文書の分散表現を用いたトピック分析手法の提案

中山悠理 [1] 小林亮太 [1,2]
[1] 東京大学大学院 新領域創成科学研究科
[2] 東京大学 数理・情報教育研究センター
{3645588575,r-koba}@edu.k.u-tokyo.ac.jp

概要

トピック分析は，多数の文書から主要なテーマを抽出する技術であり，大規模なテキストデータの分析を可能にする．トピック分析を行う代表的手法としてトピックモデル (LDA など) があり，様々な分野に応用されてきた．一方で，この手法を文書の単語数が少ないソーシャルメディアデータなどに適用すると，人間が解釈しやすいテーマ (トピック) が得れないという問題がある．本研究では文書の分散表現に基づくトピック分析手法を提案し，2 つのデータセットを用いてトピック分析手法の性能を評価した．

1 はじめに

デジタル化が進み，ニュース，Web ページ，論文，書籍など様々な形のテキストデータを利用できるようになりつつある．テキストデータを整理，理解するための 1 つの方法は，人間が実際にテキストを読んで意味解釈を行うことである．しかし，オンライン上のテキストデータは爆発的に増えているため，全てのデータを人手で意味解釈することは不可能である．

テキストデータの例として，多数のニュース記事 (文書) の集合を考えよう．それぞれのニュース記事には，政治，経済，スポーツ，科学など記事のテーマがあるだろう．トピック分析の目的は，テキストデータから記事のテーマを自動的に発見し，ニュース記事をテーマが似たいくつかのグループに分類することである．トピック分析を行う代表的手法として，トピックモデル (Latent Dirichlet Allocation, LDA) [1] が挙げられる．LDA は，1) トピックはいくつかの単語の組み合わせによって決まる，2) 文書はいくつかのトピックが組み合わさったものである，という仮定のもとで，多量のテキストデータからトピック群を抽出する統計モデルである．トピックモデルは，テキストデータの分析で広く使われており，バイオインフォマティクス [2]，科学計量学 (scientometrics) [3]，政治学 [4] などに応用されている．

Twitter や Reddit などのソーシャルメディアから得られるテキストデータの分析は，東日本大震災 [5] [6] [7] や Covid-19 [8] などの災害，あるいは，選挙 [9] や Covid-19 ワクチン [10] [11] などの社会的課題に対する人々の認識を把握することへの一助となることが期待されている．その一方で，ソーシャルメディアデータの特徴として，1) 文書 (ツイートやコメント) の単語数が少ない，2) 誤字脱字などにより出現頻度の少ない単語が多数現れる，という 2 点が挙げられる．そのため，このようなデータに LDA を適用すると，人間にとって解釈困難なトピックが得られてしまう．前者の問題 (文書の短さ) に対応するため，Author Topic Model [12] や TwitterLDA [13] など LDA を拡張したモデルが提案されている．しかしながら，ソーシャルメディアデータからトピック抽出を行うことは依然として困難な状況にある．

本研究では，文書の分散表現 (文書の埋め込み) に基づくトピック分析手法を提案する．そして，20 Newsgroup と AG's corpus の 2 つのデータセットを用いて提案手法の性能評価を行い，既存手法である LDA と比較を行った．

2 提案手法

本研究では，多量の文書を K 個のトピックに分類する手法を提案する．提案手法は，

- 文書をベクトルに埋め込む（3.1 節）．
- 得られた埋め込みベクトルを混合ガウスモデルを用いてクラスタリングを行う（3.2 節）．

の 2 段階に基づく．以降では，上の 1)，2) について

説明する．

2.1 埋め込みベクトルの計算

この節では，文書 (単語列) から埋め込みベクトルを計算する方法を説明する．自然言語処理で広く用いられている単語埋め込みは文書中の個々の単語をベクトルに変換することであるが，文書埋め込みは文書全体を 1 つのベクトルに変換することである．これにより，単語数が異なる文書群をベクトル群として表現できる．

本研究では 2 つの機械学習モデル，doc2vec [14] と Sentence-BERT (SBERT) [15] を用いた．doc2vec は，単語の埋め込みベクトルを計算する word2vec [16] から着想を得て開発されたものである．doc2vec では埋め込みベクトルの次元 D を変えることができる．今回は $D = 2, 4, \cdots, 160$ を検討した．本研究では，Python ライブラリ gensim[17] を使用して doc2vec の埋め込みベクトルを計算した．

SBERT [15] は NLI データセットから事前学習された機械学習モデルである．SBERT は，文書埋め込みベクトルを得るため，BERT [18] に pooling 処理を追加したものである．SBERT の出力ベクトルの次元は変更できないため，固定値 $D = 384$ を用いた．本研究では，hugging face を通して公開されている all-MiniLM-L6-v2 [1] を使用した．

前処理として，文書ベクトル $\mathbf{v}_i$ から平均ベクトルを引いて中心化された文書ベクトル $\bar{\mathbf{v}}_i$ を計算した．

$$\bar{\mathbf{v}}_i = \mathbf{v}_i - \mathbf{m}_{\mathbf{v}},$$

ただし，$\mathbf{m}_{\mathbf{v}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}_i$ は平均ベクトル，n はデータセットの文書数を表す．必要に応じて，中心化されたベクトル $\bar{\mathbf{v}}_i$ のノルムが 1 になるように正規化を行なった．

2.2 混合ガウス分布によるクラスタリング

次に，埋め込みベクトルを混合ガウスモデルを用いてクラスタリングした．埋め込みベクトルの確率分布 $p(\mathbf{x})$ を以下の式で表される混合ガウス分布でフィットした:

$$p(\mathbf{x} \mid \mu, \mathbf{\Sigma}, \pi) = \sum_{i=1}^{K} \pi_i \mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\right)$$

ただし，K はトピック数，$\mathcal{N}\left(\mathbf{x} \mid \boldsymbol{\mu}_i, \mathbf{\Sigma}_i\right)$ は D 次元ガウス分布 (平均ベクトル: μ_i, 共分散行列: $\mathbf{\Sigma}_i$)，π_i は混合比を表す．

1） https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

本研究では，共分散行列 Σ_i として，以下の 3 つの場合で比較を行なった:

- full: Σ_i は正定値行列 (制約なし).
- diag: Σ_i は正定値かつ対角行列.
- scalar: Σ_i は正定値かつ単位行列の定数倍.

パラメータ $\mu_i, \mathbf{\Sigma}_i, \pi_i$ は逐次的に計算することによって推定できる [19]. 混合ガウス分布のパラメータ推定は，python パッケージ scikit-learn[20] を用いて行なった．

3 実験

3.1 データセット

本研究では，20 Newsgroups(20News) [21]， AG's corpus [22] の 2 つのデータセットを用いた．20News はニュース記事のデータが 20 種類のテーマ (医学, 銃問題, 中東の政治など) に分類されたものである．このデータセットには，自然言語でない長文が含まれていたため，単語数が上位 0.01% の文書を除いて分析した．

AG's corpus は 12 万件のニュース記事を集めたものであり，ニュース記事の本文とタイトルが 4 つのテーマ (世界，科学/技術，スポーツ，ビジネス) に分類されている．データセットを，ニュース記事 (AgNews) とタイトル (AgTitle) に分けて分析を行なった．表 1 は，データセットの平均単語数，文書数を示す．

表 1 本研究で用いたデータセット

	平均単語数	文書数
20News	344.7	18770
AgNews	35.9	120000
AgTitle	8.1	120000

3.2 評価指標

本研究では，抽出されたトピックの性能を評価するための指標として，補正された相互情報量 (AMI, Adjusted Mutual Information) [23] と Coherence スコア C_V [24] を用いた．AMI は分類の類似度についての指標である．今回は人手でつけられた文書の分類結果と，トピック抽出モデルで得られた分類結果の類似度を AMI で評価し，AMI をトピックの分類精度であると考える．AMI は 1 以下の値を取る指標である．2 つの分類結果が完全に一致するときのみ AMI

は 1 をとり，2 つの分類結果の類似度はランダムな場合と同程度の場合には AMI は 0 をとる．AMI は Python ライブラリ scikit-learn[20] を用いて計算した．

Coherence score C_V はトピックの解釈性についての指標であり，トピック内の文書の単語分布により計算される．C_V は人間による解釈結果と相関があり，人にとって解釈しやすいトピックでは高い値を取る傾向にある．C_V は Python ライブラリ gensim [17] を用いて計算した．

3.3 埋め込み方法の検討

まず，埋め込み次元を変更できる doc2vec について，次元が分類精度 (AMI) に与える影響を調べた．混合ガウスモデルの共分散行列を scalar として，トピック数 K はそれぞれのデータセットと同じ数 (20News: 20, AgNews: 4) にしてトピック抽出を行なった．図 1 は，2 つのデータセット (A. 20news, B. AgNews) から分類精度 を計算した結果である．その結果，分類精度は埋め込みの次元に大きく影響を受け，20News では 80 次元，AgNews では 8 次元，とすると精度が最も高くなった．また，埋め込み次元が低次元の場合を除き，正規化 (図 1: オレンジ) は分類精度を向上させた．

次に，SBERT を用いて文書埋め込みを計算し，正規化が分類精度 (AMI) に与える影響を調べた (表 2)．doc2vec の時ほど大きな効果はないものの，正規化は SBERT を用いた場合でも分類精度を向上させることがわかった．また，最適な埋め込み次元を用いた doc2vec と SBERT で分類精度 (AMI) を比較した (表 3)．いずれのデータセットに対しても，SBERT の方が高い分類精度を達成した．

以降の解析では，データセットに応じて埋め込み次元を調整する必要がなく，高い分類精度を達成した SBERT (正規化あり) を用いて文書埋め込みを行なった．

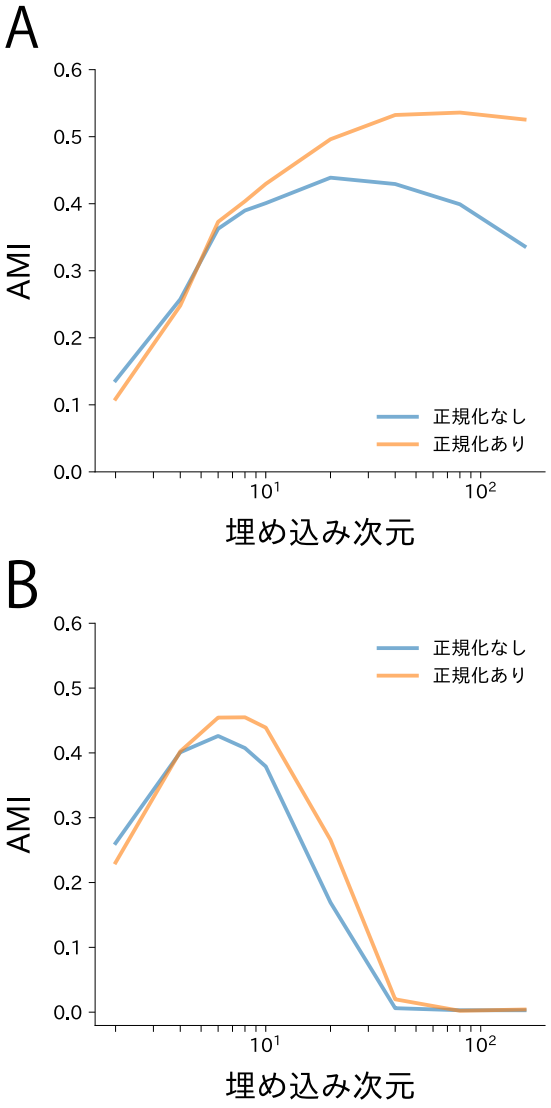

図 1 埋め込み方法 (埋め込み次元と正規化) が分類精度 (AMI) に与える影響 (A. 20news, B. AgNews)．埋め込みモデルは doc2vec を用いた．

表 2 正規化が分類精度 (AMI) に与える影響．埋め込みモデルは SBERT を用いた．

	正規化なし	正規化あり
20News	0.576	0.582
AgNews	0.579	0.582

表 3 埋め込み用いる機械学習モデルの違いによる分類精度 (AMI) の比較．太字は最も精度の高い機械学習モデルを示す．

	doc2vec	SBERT
20News	0.536	**0.582**
AgNews	0.455	**0.582**

3.4 混合ガウスモデルの共分散行列の検討

クラスタリングアルゴリズム (2.2 節) の違いが分類精度 (AMI) に与える影響を調べるため，3 種類の共分散行列 (full，diag，scalar) を用いた混合ガウスモデルでトピック抽出を行い，分類精度を比較した (表 4)．この結果，共分散行列に制限を加えない (full) と，単位行列 (scalar) に比べて分類精度がわずかに向上することがわかった．以降では，共分散行列には制限を加えずに (full) 分析を行なった．

表 4 共分散行列が分類精度 (AMI) に与える影響．太字は最も精度の高い共分散行列を示す．

	scalar	diag	full
20News	0.582	0.590	**0.593**
AgNews	0.582	0.586	**0.608**

3.5 実験結果

最後に，提案手法と既存手法である LDA を用いて，文書群のトピック分析を行い，性能比較を行なった．提案手法では，機械学習モデル SBERT を用いて埋め込みを行い，正規化をした上で，制約なし (full) の共分散行列を持つ混合ガウスモデルを使ってクラスタリングを行なった．また，LDA を用いた実験では，前処理として”I”，”is”，”a”など，文書の内容にかかわらず出現する単語（stop words) を除去し，LDA の分析には Python ライブラリ gensim [17] を用いた．

表 5 は，20News, AgNews の 2 つのデータセットについて，分類精度についての指標である AMI とトピックの解釈性についての指標である C_V を比較したものである．提案手法は，AMI，C_V のいずれの指標についても，既存手法 (LDA) に比べて高い性能を達成した．この結果は，提案手法のトピック分析結果は，既存手法に比べて，人間が分類した結果により近く，トピックもより解釈しやすいものになっていることを示唆している．

次に，ツイートのような文書の単語数が少ない場合の性能を評価するため，AG's corpus のニュースタイトルを文書としたデータセット (AgTitle) についてトピック分析を行なった．ニュース記事のデータセットと同様に，提案手法は，AMI，C_V のいずれの指標についても，既存手法 (LDA) に比べて高い性能を達成した．既存手法 (LDA) の分類精度 (AMI) はチャンスレベルと同程度になっている．この結果は，LDA はトピック分類ができていないことを示唆している．その一方で，提案手法は，AgNews (ニュース記事データ) に比べると性能が低下するものの，ニュースタイトルだけを分析した場合にも AMI: 0.50，C_V: 0.84 と高い性能を示した．この性能は，AgNews の LDA よりも高い性能であった．

表 5 提案手法 (SBERT) と既存手法 (LDA) の性能比較．太字は精度が高い手法を示す．

		AMI	C_V
20News	LDA	0.332	0.538
	提案手法	**0.593**	**0.592**
AgNews	LDA	0.408	0.806
	提案手法	**0.608**	**0.957**
AgTitle	LDA	0.023	0.705
	提案手法	**0.501**	**0.837**

4 おわりに

本研究では，テキストデータからトピックを抽出するタスクにおいて，文書埋め込みと混合ガウスモデルを組み合わせた手法を提案した．そして，人手によってトピックが付与された 20 News，AG's corpus の 2 つのデータセットを用いて，トピック分析手法の性能評価を行なった．その結果，提案手法は，既存手法である LDA に比べて高い性能であること，つまり，より人間に近いトピック分類を行い (AMI が高い)，トピックが解釈しやすい (Topic Coherence C_V が高い) ことが示された．さらに，提案手法は短い文書 (ニュースタイトルのみ) に対しても，高性能なトピックを抽出できることを示唆する結果が得られた．

謝辞

本研究は，JSPS 科研費 JP18K11560, JP19H01133, JP21H03559, JP21H04571, JP22H03695, AMED JP21wm0525004，JST さきがけ JPMJPR1925 の支援を受けたものである．

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] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. An overview of topic modeling and its current ap-

plications in bioinformatics. **SpringerPlus**, Vol. 5, No. 1, p. 1608.

[3] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. **Proceedings of the National Academy of Sciences**, Vol. 101, No. suppl_1, pp. 5228–5235, 2004.

[4] Justin Grimmer and Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. **Political Analysis**, Vol. 21, No. 3, pp. 267–297.

[5] Son Doan, Bao-Khanh Ho Vo, and Nigel Collier. An Analysis of Twitter Messages in the 2011 Tohoku Earthquake. In Patty Kostkova, Martin Szomszor, and David Fowler, editors, **Electronic Healthcare**, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 58–66. Springer.

[6] Misako Takayasu, Kazuya Sato, Yukie Sano, Kenta Yamada, Wataru Miura, and Hideki Takayasu. Rumor Diffusion and Convergence during the 3.11 Earthquake: A Twitter Case Study. **PLOS ONE**, Vol. 10, No. 4, p. e0121443.

[7] Takako Hashimoto, David Lawrence Shepard, Tetsuji Kuboyama, Kilho Shin, Ryota Kobayashi, and Takeaki Uno. Analyzing temporal patterns of topic diversity using graph clustering. **The Journal of Supercomputing**, Vol. 77, No. 5, pp. 4375–4388.

[8] Shu-Feng Tsao, Helen Chen, Therese Tisseverasinghe, Yang Yang, Lianghua Li, and Zahid A. Butt. What social media told us in the time of COVID-19: A scoping review. **The Lancet Digital Health**, Vol. 3, No. 3, pp. e175–e194.

[9] Hao Wang, Doğan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In **Proceedings of the ACL 2012 System Demonstrations**, pp. 115–120.

[10] Ryota Kobayashi, Yuka Takedomi, Yuri Nakayama, Towa Suda, Takeaki Uno, Takako Hashimoto, Masashi Toyoda, Naoki Yoshinaga, Masaru Kitsuregawa, and Luis E. C. Rocha. Evolution of public opinion on covid-19 vaccination in japan: Large-scale twitter data analysis. **Journal of Medical Internet Research**, Vol. 24, No. 12, p. e41928.

[11] Takako Hashimoto, Takeaki Uno, Yuka Takedomi, David Shepard, Masashi Toyoda, Naoki Yoshinaga, Masaru Kitsuregawa, and Ryota Kobayashi. Two-stage clustering method for discovering people's perceptions: A case study of the covid-19 vaccine from twitter. In **2021 IEEE International Conference on Big Data (Big Data)**, pp. 614–621, 2021.

[12] 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**, pp. 487–494. AUAI Press.

[13] Wayne ZHAO, Jing JIANG, Jianshu WENG, Jing HE, Ee Peng LIM, Hongfei YAN, and Xiaoming LI. TwitterLDA. **SMU Research Data**.

[14] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In **Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32**, p. II–1188–II–1196. JMLR.org, 2014.

[15] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing**. Association for Computational Linguistics, 11 2019.

[16] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In **1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings**, 2013.

[17] Radim Rehurek and Petr Sojka. Gensim–python framework for vector space modelling. **NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic**, Vol. 3, No. 2.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[19] Christopher M. Bishop. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Springer-Verlag, Berlin, Heidelberg, 2006.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, Vol. 12, pp. 2825–2830.

[21] Youngjoong Ko. A study of term weighting schemes using class information for text classification. In **Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 1029–1030. Association for Computing Machinery.

[22] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In **Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1**, p. 649–657, Cambridge, MA, USA, 2015. MIT Press.

[23] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In **Proceedings of the 26th Annual International Conference on Machine Learning**, pp. 1073–1080. Association for Computing Machinery.

[24] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In **Proceedings of the Eighth ACM International Conference on Web Search and Data Mining**, pp. 399–408. Association for Computing Machinery.