(z_j)}}{\sum_{d_j}\theta_{d_j}^{(z_j)}} \qquad (22)
$$

この仮定のもと、引用 $c_{i,j}\in\{0,1\}$ の尤度は以下のように与えられる。

$$
\begin{aligned}
&p(c_{i,j}=1|\boldsymbol{\theta},\boldsymbol{\alpha})\\
&=\sum_k^K\sum_{k'}^K p(d_j|\boldsymbol{\phi}_{k'})p(z_j=k'|\boldsymbol{a}_k^{(t_i)})p(z_i=k|\boldsymbol{\theta}_i) \qquad (23)
\end{aligned}
$$

ただし、$c_{i,j}=1$ は文書 d_i が文書 d_j を引用することを示す。文書と引用の対数尤度に対する変分下限は以下の式で表される。

$$
L = L_{doc} + \sum_i^D\sum_j^D \mathrm{BCE}(p(c_{i,j}|\boldsymbol{\theta},\boldsymbol{\alpha}),c_{i,j}) \qquad (24)
$$

ただし、L_{doc} は式 (18) により定義され、BCE は二値の交差エントロピーである。詳細な変分下限の導出は付録 A.1 に記す。

5 実験

5.1 実験設定

本実験では、*The Semantic Scholar Open Research Corpus* (S2ORC [8])[1] をもとに構築した *ACL* 及び *CS* のデータセットを用いる。*ACL* データセットは ACL 系列の国際会議で発表された論文群であり、*CS* デー

1) https://github.com/allenai/s2orc

表 1　各モデルの評価結果。Perplexity は低いほど良く、Coherence と Diversity は高いほど良い。

Dataset	ACL			CS		
	Perplexity	Coherence	Diversity	Perplexity	Coherence	Diversity
ETM [5]	1,586.4	0.029	**0.896**	3,009.6	0.025	0.948
D-ETM [1]	1,150.0	0.092	0.800	2,412.5	0.085	**0.953**
提案モデル	1,080.6	0.085	0.865	2,184.0	0.077	0.919
引用正則化項ありの提案モデル	**1,066.7**	**0.097**	0.882	**2,146.7**	**0.103**	0.944

表 2　各データセットの統計量。

データセット	ACL	CS
時刻数	7	7
語彙数	5,540	10,449
訓練用文書の数	14,110	23,991
検証用文書の数	4,704	7,997
評価用文書の数	4,704	7,998

タセットは、研究分野に “Computer Science” を含む論文群である。各データセットの詳細な情報を、付録 A.2 に記載した。データセットに関する具体的な数値を表 2 にまとめる。ベースラインには、ETM [5] 及び D-ETM [1] を用いる。提案モデルに対する引用データ制約の効果を測定するため、提案モデルに対する引用正則化項の有無による結果も比較する。詳細な実験設定は付録 A.3 に記載する。

5.2　実験結果

トピックモデルの性能を以下の 3 つの基準で定量的に評価した。

Perplexity　まず、生成モデルとしての汎化能力を評価するため、Perplexity [9] を用いた。両データセットにおいて、提案モデルは、従来モデルを上回ることを示した（表 1）。また、引用正則化項ありの提案モデルは、引用正則化項なしの提案モデルを上回ることから、引用情報がトピックモデルの汎化能力に寄与することが確認された。

Coherence　トピックの解釈可能性を評価するため、トピックの Coherence を平均自己相互情報量 [10] で測定した。具体的には、トピック k の頻出語上位 10 個の単語中の各 2 単語について、正規化自己相互情報量報を計算した。各モデルの全トピックの平均 Coherence を表 1 に示す。引用正則化項ありの提案モデルが従来モデルを上回った他、引用正則化項なしのトピックモデルも D-ETM と競合し、十分にトピックが解釈可能であることが確認された。

Diversity　トピックの多様性を評価するため、全トピックの頻出語上位 25 語において重複していない単語の割合を算出し、トピック間の Diversity を測定する。各モデルにおける Diversity の平均値を表 1 に示す。両モデルは従来モデルと遜色なく、十分にトピックの多様性が担保されることを確認した。

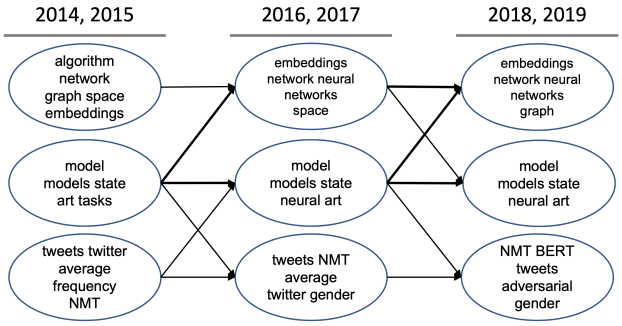

図 3　*ACL* データセットにおけるトピック遷移の例。各トピックの頻出語上位 5 つを示す。Attention の重みが 0.05 以上であれば太線、0.02 以上 0.05 未満であれば細線でトピック間の依存関係を表現した。

5.3　トピック遷移の可視化

ACL データセットを用いて、提案モデルがトピックの遷移を捉えられているか確認した。図 3 にその例を示す。上段のトピックは「グラフ」、中段のトピックは「ニューラルネットワーク」、下段のトピックは「ソーシャルメディア」を表している。トピックの頻出語を確認することで、そのトピックにおける頻出単語の変遷を追うことができる一方、アテンションの重みを確認することでトピックの分化や統合過程を追うことができる。

6　おわりに

本研究では、トピック間の依存関係を self-attention 機構でモデル化したトピックモデルを提案した。評価実験の結果、提案モデルは複数の指標について従来モデルの性能を上回った。また、文書間の引用関係を考慮した正則化を加えることで、さらなる性能の向上が示唆された。また、トピック遷移を追うことで、提案モデルが実際の学術文献のトピックの動向調査に利用できることを確認した。

謝辞

本研究は、NEDO JPNP20006、JST ACT-X JPMJAX1904 及び JST CREST JPMJCR21D1 の支援を受けたものである。

参考文献

[1] Adji B Dieng, Francisco JR Ruiz, and David M Blei. The dynamic embedded topic model. **arXiv preprint arXiv:1907.05545**, 2019.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. **Journal of machine Learning research**, Vol. 3, No. Jan, pp. 993–1022, 2003.

[3] David M. Blei and John D. Lafferty. Dynamic topic models. In **Proceedings of the 23rd International Conference on Machine Learning**, ICML '06, p. 113–120, New York, NY, USA, 2006. Association for Computing Machinery.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[5] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 439–453, 2020.

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. **Journal of the American statistical Association**, Vol. 112, No. 518, pp. 859–877, 2017.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

[8] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics.

[9] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In **Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence**, UAI '04, p. 487–494, Arlington, Virginia, USA, 2004. AUAI Press.

[10] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In **Proceedings of the 2011 conference on empirical methods in natural language processing**, pp. 262–272, 2011.

[11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. **CoRR**, Vol. arXiv:1412.6980v9, , 2014.

[12] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. **arXiv preprint arXiv:1607.06450**, 2016.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. **Advances in neural information processing systems**, Vol. 26, , 2013.

A 付録

A.1 引用正則化項ありの提案モデルの変分下限の導出

この章では、式 (24) の導出を行う。文書と引用の尤度は以下の式で計算される。

$$
\begin{aligned}
&p(\boldsymbol{w}_{1:D}, c_{1,1}, \ldots, c_{D,D}|\sigma, \delta, \gamma) \\
&= \int \left\{ \prod_i \prod_n (\boldsymbol{\beta}^{(t_i)} \cdot \boldsymbol{\theta}_i)_{w_{i,n}} p(\boldsymbol{\theta}_i|\boldsymbol{\eta}_{t_i}) \right\} \\
&\quad \left\{ \prod_i \prod_j p(c_{i,j}|\boldsymbol{\theta}, \boldsymbol{\alpha}) \right\} \\
&\quad \left\{ \prod_t \prod_k p(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1}) p(\boldsymbol{\alpha}_k^t|\boldsymbol{\alpha}_k^{t-1}) \right\} d\boldsymbol{\theta} d\boldsymbol{\eta} d\boldsymbol{\alpha} \qquad (25)
\end{aligned}
$$

近似事後分布 $q(\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\alpha})$ は、式 (17) のように計算される。よって、文書と引用の対数尤度に対する変分下限は以下のように表される。

$$
\begin{aligned}
L &= \sum_i \mathrm{E}_{q(\boldsymbol{\theta}_i) q(\boldsymbol{\eta}_t) q(\boldsymbol{\alpha}_k^{(t)})} \left[\boldsymbol{w}_i^\top \log(\boldsymbol{\beta}^{(t_i)} \cdot \boldsymbol{\theta}_i) \right] \\
&+ \sum_i \sum_j \mathrm{E}_{q(\boldsymbol{\theta}_i) q(\boldsymbol{\eta}_t) q(\boldsymbol{\alpha}_k^{(t)})} \left[\log p(c_{i,j}|\boldsymbol{\theta}, \boldsymbol{\alpha}) \right] \\
&- \sum_i \mathrm{D}_{\mathrm{KL}} \left[q(\boldsymbol{\theta}_i|\boldsymbol{\eta}_{t_i}, \boldsymbol{w}_i) || p(\boldsymbol{\theta}_i|\boldsymbol{\eta}_{t_i}) \right] \\
&- \sum_t \mathrm{D}_{\mathrm{KL}} \left[q(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1}, \tilde{\boldsymbol{w}}_t) || p(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1}) \right] \\
&- \sum_t \sum_k \mathrm{D}_{\mathrm{KL}} \left[q(\boldsymbol{\alpha}_k^{(t)}|\boldsymbol{\alpha}_{1:K}^{(1:t-1)}) || p(\boldsymbol{\alpha}_k^{(t)}|\boldsymbol{\alpha}_k^{(t-1)}) \right] \\
&= L_{doc} + \sum_i^D \sum_j^D \mathrm{BCE}(p(c_{i,j}|\boldsymbol{\theta}, \boldsymbol{\alpha}), c_{i,j})
\end{aligned}
$$

A.2 実験に用いたデータセット

本節では、各データセットの作成手順について説明する。S2ORC は、136,595,995 の論文群を保有しており、各論文は論文 ID、出版年、要約、被引用論文 ID、ACL ID、研究分野などを保持する。*ACL* データセットについては、2006 年から 2019 年までの ACL ID が “None” でない論文を抽出する。*CS* データセットについては、2006 年から 2019 年までの研究分野が “Computer Science” である論文を抽出し、被引用論文数の多い上位 40,000 本を抽出する。

得られた論文群は、3:1:1 の比率でランダムに訓練用、検証用、評価用文書データセットに分類する。訓練用文書データセットに対して、出現数が 10 未満の語彙、70%以上の論文に共通して出現する語彙、数字を除去して得られた文書を Bag-of-Words として保持する。各時刻において十分な量の文書を確保するため、2 年分の文書群をまとめて一つの時刻の文書群とする。例えば、2006 年と 2007 年の文書群をまとめて、それらを時刻 $t=0$ とする。

A.3 実装詳細

各モデルのハイパーパラメータは、*ACL* の検証データセットの Perplexity に基づいてチューニングされている。全ての実験でトピック数 K は 20 と設定し、トピック埋め込みの推論以外のパラメータは D-ETM の実装[2] に従って設定した。勾配降下法は Adam [11] を用いて、バッチサイズ 512 でモデルを学習させる。学習率については、提案モデルと ETM については 6.0×10^{-4}、D-ETM については 8.0×10^{-4} である。より良い実験結果を得るため、各モデルについて学習率の減衰を適用し、ロスが収束するまで学習を継続させている。

DSNTM 内部の self-attention 機構の f_q, f_k, f_v 及びトピック埋め込みの変分パラメータ構築の f_μ, f_σ は、全てに次元数が 300 の隠れ層の MLP を利用する。また、残差接続後の正規化は、レイヤー正規化 [12] を指す。

各モデルとも、変文推論の事後分布の分散のパラメータは $\delta^2 = \sigma^2 = 0.005$ で、$\gamma^2 = 1$ とした。単語埋め込みについては、skip-gram [13] の埋め込み表現を利用しており、その次元数は 300 である。

トピック比率 $\boldsymbol{\theta}$ の推論 $q(\boldsymbol{\theta}_d|\boldsymbol{\eta}_{t_d}, \boldsymbol{w}_d)$ は、ReLU 活性化と 800 次元の隠れ層を持つ 2 層の MLP、その後に θ_d の平均 f_μ と分散 f_σ が続く構成である。トピック比率の平均 $\boldsymbol{\eta}$ の推論 $q(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1}, \tilde{\boldsymbol{w}}_t)$ を構築するため、bag-of-words 表現 $\tilde{\boldsymbol{w}}_t$ を 400 次元の低次元空間に線形写像し、それを、400 次元の 3 層の隠れ層を持つ LSTM の入力として利用する。LSTM の出力は、直前のトピック比率の平均 $\boldsymbol{\eta}_{t-1}$ と結合され、その結果を K 次元空間に線形写像し、$\boldsymbol{\eta}_t$ の平均と分散を得る。

2) https://github.com/adjidieng/DETM