# 大規模言語モデルによるトピックモデルの補正後処理の検討

森 承宇[1, a)]
[1]株式会社日立製作所　研究開発グループ
[a)]sho.mori.jh@hitachi.com

## 概要

本研究は，生成 AI の利用制約下でも高品質なテキスト分類を可能にするため，Guided Topic Model（Guided TM）に対する LLM 後処理補正を提案する。提案手法では，Guided TM により生成されるトピックを分類ラベルと対応付け，ラベル数に応じてトピックを分割し語句を選別する補正を行う。組織が保有する情報の疑似データでの評価によると，Coherence や Diversity などの TM の評価指標の向上は見られなかった一方で，文書分類の評価指標である正解率は向上し，トピック内のラベル混在が解消されることが確認された。

## 1 背景

テキストデータの分類タスクは，昨今 chatGPT のような大規模言語モデル(LLM: large language model)に適切なプロンプトを与えることで，訓練データを追加で与えることなく優れた性能を示すことが期待できるが[1]，組織のガバナンス要件等により生成 AI のクラウドサービスが利用不可であることやタスク実行時の計算負荷が高いことによる制約のため，LLM を利用することが難しいユースケースがある。

軽量にテキストデータの分類を実現するモデルの一つとして，トピックモデル(TM: Topic Model)は，テキストデータの大規模な解析を可能にする手法として注目されている。TM は，テキストデータから潜在的なトピックを，トピック語句及びその重要度を表すスコア値の集合の形で表されるトピック情報を複数出力するタスクである。分類のために用いる分類ラベルをテキストデータに対して付与せずとも教師なしでモデルを構築できるため，ラベルコストなどの工数をかけることなく分類モデルを構築することができる。代表的な手法として，従来は潜在ディリクレ配分法(LDA: Latent Dirichlet Allocation) [2]に代表される統計的確率モデルを用いる手法があるが，近年では LLM のようにニューラルネットワークを用いた手法[3-7]が研究され，高い性能を示している。

このように優れた性能を示す TM 手法が開発されてきているが，テキスト分類の実運用において TM は以下の課題がある。

- 予め分類ラベルを想定している場合，分類ラベルと TM の出力であるトピック情報を適切に対応させること難しい。
- その結果として，類似度の高い分類ラベルが 2 つ以上混在するトピック情報や，類似度の高い分類ラベルが 1 つも含まれていないトピック情報が生成される。

前述の課題に対処する手法の一つとして，TM の入力に分類ラベルと直接関連する語の集合であるシード語句を追加で与えることで TM の方向付けを可能にする Guided TM という方法がある[8-9]が，その効果は限定的である。

本研究では，TM の出力であるトピック情報に対し，LLM による補正処理を後処理として付加することで，前述の課題を解決する手法を提案する。近年，LLM を用いた TM に関する研究は増加しているものの[3-7]，提案手法のように生成されたトピック情報を後処理として補正することに関する研究事例は少ない[10-11]。さらに，本手法は TM において生じる課題を解決するだけでなく，既存の TM 手法[2-9]と独立に適用可能である点も強みである。

## 2 提案手法

前記の課題が発生する理由として，モデル構築時及びモデル使用時に TM は分類ラベルそれ自体を入力しないため，出力されるトピック情報に対してトピック情報―分類ラベル間の類似度が 1 番高い分類ラベルと，2 番目に高い分類ラベルの類似度が近い場合があることが挙げられる。本研究で提案する LLM による補正処理である「トピック情報分割処理」は，1 つのトピック情報に 1 つの分類ラベルが対応するようにトピック情報の分割することで，こ

のような状況を改善することが期待される。提案処理を含む処理フローについて，図 1 に示す。

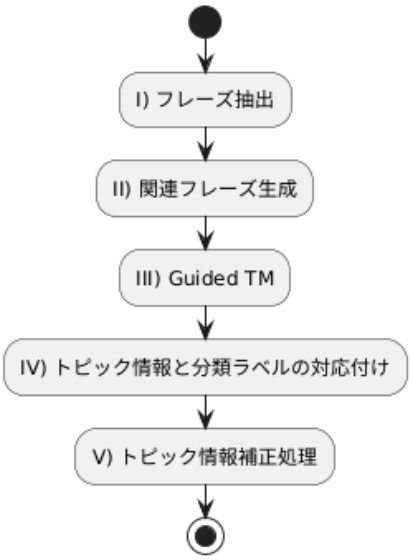


図 1 提案手法の処理

I) フレーズ抽出
フレーズ抽出を行うにあたり，まず文章に対して形態素解析を実施した。日本語の形態素解析には MeCab[12]を用いた。数字やメールアドレスなどについては正規表現を用いて解析対象から除いたうえで，形態素解析で抽出される名詞句，及び，名詞句の連なりを解析対象のフレーズとして抽出した。

II) 関連フレーズ生成
抽出した頻出フレーズと分類ラベルを入力として，LLM を用いてフレーズと分類ラベルの類似度を計算することで，各分類ラベルと関連性の高いフレーズ集合を得た。

III) Guided TM
II)で得たフレーズをシード語句とし，シード語句とテキストデータを入力として Guided TM を適用して，トピック情報を生成した。

IV) トピック情報と分類ラベルの対応付け
III)で生成された各トピック情報について各分類ラベルのシード語句に付与されているスコア値を合計することで，トピック情報―分類ラベル間の類似度を計算し，分類ラベル毎に最も類似度の高いトピック情報を対応付けた。

V) トピック情報補正処理
提案処理の概要を図 2 に示す。提案処理は，下記で構成される。

V-1) 類似度が所定の閾値よりも高く算出された分類ラベルを 2 以上含むトピック情報を 1 つ選択する。

V-2) 選択したトピック情報を，類似度の高い分類ラベルと同数だけ生成する（トピック語句とスコア値をまだ含まない）。

V-3) 選択したトピック情報の各トピック語句について，各分類ラベルとの類似度を算出し，所定の閾値よりも高いトピック語句のみ V-2)で複製したトピック情報に登録する。

V-4) 上記 V-1) ~ V-3)の処理を対象のトピック情報すべてを対象として繰り返す。

得られたトピック情報に基づき，各テキストデータに対して各トピック情報に当てはまる確率分布であるトピック分布を算出し，最も確率の高いトピック情報を文書分類結果とする。

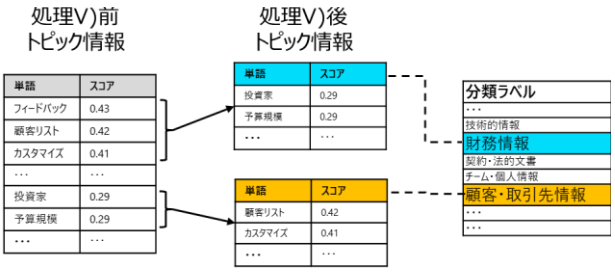


図 2 トピック情報分割処理(処理 V)の概要

## 3 検証方法

提案手法の評価方法について，データセット，比較条件，評価指標を示す。

データセットについては，日本語文書として組織が保有するデータの疑似データを評価に用いた。データセットの詳細については，付録 A に記す。

比較条件については，2.提案手法の IV) まで行った Guided TM モデル(以降，M1 と略記)と，提案処理 V)を適用したモデル(以降，M2)を評価に用いた。Guided TM は，BERTopic[7]を適用した。BERTopic を用いる場合は文章の埋め込み表現を作成するための LLM が必要なため，日本語の場合は sentence-bert-base-ja-mean-tokens-v2 [13] の Sentence-BERT モデルを用いた。なお，テキストデータの正規化，ストップワードの設定，形態素解析における品詞フィルタなどの前処理を含む他の条件は，諸条件で統一して行った。

評価指標については，TM の評価指標と，文書分類の評価指標を用いた。TM の評価指標については，トピック語句間の意味的一貫性や解釈のしやすさを定量化する指標である Coherence (Umass Coherence)，トピック情報間でトピック語句の重複がどれだけ少ないかを示す指標である Diversity を評価した。文書分類の評価指標としては，分類一致の正解率を評価した。

## 4 結果

表 1, 2 に提案手法を適用した場合(M1)，適用しなかった場合(M2)における，TM の評価結果，及び，文書分類の評価結果を示す。また，M2 の対象となったトピック情報の文書分類結果の詳細は表 3 に示す。表 3 の各列は分類ラベルの略記(付録 A 参照)であり，各行は生成されるトピック情報を識別するための記号を表す。なお，「トピック Z」はいずれのトピックにも分類されなかったことを表す。表 3 は，各分類ラベルが付されている文書が各々のトピック情報であると分類された数を表す。

表 1. TM の評価結果

| | M1 (Guided TM) | M2 (提案手法) |
|---|---|---|
| Coherence | 0.31 | 0.30 |
| Diversity | 0.4911 | 0.4787 |

表 2. 文書分類の評価結果

| | M1 (Guided TM) | M2 (提案手法) |
|---|---|---|
| 正解率 | 0.808 | 0.936 |

## 5. 結果の検討

TM の評価結果(表 1)によると，提案処理の適用有無による指標変化は全体的に小さいが，Diversity 値が低下する傾向が観察された。これは，複数の分類ラベルにまたがる語句が提案処理後も複数トピックに残る結果，複数のトピック情報に共通して記載される語句数が増加したためと考えられる。

文書分類の評価結果(表 2)によると，提案処理の適用することで正解率が 12.8%向上した。実際に表 3 によると，提案処理後はトピック A, B, H, J 内での分類ラベルの混在が解消され，分類の一意性が向上したことが確かめられた。

今後は，その他の公開データセット([14]など)で評価を行ったうえで，提案処理で選定されるトピック情報の候補数の調整や，類似度の高い分類ラベルが 1 つも含まれていないトピック情報の検出及びその発生を抑制するための手法の検討など，精度向上に向けた更なる改善が必要である。

表 3. M2 が適用されたトピック情報の文書分類結果の詳細

| M1 | M2 | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o | p | q | r | s | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Z | Z | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 20 | 4 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 3 |
| A | - | 0 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| - | A_b | 0 | 19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| - | A_i | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | - | 0 | 0 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 |
| - | B_e | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| - | B_g | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 |
| H | - | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 11 | 0 | 0 | 0 | 0 | 0 | 0 |
| - | H_m | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| - | H_n | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 |
| J | - | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 5 |
| - | J_s | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 1 |
| - | J_t | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |

## 参考文献


[1] OpenAI, “chatGPT”, https://chat.openai.com, accessed 2025-09-03.
[2] Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Volume 3, Pages 993-1022, 2003.
[3] Yishu Miao, Edward Grefenstette, Phil Blunsom, “Discovering Discrete Latent Topics with Neural Variational Inference”, In Proceedings of the 34th International Conference on Machine Learning, Volume 70, 2017.
[4] Akash Srivastava, Charles Sutton, “Autoencoding Variational Inference For Topic Models”, In International Conference on Learning Representations, 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”, Volume 8, Pages 439–453, 2020.
[6] Bianchi Federico, Terragni Silvia, Hoby Dirk, Nozza Debora, Fersini Elisabetta, “Crosslingual Contextualized Topic Models with Zero-shot Learning”, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Pages 1676–1683, 2021.
[7] Grootendorst, Maarten, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure”, arXiv preprint arXiv:2203.05794, 2022.
[8] “GuidedLDA: Guided Topic modeling with latent Dirichlet allocation”, https://github.com/vi3k6i5/guidedlda, accessed 2025-09-05.
[9] “Guided Topic Modeling”, https://maartengr.github.io/BERTopic/getting_started/guided/guided.html, accessed 2025-09-05.
[10] Tian Nie, Yi Ding, Chen Zhao, Youchao Lin, T. Utsuro, “A Method of Subtopic Classification of Search Engine Suggests by Integrating a Topic Model and Word Embeddings”, International Journal of Software Innovation, 2018.
[11] Arik Reuter, Bishnu Khadka, Anton Thielmann, Christoph Weisser, Sebastian Fischer, Benjamin Säfken, “GPTopic: Dynamic and Interactive Topic Representations”, arXiv preprint arXiv:2403.03628, 2025.
[12] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto, “Applying Conditional Random Fields to Japanese Morphological Analysis”, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pages 230-237, 2004.
[13] “sonoisa/sentence-bert-base-ja-mean-tokens-v2”, https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2, 2021, accessed 2025-09-03.
[14] “fetch_20newsgroups”, https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_20newsgroups.html, accessed 2025-12-11.


## A 付録

組織が保有するデータの疑似データとして，筆者があらかじめ定めた 20 件の分類ラベルに基づき，分類ラベルのいずれかについて記載された 300 字程度のテキスト 3300 件を用意した。TM を構築するための訓練データとして 3000 件，訓練データを除く残りのデータを文書分類の評価データとして用いた。

本文では，各分類ラベル名について，以下のように略記する。

a. プロジェクト計画・スケジュール
b. 技術的情報
c. 財務情報
d. 契約・法的文書
e. チーム・個人情報
f. 顧客・取引先情報
g. 市場戦略・営業情報
h. セキュリティ情報
i. 未公開情報
j. その他組織が保有する情報
k. プロジェクトの概要
l. 公開可能なスケジュール
m. 製品やサービスの公開情報
n. プロモーション活動情報
o. 公開パートナー情報
p. 公開財務情報
q. チームや組織に関する情報
r. 市場・業界情報
s. 成果物に関する情報
t. その他の一般情報