

## トピック文生成による教師なし意見要約

磯沼 大<sup>1</sup> 森 純一郎<sup>1</sup> ボレガラ ダヌシカ<sup>2</sup> 坂田 一郎<sup>1</sup><sup>1</sup> 東京大学 <sup>2</sup> リヴァプール大学

{isonuma, isakata}@ipr-ctr.t.u-tokyo.ac.jp

mori@mi.u-tokyo.ac.jp

danushka@liverpool.ac.uk

## 1 はじめに

近年、EC 上の商品レビューや SNS 投稿の急激な増加により、それらに含まれる意見を集約し、知見として利用する取り組みが注目されている。こうした意見文書の要約では、抽出型要約よりも言い換えや一般化を交える生成型要約がより有効なことが知られているが [1, 2]、既存の生成型要約手法はほぼ全て教師ありアプローチであり、大量の参照要約の用意が困難な意見文書への適用方法が課題となっている。こうした経緯から、近年意見文書に対する教師なし生成型要約が着目されている [3, 4, 5]。

教師なしアプローチでは、要約の潜在表現を参照要約なしにいかにかに獲得するかが鍵となるが、本研究では多様な詳細度合いを持つトピックを捉えることで要約文を生成する。例えば図 1 に示したあるレビューの要約は、「料理」、「サービス」、「場所」の各トピックについて詳述し、最後は「全体」について述べている。このように、要約は多様なトピックで構成され、あるトピックは詳述され、他のトピックは簡潔に記述されていることが観察される。そこで本研究では、文書から木構造上のトピックを推定し、根からは全体的な、葉に近づくにつれより詳細なトピックに関する要約文（トピック文）を生成する。トピック文から、要約として相応しいトピックと詳細度合いを持つ文を選択することで、意見文書の要約が教師なしに得られることを示す。抽出型要約においては、トピック [6, 7] やその木構造 [8, 9] の有効性が従来研究で示されているが、本研究はそれらの教師なし生成型要約への応用を提案する。

トピック文生成の文脈では、Wang らは文書中の文の潜在分布を混合ガウス分布で表現すると、その構成要素である各単峰ガウス分布がトピック文の潜在分布として機能することを明らかにした [10]。一方、本研究のように多様な詳細度合いを持つトピック文を生成するためには、文の詳細度合いを潜在空

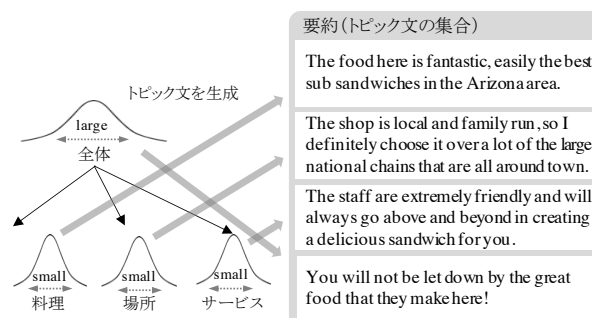


図 1 本研究の概要。右側はあるレストランレビューの人手要約であり、各文が木構造上の各トピックに対応することがわかる。トピック文の潜在分布の分散は葉に近づくほど小さくなり、より具体的なトピックに対応する。

間でモデル化する必要がある、その方法は明らかでない。そこで本研究では、トピック文の潜在表現をガウス分布で表現する際に、子の分布の分散が親よりも小さくなるようにモデルを構築することで、葉に近づくにつれより詳細な文を生成する（図 1）。単語の潜在分布としてガウス分布を用いる Gaussian word embedding では、「犬」のような具体的な単語は、「動物」といった一般的な単語よりも小さい分散を持つことが示されている [11]。文においても同様に、具体的な文は意味の分散が小さいため、その潜在分布は分散が小さいと考えられる。子の分布の分散を親よりも小さくするために、本研究では再帰的混合ガウス分布（再帰的 GMM）を文書中の文の潜在表現の事前分布として導入する。再帰的 GMM は、木構造上の各トピックに対応するガウス分布で構成され、子の事前分布に親の事後分布を設定して構築される。これにより、根の潜在分布の分散は大きく、葉に近づくにつれ分散は小さくなる。

評価実験では、提案法の要約性能が既存の教師なし生成型要約手法と競合することを確認した。また、トピック文の詳細度合いはその潜在分布の分散の大きさに依存し、葉に近づくほど分散が小さく、より具体的な文が生成されることを示した。

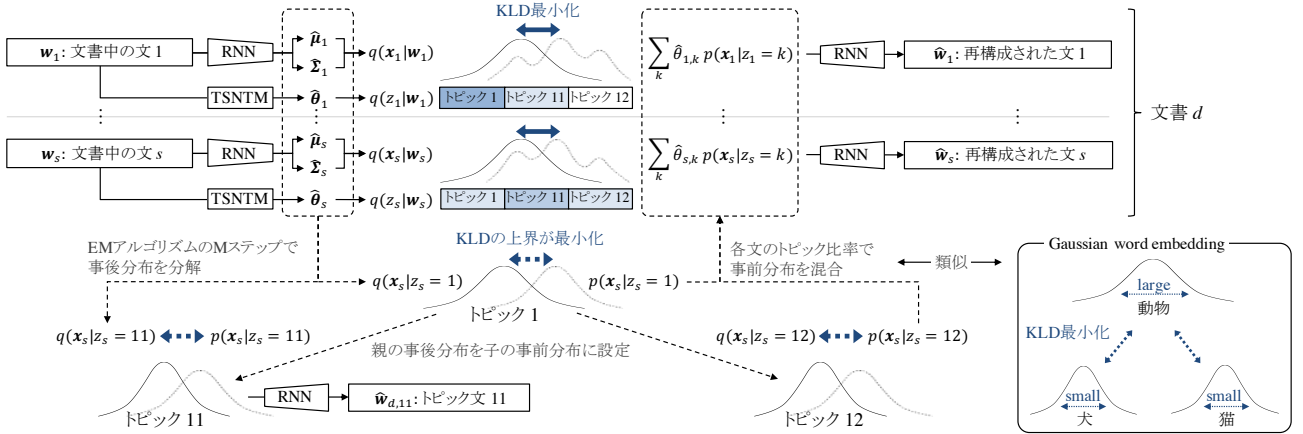


図2 提案法概要。文書の文の事前分布に再帰的 GMM を、それを構成するガウス分布をトピック文の事前分布に設定。

## 2 提案法

本節では、文書からトピック文を生成し、要約を得る方法を説明する。提案法の概要を図2に示す。

### 2.1 文書の生成過程

提案法では以下の文書の生成過程を仮定する。

各文書インデックス  $d \in \{1, \dots, D\}$  について:

文書  $d$  の各文インデックス  $s \in \{1, \dots, S_d\}$  について:

1. 文のトピック  $z_s \in \{1, 11, 12, \dots, K\}$  をサンプル:

$$z_s \sim \text{Mult}(\theta) \quad (1)$$

2. 文の潜在表現  $\mathbf{x}_s \in \mathbb{R}^H$  をサンプル:

$$\mathbf{x}_s | z_s \sim \prod_{k=1}^K p(\mathbf{x}_s | z_s = k)^{\delta(z_s=k)} \quad (2)$$

3. 尤度が最大となる文  $\mathbf{w}_s = (w_s^1, \dots, w_s^{N_s})$  を得る。

$$\mathbf{w}_s | \mathbf{x}_s = \arg \max_{\mathbf{w}_s} p(\mathbf{w}_s | \mathbf{x}_s) \quad (3)$$

ただし、 $p(\mathbf{w}_s | \mathbf{x}_s) = \prod_{n=1}^{N_s} p(w_s^n | \mathbf{w}_s^{<n}, \mathbf{x}_s)$  は RNN で得る。トピック分布は木構造上に定義され、その事前分布を一様分布に設定する:  $\theta_k = K^{-1}$ 。文書中の文の事前分布として再帰的 GMM を設定し (2)、各要素はトピック文の事前分布  $p(\mathbf{x}_s | z_s = k)$  に対応する:

$$p(\mathbf{x}_s | z_s = 1) = N(\mathbf{x}_s | \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \quad (4)$$

$$p(\mathbf{x}_s | z_s = k) = q(\mathbf{x}_s | z_s = \text{par}(k)) \\ = N(\mathbf{x}_s | \hat{\boldsymbol{\mu}}_{d, \text{par}(k)}, \hat{\boldsymbol{\Sigma}}_{d, \text{par}(k)}) \quad (k \neq 1) \quad (5)$$

ただし、 $\text{par}(k)$  はトピック  $k$  の親を示す。 $q(\mathbf{x}_s | z_s = \text{par}(k))$  は親トピック文の推定事後分布であり、子トピック文の事前分布として設定される。

この時、文書の尤度とその対数の変分下限は、それぞれ下記式で表される:

$$p(\mathbf{W}_{1:S_d}) = \prod_{s=1}^{S_d} \left\{ \int p(\mathbf{w}_s | \mathbf{x}_s) p(\mathbf{x}_s | z_s) p(z_s) d\mathbf{x}_s dz_s \right\} \quad (6)$$

$$L_d = \sum_{s=1}^{S_d} \left\{ \mathbf{E}_{q(\mathbf{x}_s | \mathbf{w}_s)} [\log p(\mathbf{w}_s | \mathbf{x}_s)] - \text{D}_{\text{KL}}[q(z_s | \mathbf{w}_s) | p(z_s)] \right\} \\ - \sum_{k=1}^K \sum_{s=1}^{S_d} \left\{ \hat{\theta}_{s,k} \text{D}_{\text{KL}}[q(\mathbf{x}_s | \mathbf{w}_s) | p(\mathbf{x}_s | z_s = k)] \right\} \quad (7)$$

ただし  $q(\mathbf{x}_s | \mathbf{w}_s) = N(\mathbf{x}_s | \hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\Sigma}}_s)$  は文  $s$  の潜在表現の事後分布で、 $\text{RNN} f_{\boldsymbol{\mu}}, f_{\boldsymbol{\Sigma}}$  により推定される。 $\hat{\theta}_{s,k} = q(z_s = k | \mathbf{w}_s)$  はトピック分布の事後分布であり、本研究では木構造ニューラルトピックモデル (TSNTM [12]) により推定する (A.1 節にて詳述)。

### 2.2 トピック文の生成

トピック文の潜在表現の事後分布を式 (8) で与え、 $\sum_{s=1}^{S_d} \mathbf{E}_{q(\mathbf{x}_s | \mathbf{w}_s) q(z_s | \mathbf{w}_s)} [\log q(\mathbf{x}_s | z_s)]$  を最大化するパラメータを EM アルゴリズムの M ステップで得る:

$$q(\mathbf{x}_s | z_s) = \prod_{k=1}^K N(\mathbf{x}_s | \hat{\boldsymbol{\mu}}_{d,k}, \hat{\boldsymbol{\Sigma}}_{d,k})^{\delta(z_s=k)} \quad (8)$$

$$\hat{\boldsymbol{\mu}}_{d,k} = \frac{\sum_{s=1}^{S_d} \hat{\theta}_{s,k} \mathbf{E}_{q(\mathbf{x}_s | \mathbf{w}_s)} [\mathbf{x}_s]}{\sum_{s=1}^{S_d} \hat{\theta}_{s,k}} = \frac{\sum_{s=1}^{S_d} \hat{\theta}_{s,k} \hat{\boldsymbol{\mu}}_s}{\sum_{s=1}^{S_d} \hat{\theta}_{s,k}} \quad (9)$$

$$\hat{\boldsymbol{\Sigma}}_{d,k} = \frac{\sum_{s=1}^{S_d} \hat{\theta}_{s,k} \mathbf{E}_{q(\mathbf{x}_s | \mathbf{w}_s)} [(\mathbf{x}_s - \hat{\boldsymbol{\mu}}_{d,k})(\mathbf{x}_s - \hat{\boldsymbol{\mu}}_{d,k})^T]}{\sum_{s=1}^{S_d} \hat{\theta}_{s,k}} \\ = \frac{\sum_{s=1}^{S_d} \hat{\theta}_{s,k} \{ \hat{\boldsymbol{\Sigma}}_s + (\hat{\boldsymbol{\mu}}_s - \hat{\boldsymbol{\mu}}_{d,k})(\hat{\boldsymbol{\mu}}_s - \hat{\boldsymbol{\mu}}_{d,k})^T \}}{\sum_{s=1}^{S_d} \hat{\theta}_{s,k}} \quad (10)$$

これらトピック文の事後分布の平均から、各文書毎にトピック文を生成する:  $\hat{\mathbf{w}}_{d,k} \sim \text{RNN}(\hat{\boldsymbol{\mu}}_{d,k})$ 。既存研究 [4] と同様に、潜在分布の平均を用いることで、各トピックの要約文が得られることを期待する。

表1 評価データセットにおける各モデルの ROUGE (F1) 値 (%)。

データセット	Yelp Dataset Challenge			Amazon Product Reviews		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
Multi-Lead-1	27.42	3.74	14.34	<u>30.32</u>	<b>5.85</b>	15.96
LexRank [13]	26.53	3.30	14.54	<u>31.42</u>	<u>5.17</u>	16.67
Opinosis [14]	25.80	2.92	14.57	28.90	4.11	16.33
MeanSum [3]	<u>28.66</u>	3.73	15.77	30.16	4.51	17.76
Copycat [4]	<u>28.95</u>	<u>4.80</u>	<b>17.76</b>	<b>31.84</b>	<u>5.79</u>	<b>20.00</b>
提案法	<b>29.99</b>	<b>5.03</b>	<u>17.39</u>	<u>31.31</u>	<u>5.64</u>	18.03

## 2.3 Gaussian Word Embedding との関連

こうして得られたトピック文が、上位トピックほどより一般的な内容になることを Gaussian Word Embedding との関連を交えて説明する。親子関係にあるトピック文の事後分布間の KL ダイバージェンス (KLD) は、式 (7) に示した変分下限の  $x_s$  に関する項によって上から抑えられる (A.2 節にて証明)。

$$\begin{aligned}
 & \sum_{s=1}^{S_d} \hat{\theta}_{s,k} \text{D}_{\text{KL}}[q(x_s | w_s) | p(x_s | z_s = k)] \\
 & \geq \sum_{s=1}^{S_d} \hat{\theta}_{s,k} \text{D}_{\text{KL}}[q(x_s | z_s = k) | p(x_s | z_s = k)] \quad (11) \\
 & = \sum_{s=1}^{S_d} \hat{\theta}_{s,k} \text{D}_{\text{KL}}[q(x_s | z_s = k) | q(x_s | z_s = \text{par}(k))]
 \end{aligned}$$

一方、Gaussian word embedding では、単語のタクソノミー上において、親と子の潜在分布間の KLD を最小化することで、「動物」といった一般的な単語の分散は大きく、「犬」や「猫」などの具体的な単語の分散は小さくなることが報告されている [11]。これは「動物」は「犬」を指すことも「猫」を指すこともあり、意味の分散が大きいと説明できる。本研究も同様に、変分下限の最大化により、親子の潜在分布間の KLD の上界が最小化され、親トピック文の分散は大きく、子トピック文の分散が小さくなる。これも例えば、「I love this restaurant.” など全体的な内容に関する文は、「料理」を指すことも「サービス」を指すこともあり、意味の分散が大きくなると説明できる。したがって、上位トピックほど潜在分布の分散が大きく、より一般的な文が生成される。

## 2.4 トピック文の抽出

最後に、生成されたトピック文から要約に適切な文を抽出する。既存研究 [3] では、要約候補と入力文書間の ROUGE 値は、参照要約との ROUGE 値と高い相関があり、参照要約が存在しない状況下での要約候補評価に有用なことが報告されている。この知見をもとに、入力文書との ROUGE-1 (F1) を最大化するトピック文集合をビーム探索で選択する。

The food is good.	全体
It's a great place to go and have a great experience.	場所
If you are in the area, you need to try this place.	
I would recommend this place to anyone looking for a good pizza joint.	
The food is delicious and the service is top notch.	サービス
I've had a few of the burgers, and the pulled pork sandwich.	料理

参照要約: The food here is fantastic, easily the best sub sandwiches in the Arizona area. The shop is local and family run, so I definitely choose it over a lot of the large national chains that are all around town. The staff are extremely friendly and will always go above and beyond in creating a delicious sandwich for you. You will not be let down by the great food that they make here!

図3 レストランレビューから生成された要約例。

## 3 実験

### 3.1 実験設定

本実験では、Yelp と Amazon のレビューとそれを人手で要約した評価用データ [3, 4] を用いて、複数文書要約生成タスクにおける性能評価を行った。ベースラインには、各文書の先頭の文を抽出する Multi-Lead-1 に加え、教師なし抽出型手法である LexRank [13] と Opinosis [14]、及び教師なし生成型手法の MeanSum [3] と Copycat [4] を用いた。エンコーダ・デコーダには GRU-RNN を用い、木構造は 3 階層で固定され、それぞれの親トピックは 4 つの子トピックを持つ。詳細な設定は A.3 節にて補足する。

### 3.2 要約生成の性能評価

表 1 に、評価データにおける ROUGE 値を示す。太字はモデルの内最大値を、下線は最大値との差が approximate randomization test において統計的に有意とは言えない ( $p < 0.05$ ) ことを示す。いずれのデータセットにおいても、ほぼ全ての評価指標において、提案法は MeanSum を上回り、最新の教師なし要約モデルである Copycat に対しては競合する性能が得られた。図 1 に示したレビューについて、実際に生成された要約を図 3 に示す。参照要約は「料理」、「サービス」、「場所」について述べており、実

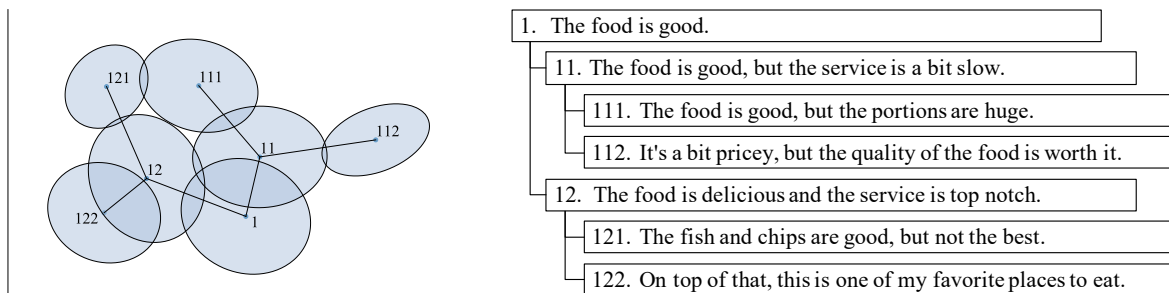


図4 主成分分析により2次元空間において表したトピック文の潜在分布。各点は潜在分布の平均を示しており、それぞれの円は平均からのマハラノビス距離が等しい座標群を示す。

表2 トピック文の平均類似度（高いほど類似）。

親子間類似度	Yelp	Amazon
親子ペア	1.83	1.73
非親子ペア	1.25	1.24

表3 トピック文の詳細度合い（高いほど詳細）。

階層毎の詳細度合い	Yelp	Amazon
第1階層	1.37	1.29
第2階層	1.67	1.45
第3階層	1.66	1.50

表4 各階層における分散共分散行列の行列式の対数。

階層毎の $\log  \hat{\Sigma}_{d,k} $ の平均値	Yelp	Amazon
第1階層	21.5	19.0
第2階層	19.6	18.1
第3階層	15.6	13.3

際に得られた要約においても同様のトピックについて述べていることが分かる。また、参照要約の最後の文は、結論として料理の美味しさに関して述べているが、提案法も同様に根トピックから料理の美味しさに関する内容のトピック文を生成している。

### 3.3 トピック文と木構造の定量分析

得られたトピック文と木構造について、以下2つの観点から、定量的な分析を行った。

**親子間類似度** 親子関係にあるトピック文は類似するトピックについて述べていることが望ましい。この特性を確認するために、親子ペアと非親子ペアのトピック文の類似度を、semantic textual similarity タスク [15] で最高精度を達成している ALBERT [16] を用いて評価した。表2に示すように、親子関係にあるトピック文は、親子でないトピック文のペアより類似していることが確認された。

**階層毎の詳細度合い** 提案法では、階層が深くなるにつれ、トピック文の内容がより具体的になることを期待している。この特性を検証するために、文の詳細度合い推定タスク [17] にて ALBERT を学習し、各階層のトピック文の詳細度合いを計測した。表3に示すように、階層が深いほど詳細度合いは高く、上位トピックからは一般的な、下位トピックからは詳細な内容の文が生成される傾向を確認した。

### 3.4 トピック文の潜在分布の分析

図4に、あるレストランレビューに関するトピック文の潜在分布について、その主成分ベクトル空間を可視化した。仮説通り子トピック文の潜在分布は親のトピック文の分布に近い位置にあり、親子関係にあるトピック文が類似したトピックを持つものと考えられる。また表4に示すように、各階層における分散共分散行列の行列式の対数は、階層が深くなるにつれ小さくなることが確認された。したがって、葉に近づくにつれ潜在分布の分散が小さくなり、具体的な文が生成されることがわかる。

## 4 おわりに

本研究では、文書から木構造上のトピックを推定し、トピック毎に要約文を生成することで、意見文書の要約が教師なしに得られることを示した。評価実験において、提案法の性能は最新の教師なし生成型要約手法と競合することを確認した。また、根の文の潜在分布は分散が大きく一般的な文が生成される一方、葉に近づくにつれ分散が小さくなり具体的な文が生成されるといった特性を確認した。これは単語の潜在表現にガウス分布を用いた Gaussian word embedding にて報告された特性と類似しており、要約のみならず、質問応答や対話生成などの文の詳細度合いを考慮する他タスクにも有用な知見である。

**謝辞** 本研究は、JST ACT-X JPMJAX1904、JST CREST JPMJCR1513 及び JSPS 特別研究員奨励費 20J10726 の支援を受けたものである。

## 参考文献

- [1]Giuseppe Carenini, Jackie Chi Kit Cheung, and Adam Pauls. Multi-document summarization of evaluative text. *Computational Intelligence*, 29(4):545–576, 2013.
- [2]Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Bitá Nejat. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1602–1613, 2014.
- [3]Eric Chu and Peter Liu. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1223–1232, 2019.
- [4]Arthur Braźinskas, Mirella Lapata, and Ivan Titov. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, 2020.
- [5]Reinald Kim Amplayo and Mirella Lapata. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, 2020.
- [6]Ivan Titov and Ryan McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 308–316, 2008.
- [7]Stefanos Angelidis and Mirella Lapata. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, 2018.
- [8]Asli Celikyilmaz and Dilek Hakkani-Tur. A hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 815–824, 2010.
- [9]Asli Celikyilmaz and Dilek Hakkani-Tur. Discovery of topically coherent sentences for extractive summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 491–499, 2011.
- [10]Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 166–177, 2019.
- [11]Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.
- [12]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*, pages 800–806, 2020.
- [13]Günes Erkan and Dragomir R Radev. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 365–371, 2004.
- [14]Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 340–348, 2010.
- [15]Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 1–14, 2017.
- [16]Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.
- [17]Annie Louis and Ani Nenkova. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 605–613, 2011.
- [18]John Paisley, Chong Wang, David M Blei, and Michael I Jordan. Nested hierarchical dirichlet processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):256–270, 2014.
- [19]David Alvarez-Melis and Tommi S Jaakkola. Tree-structured decoding with doubly-recurrent neural networks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- [20]Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, arXiv:1412.6980v9, 2014.

## A 付録

### A.1 木構造上のトピック分布の推定

本研究では、勾配降下法により学習可能な木構造ニューラルトピックモデル (TSNTM [12]) を用いることで、文の木構造上のトピック分布を推定する。

前置きとして、木構造トピックモデルの一つである hierarchical Dirichlet process [18] は、各文に関する木構造上のパス分布  $\pi_s$  と階層分布  $\phi_s$  を下記式でモデル化する。

$$v_{s,k} \sim \text{Beta}(1, \gamma), \pi_{s,k} = \pi_{s,par(k)} v_{s,k} \prod_{j \in \text{Sib}(k)} (1 - v_{s,j}) \quad (12)$$

$$\eta_{s,k} \sim \text{Beta}(\alpha, \beta), \phi_{s,k} = \eta_{s,k} \prod_{j \in \text{Anc}(k)} (1 - \eta_{s,j}) \quad (13)$$

$$\theta_{s,k} = \pi_{s,k} \cdot \phi_{s,k} \quad (14)$$

ここで、 $\text{Sib}(k)$  及び  $\text{Anc}(k)$  はそれぞれトピック  $k$  の前の兄弟のトピック集合及び祖先のトピック集合を指す。 $\pi_{s,k}$  は文  $s$  が根トピックからトピック  $k$  を通るパスを選択する確率を示す。一方、 $\phi_{s,k}$  は当該パス上で文  $s$  がトピック  $k$  の祖先のトピック  $j \in \text{Anc}(k)$  を選択せず、トピック  $k$  を選択する確率を示す。文  $s$  がトピック  $k$  を選択する確率  $\theta_{s,k}$  はこれら2つの確率の積で表される。

一方、TSNTM は、doubly-recurrent neural networks (DRNN [19]) を用いて文の潜在表現  $\mathbf{y}_s = \text{RNN}(\mathbf{w}_s)$  をパス分布  $\pi_s$  及び階層分布  $\phi_s$  に変換する。DRNN は親子と兄弟間の2つの RNN デコーダで構成され、トピック  $k$  の中間層は式 (15) で表される。この中間層から得られた棒折比率  $v_s$  で式 12 を置き換え、パス分布を得る。

$$\mathbf{h}_k = \tanh(\mathbf{W}_p \mathbf{h}_{par(k)} + \mathbf{W}_s \mathbf{h}_{k-1}) \quad (15)$$

$$v_{s,k} = \text{sigmoid}(\mathbf{h}_k^\top \mathbf{y}_s) \quad (16)$$

ただし  $\mathbf{h}_{par(k)}$  及び  $\mathbf{h}_{k-1}$  はそれぞれトピック  $k$  の親とその前の兄弟である。同様に、階層分布もまた DRNN により棒折比率  $\eta_s$  を算出して得られる。

### A.2 式 (11) の導出

**命題**  $q(\mathbf{x}_s | z_s)$  が式 (8)、(9)、(10) により与えられる時、式 (17) が成り立つ:

$$\sum_s \hat{\theta}_{s,k} \text{D}_{\text{KL}}[q(\mathbf{x}_s | \mathbf{w}_s) | p(\mathbf{x}_s | z_s = k)] - \sum_s \hat{\theta}_{s,k} \text{D}_{\text{KL}}[q(\mathbf{x}_s | z_s = k) | p(\mathbf{x}_s | z_s = k)] \geq 0 \quad (17)$$

**証明** 式 (17) の第1項は以下のように展開できる。

$$\begin{aligned} & \sum_s \hat{\theta}_{s,k} \text{D}_{\text{KL}}[q(\mathbf{x}_s | \mathbf{w}_s) | p(\mathbf{x}_s | z_s = k)] \\ &= \sum_s \hat{\theta}_{s,k} \text{D}_{\text{KL}}[N(\hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\Sigma}}_s) | N(\hat{\boldsymbol{\mu}}_{d,par(k)}, \hat{\boldsymbol{\Sigma}}_{d,par(k)})] \\ &= \frac{1}{2} \sum_s \hat{\theta}_{s,k} \left\{ \log |\hat{\boldsymbol{\Sigma}}_{d,par(k)}| - \log |\hat{\boldsymbol{\Sigma}}_s| + \text{Tr}[\hat{\boldsymbol{\Sigma}}_{d,par(k)}^{-1} \hat{\boldsymbol{\Sigma}}_s] \right. \\ & \quad \left. + (\hat{\boldsymbol{\mu}}_s - \hat{\boldsymbol{\mu}}_{d,par(k)})^\top \hat{\boldsymbol{\Sigma}}_{d,par(k)}^{-1} (\hat{\boldsymbol{\mu}}_s - \hat{\boldsymbol{\mu}}_{d,par(k)}) - d \right\} \\ &= \frac{1}{2} \sum_s \hat{\theta}_{s,k} \left\{ C_{d,par(k)} - \log |\hat{\boldsymbol{\Sigma}}_s| + \text{Tr}[\hat{\boldsymbol{\Sigma}}_{d,par(k)}^{-1} \hat{\boldsymbol{\Sigma}}_s] \right. \\ & \quad \left. + \text{Tr}[\hat{\boldsymbol{\Sigma}}_{d,par(k)}^{-1} \hat{\boldsymbol{\mu}}_s \hat{\boldsymbol{\mu}}_s^\top] - 2 \hat{\boldsymbol{\mu}}_{d,par(k)}^\top \hat{\boldsymbol{\Sigma}}_{d,par(k)}^{-1} \hat{\boldsymbol{\mu}}_{d,k} \right\} \end{aligned} \quad (18)$$

ただし、式 (9) から  $\sum_s \hat{\theta}_{s,k} \hat{\boldsymbol{\mu}}_{d,k} = \sum_s \hat{\theta}_{s,k} \hat{\boldsymbol{\mu}}_s$  を用いた。

第2項についても同様に、以下のように展開できる。

$$\begin{aligned} & \sum_s \hat{\theta}_{s,k} \text{D}_{\text{KL}}[q(\mathbf{x}_s | z_s = k) | p(\mathbf{x}_s | z_s = k)] \\ &= \frac{1}{2} \sum_s \hat{\theta}_{s,k} \left\{ C_{d,par(k)} - \log |\hat{\boldsymbol{\Sigma}}_{d,k}| + \text{Tr}[\hat{\boldsymbol{\Sigma}}_{d,par(k)}^{-1} \hat{\boldsymbol{\Sigma}}_{d,k}] \right. \\ & \quad \left. + \text{Tr}[\hat{\boldsymbol{\Sigma}}_{d,par(k)}^{-1} \hat{\boldsymbol{\mu}}_{d,k} \hat{\boldsymbol{\mu}}_{d,k}^\top] - 2 \hat{\boldsymbol{\mu}}_{d,par(k)}^\top \hat{\boldsymbol{\Sigma}}_{d,par(k)}^{-1} \hat{\boldsymbol{\mu}}_{d,k} \right\} \end{aligned} \quad (19)$$

表5 データセット中のレビュー文書/要約の件数。

データセット	Yelp	Amazon
学習	1,012,280/-	4,566,519/-
検証	800/100	672/84
評価	800/100	768/96

したがって、式 (17) は以下のように整理される。

$$\begin{aligned} & \sum_s \hat{\theta}_{s,k} \text{D}_{\text{KL}}[q(\mathbf{x}_s | \mathbf{w}_s) | p(\mathbf{x}_s | z_s = k)] \\ & - \sum_s \hat{\theta}_{s,k} \text{D}_{\text{KL}}[q(\mathbf{x}_s | z_s = k) | p(\mathbf{x}_s | z_s = k)] \\ &= \frac{1}{2} \sum_s \hat{\theta}_{s,k} \left\{ -\log |\hat{\boldsymbol{\Sigma}}_s| + \log |\hat{\boldsymbol{\Sigma}}_{d,k}| \right. \\ & \quad \left. + \text{Tr}[\hat{\boldsymbol{\Sigma}}_{d,par(k)}^{-1} (\hat{\boldsymbol{\Sigma}}_s + \hat{\boldsymbol{\mu}}_s \hat{\boldsymbol{\mu}}_s^\top - \hat{\boldsymbol{\Sigma}}_{d,k} - \hat{\boldsymbol{\mu}}_{d,k} \hat{\boldsymbol{\mu}}_{d,k}^\top)] \right\} \\ &= \frac{1}{2} \sum_s \hat{\theta}_{s,k} \left\{ -\log |\hat{\boldsymbol{\Sigma}}_s| + \log |\hat{\boldsymbol{\Sigma}}_{d,k}| \right\} \end{aligned} \quad (20)$$

ただし、式 (10) より  $\sum_s \hat{\theta}_{s,k} \{\hat{\boldsymbol{\Sigma}}_s + \hat{\boldsymbol{\mu}}_s \hat{\boldsymbol{\mu}}_s^\top\} = \sum_s \hat{\theta}_{s,k} \{\hat{\boldsymbol{\Sigma}}_{d,k} + \hat{\boldsymbol{\mu}}_{d,k} \hat{\boldsymbol{\mu}}_{d,k}^\top\}$  を用いた。すなわち、与式はエントロピーの重み付け和の大小比較に帰着される。ここで一般に、 $-\int q_1(\mathbf{x}) \log q_2(\mathbf{x}) d\mathbf{x} \geq -\int q_1(\mathbf{x}) \log q_1(\mathbf{x}) d\mathbf{x}$  が成立することから、式 (21) が成り立つ。

$$\begin{aligned} & \sum_s \hat{\theta}_{s,k} \left\{ -\int q(\mathbf{x}_s | \mathbf{w}_s) \log q(\mathbf{x}_s | z_s = k) d\mathbf{x}_s \right\} \\ & \geq \sum_s \hat{\theta}_{s,k} \left\{ -\int q(\mathbf{x}_s | \mathbf{w}_s) \log q(\mathbf{x}_s | \mathbf{w}_s) d\mathbf{x}_s \right\} \end{aligned} \quad (21)$$

右辺はガウス分布のエントロピーの重み付け和であることから、以下のように書ける。

$$\begin{aligned} & \sum_s \hat{\theta}_{s,k} \left\{ -\int q(\mathbf{x}_s | \mathbf{w}_s) \log q(\mathbf{x}_s | \mathbf{w}_s) d\mathbf{x}_s \right\} \\ &= \frac{1}{2} \sum_s \hat{\theta}_{s,k} \left\{ \log |\hat{\boldsymbol{\Sigma}}_s| + d \log 2\pi + d \right\} \end{aligned} \quad (22)$$

一方、左辺は以下のように展開できる。

$$\begin{aligned} & \sum_s \hat{\theta}_{s,k} \left\{ -\int q(\mathbf{x}_s | z_s = k) \log q(\mathbf{x}_s | \mathbf{w}_s) d\mathbf{x}_s \right\} \\ &= \frac{1}{2} \sum_s \hat{\theta}_{s,k} \left\{ \log |\hat{\boldsymbol{\Sigma}}_{d,k}| + d \log 2\pi \right. \\ & \quad \left. + \text{E}_{q(\mathbf{x}_s | \mathbf{w}_s)}[(\mathbf{x}_s - \hat{\boldsymbol{\mu}}_{d,k})^\top \hat{\boldsymbol{\Sigma}}_{d,k}^{-1} (\mathbf{x}_s - \hat{\boldsymbol{\mu}}_{d,k})] \right\} \end{aligned} \quad (23)$$

よって、式 (23) の最後の項は以下のように整理できる。

$$\begin{aligned} & \sum_s \hat{\theta}_{s,k} \left\{ \text{E}_{q(\mathbf{x}_s | \mathbf{w}_s)}[(\mathbf{x}_s - \hat{\boldsymbol{\mu}}_{d,k})^\top \hat{\boldsymbol{\Sigma}}_{d,k}^{-1} (\mathbf{x}_s - \hat{\boldsymbol{\mu}}_{d,k})] \right\} \\ &= \sum_s \hat{\theta}_{s,k} \left\{ \text{E}_{q(\mathbf{x}_s | \mathbf{w}_s)}[\text{Tr}(\hat{\boldsymbol{\Sigma}}_{d,k}^{-1} (\mathbf{x}_s - \hat{\boldsymbol{\mu}}_{d,k})(\mathbf{x}_s - \hat{\boldsymbol{\mu}}_{d,k})^\top)] \right\} \\ &= \text{Tr}[\hat{\boldsymbol{\Sigma}}_{d,k}^{-1} \sum_s \hat{\theta}_{s,k} \left\{ \text{E}_{q(\mathbf{x}_s | \mathbf{w}_s)}[(\mathbf{x}_s - \hat{\boldsymbol{\mu}}_{d,k})(\mathbf{x}_s - \hat{\boldsymbol{\mu}}_{d,k})^\top] \right\}] \\ &= \text{Tr}[\hat{\boldsymbol{\Sigma}}_{d,k}^{-1} \sum_s \hat{\theta}_{s,k} \hat{\boldsymbol{\Sigma}}_{d,k}] \\ &= \sum_s \hat{\theta}_{s,k} d \end{aligned} \quad (24)$$

したがって、式 (21)、(22)、(23)、(24) を整理すると、 $\sum_s \hat{\theta}_{s,k} \{\log |\hat{\boldsymbol{\Sigma}}_{d,k}|\} \geq \sum_s \hat{\theta}_{s,k} \{\log |\hat{\boldsymbol{\Sigma}}_s|\}$  が成り立ち、命題が成立することが示された。

### A.3 実験設定詳細

データセット中の文書及び要約の件数を表5に示す。参照要約は Amazon Mechanical Turk にて、各製品について8件のレビューの要約を依頼し作成された [3, 4]。ハイパーパラメータは検証データにおける性能をもとに決定され、単語の潜在表現、隠れ層及び文の潜在表現の次元数はそれぞれ200、400及び32である。勾配降下法は Adam [20] を用い、学習率は  $10^{-3}$  でバッチサイズは4である。