SNS 投稿文とハッシュタグの共通埋め込み空間の構築

仲田明良 [1]　狩野芳伸 [1]
[1] 静岡大学大学院総合科学技術研究科
{anakada,kano}@kanolab.net

概要

本研究は，SNS 投稿文とハッシュタグを共通の埋め込み空間に配置し，共通空間でのベクトル類似度に基づいて投稿文に適したタグを選択可能にする手法を提案する．ハッシュタグの埋め込みは，投稿内で観測されるハッシュタグの共起構造から学習し，投稿文の埋め込みは言語モデルにより得る．双方の埋め込みを対照学習により埋め込み空間上で対応付けし，「空間上で近い＝意味的に対応する」空間を構築する．X（旧 Twitter）から収集したユーザの投稿を用い，2 候補から正しいタグを選択するタスクで評価した結果，提案手法はベースラインを上回り精度 0.849 を得た．

1　はじめに

SNS では，膨大な投稿から特定トピックの情報へ素早く辿り着くためにハッシュタグ検索が「入口」として機能する．一方で，ハッシュタグはすべての投稿に付与されるわけではなく，タグ検索だけでは拾いきれない投稿が残る．そのため，投稿内容から適切なハッシュタグを自動的に推定・推薦する研究が行われてきた．従来研究では，投稿文を入力として，ハッシュタグを出力するマルチラベル分類として扱うことが多いが，この定式化では対象タグ集合を事前に固定する必要があり，SNS 上で継続的に生まれる新規タグや流行の変化に追随しにくく，集合外のタグを扱えない．

そこで本研究では，マルチラベル分類ではなく「投稿文とハッシュタグの意味的な近さ」に基づいてタグを選択する問題として捉える．その際，投稿文とタグ候補を同一の尺度で比較できるよう，両者を共通の埋め込み空間に配置して類似度を距離によって直接計算できるようにする．

投稿文もハッシュタグも同一の言語モデルで埋め込めば十分にも思えるが，ハッシュタグは文脈を持つ自然文ではなく短い記号列であり，表層文字列だけでなく「どのタグと一緒に使われるか」といった共起関係にも意味が反映されると考えられる．そこで本研究では，ハッシュタグ共起グラフに基づくグラフ埋め込みによりタグ表現を学習し，さらに投稿文と付与ハッシュタグを正例とする対照学習により，投稿文表現とタグ表現を対応付けして共通空間上で直接比較可能にする．

実験では，X（旧 Twitter）の投稿データを用い，二択のハッシュタグ選択タスクにおいて，投稿文・ハッシュタグを同一モデルで埋め込むベースラインと比較し高い精度を確認した．

2　関連研究

2.1　ハッシュタグ予測に関する研究

SNS の投稿文からハッシュタグを予測する研究は，投稿文を入力としてハッシュタグを出力するマルチラベル分類問題として扱われることが多い [1]．Zhang ら [2] は，X（旧 Twitter）の投稿を用いて事前学習した BERT ベースの多言語モデルを用いて，各言語で頻出するハッシュタグ集合に対する予測を行った．

しかし，これらの研究はいずれも，あらかじめ定義された有限個のハッシュタグ集合に基づく分類問題として設計されている．これに対し，本研究ではハッシュタグ予測を分類問題として扱わず，投稿文とハッシュタグを共通の埋め込み空間に写像し，類似度に基づいてハッシュタグを選択する枠組みを採用することで，より柔軟なハッシュタグ選択を可能にする．

2.2　グラフ埋め込み手法とその応用

グラフ構造を用いてノード間の関係性を表現し，その構造情報から埋め込みを学習する手法は広く研究されてきた [3]．Hamilton ら [4] が提案した GNN の一種である GraphSAGE（Graph Sample and Aggregate）は，ノード近傍の情報を集約すること

で，規模の大きいグラフにも適用しやすい点に特徴がある．GraphSAGE は，リンク予測タスクにも応用されており，観測されたエッジを正例として，負例と区別するように埋め込みを学習することで，ノード間の関係性を捉えた埋め込みを得ることができる．

本研究では，GraphSAGE を利用し，ハッシュタグ間の関係性を表現するための埋め込みを学習する．学習されたハッシュタグ埋め込みは，投稿文とハッシュタグを対応付けるためのタグ側の初期表現として利用され，後段の対照学習において投稿文埋め込みとの対応付けを行う．

2.3 テキスト表現学習における対照学習

対照学習 [5] は，対応関係にあるデータ対を近づけ，対応しない対を遠ざけることで表現を学習する枠組みとして提案されている．この枠組みは，自然言語処理分野でも利用されており，教師あり・教師なしの両設定で文埋め込み学習に応用されている．たとえば教師あり SimCSE は NLI データセットにおける文対を用いて対照学習を行うことで，文埋め込みを学習する手法である [6]．

これらの研究は，主に文同士の意味的類似性を捉えることを目的としているが，対照学習は「意味的に対応するペア」を与えられるなら，粒度の異なるテキスト単位を共通空間に対応付ける問題にも適用できると考えられる．本研究ではこの考えに基づき，投稿文とハッシュタグという粒度の異なるテキスト単位に対して対照学習を行い，両者の対応関係を学習する．

3 提案手法

本研究では，投稿文と関連するハッシュタグが埋め込み空間上で近傍に配置されるような，共通の埋め込み空間を構築する手法を提案する．提案手法は，(i) ハッシュタグ間の関係性を考慮した埋め込み学習と，(ii) 投稿文とハッシュタグの対応関係を学習する対照学習の 2 段階から構成される．

3.1 ハッシュタグ埋め込みの学習

1 つの投稿文に含まれる複数のハッシュタグは，その投稿内容に基づく意味的な関連を持つと考えられる．本研究では，この仮定に基づき，ハッシュタグ間の関係性を表現するための埋め込みモデルを学習する．

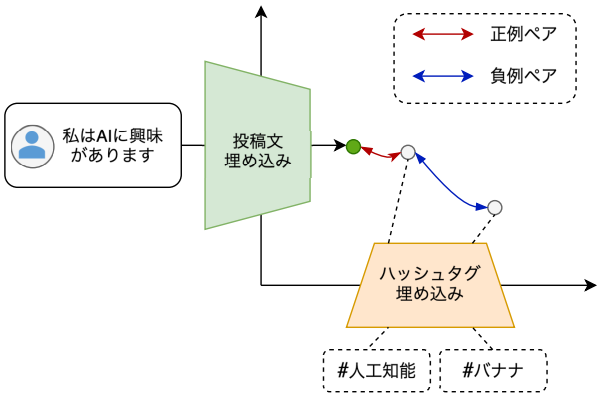

図 1 学習イメージ．投稿文と付与されていたハッシュタグを正例ペアとする．図中の投稿例は，筆者による作例．

ハッシュタグ共起グラフの構築 ハッシュタグをノードとする無向グラフを構築する．1 つの投稿に複数のハッシュタグが含まれる場合，その全組合せにエッジを張る．また，共起回数が閾値[1)]未満のペアはノイズになり得るため，エッジとして採用しない．

学習方法 ハッシュタグ埋め込みはリンク予測タスクとして学習する．実際に存在するエッジを正例エッジとし，正例エッジの片側ノードを固定した上で，他方をランダムに置換して負例エッジを生成する．正例エッジと負例エッジのスコア差が一定のマージン[2)]以上となるよう，マージン付きランキング損失を用いて学習を行う．スコア関数には，ハッシュタグ埋め込み同士の内積を用いる．

グラフの埋め込み手法には，GraphSAGE を用い，学習したハッシュタグ埋め込みを，次節の対照学習におけるハッシュタグエンコーダの初期表現として利用する．

3.2 投稿文とハッシュタグの対照学習

3.1 節で得たハッシュタグ埋め込みを投稿文の埋め込みと比較可能にするために対照学習を行う．図 1 に提案手法の学習イメージを示す．

投稿文 x_i に対し，その投稿内で付与されていたハッシュタグ h_i を正例ペアとし，ミニバッチ内の正例以外のランダムなハッシュタグをバッチ内負例として，投稿文とハッシュタグの対応関係を学習する．投稿文エンコーダを $f(\cdot)$，ハッシュタグエンコーダを $g(\cdot)$ とすると，投稿文 x_i およびハッシュタグ h_i は，それぞれ

1) 本研究では，閾値を 5 とする．
2) 本研究では，マージン γ を 1 とする．

$$z_{x_i} = f(x_i), z_{h_i} = g(h_i) \tag{1}$$

として共通の埋め込み空間上の表現に変換される．本研究では，対照学習において投稿文エンコーダ $f(\cdot)$ とハッシュタグエンコーダ $g(\cdot)$ の双方を更新する．投稿文 x_i，正例となるハッシュタグ h_i から成るデータセット $D = \{(x_i, h_i)\}_{i=1}^{m}$ が与えられた時，対照学習の損失関数は以下のようになる．

$$\ell_i = -\log \frac{e^{sim(z_{x_i}, z_{h_i})/\tau}}{\sum_{j=1}^{N}(e^{sim(z_{x_i}, z_{h_j})/\tau})} \tag{2}$$

ここで，N はバッチサイズ，τ は温度付きソフトマックス関数の温度パラメータである．また，ベクトル同士の類似度を計算する関数 $sim(\cdot)$ には，コサイン類似度を用いる．

4 実験

4.1 データ収集

本研究で使用するデータは，Twitter API V2 の Academic Research アクセス[3]を用いて取得した．ランダムに抽出した 15,000 ユーザを対象とし，2006 年 3 月 21 日〜2021 年 8 月 31 日の期間に投稿された投稿文を収集した．すべての投稿文に共通する前処理として，リツイート，画像・動画を含むツイートは除外した．

4.2 データセット

ハッシュタグ埋め込みモデルの学習には，4.1 節で収集した 15,000 ユーザの投稿文に含まれる 28,865 種類のハッシュタグを使用した．

投稿文とハッシュタグの対照学習のため，データセットはユーザ単位で訓練・検証・評価が 7:1:2 になるようにユーザを分割した（それぞれ 10,500 / 1,500 / 3,000 ユーザ）．このようにユーザ単位で分割することで，同一ユーザに由来する投稿が異なるデータセットに含まれることを防ぎ，ユーザ固有の投稿傾向に起因する過学習を抑制する．各データセットに含まれる投稿数を表 1 に示す．

ハッシュタグ選択タスク　提案手法の有効性を評価するために，ハッシュタグ選択タスクを設計した．本タスクは，評価データセットから抽出した投稿文に対し，正例となるハッシュタグと負例となるハッシュタグの 2 つの候補を提示し，正しいハッシュタグを選択するタスクである．正例ハッシュタグは，投稿文に実際に付与されていたハッシュタグからランダムに 1 つ選択する．負例ハッシュタグは，データセットに含まれる全ハッシュタグの中からランダムに選択する．モデルは，投稿文の埋め込みと各ハッシュタグの埋め込みのコサイン類似度を計算し，類似度が高い方を予測結果として選択する．評価データセットのために割り当てた 3,000 ユーザの投稿文から 30,000 ペアを抽出し，評価に用いた．

表 1　各データセットの投稿数

データセット	投稿数
訓練データセット	14,348,691
検証データセット	2,050,260
評価データセット	4,140,965
合計	20,539,916

4.3 ハッシュタグ埋め込みモデル

本研究で学習するハッシュタグ埋め込みモデルの出力次元数は，投稿文の埋め込みモデルと同一の 768 次元とした．これは，後段の対照学習において，ハッシュタグの埋め込みと投稿文の埋め込みを同一の次元数で扱い，類似度の計算を行うためである．ハッシュタグ埋め込みの学習には GraphSAGE を用いた．GraphSAGE の学習時の各種パラメータ設定は，付録 A.1 に示す．

4.4 投稿文とハッシュタグの対照学習

投稿文埋め込みモデル　投稿文の埋め込みモデルのベースには，東北大学が公開している事前学習済みの日本語 BERT モデル[4]を用いた [7]．投稿文の表現として，[CLS] トークンに対応する埋め込みを使用し，出力される次元数は，ベースモデルと同じ 768 次元である．投稿文とハッシュタグの対照学習においては，3.2 節で定義した損失関数に基づき，投稿文エンコーダとハッシュタグエンコーダの双方を更新する．学習時の各種パラメータ設定は，付録 A.1 に示す．

ベースライン　提案手法の有効性を評価するために，ハッシュタグ選択タスクの比較対象として以下のモデルを設定した．

3） https://developer.twitter.com/en/docs/twitter-api

4） tohoku-nlp/bert-base-japanese-v3

1 つ目の比較モデルとして，投稿文埋め込みモデルのベースとなる日本語 BERT モデルを用い，[CLS] トークンに対応する埋め込みをそのまま使用した．2 つ目の比較モデルとして，OpenAI が公開している埋め込み特化モデル[5]を用いた [8]．なお，これらのモデルは，投稿文とハッシュタグの双方を同一のモデルに入力し，得られた埋め込み間のコサイン類似度に基づいてハッシュタグを選択する．ハッシュタグは，「#」を除去した文字列として入力した．

5 実験結果

提案手法が，投稿文とハッシュタグを共通の埋め込み空間に配置し，「空間上で近い＝意味的に対応する」という関係に基づいてハッシュタグを選択できることを検証する．

5.1 ハッシュタグ埋め込みモデルの検証

ハッシュタグ埋め込みモデルがハッシュタグの意味的関係を適切に捉えているかを確認するため，基準となるハッシュタグに対し，埋め込み空間上で近傍に位置するハッシュタグを抽出した．類似度の計算にはコサイン類似度を用い，上位 5 件の近傍ハッシュタグを抽出した．

表 2 に，基準ハッシュタグに対する近傍ハッシュタグの例を示す．基準となるハッシュタグと意味的に関連性の高いハッシュタグが埋め込み空間上で近傍に配置されていることが確認できる．

この結果は，同一投稿内で共起するハッシュタグは，投稿内容に基づく意味的な関連性を持つというハッシュタグ埋め込みモデル学習時の前提が妥当であることを示唆しており，ハッシュタグ側の初期表現として適切に機能すると考えられる．

表 2 近傍ハッシュタグの検索結果 (Top5).

検索ハッシュタグ：# コロナウイルス	
1	# コロナ
2	# 新型コロナウイルス
3	# 新型コロナ
4	# 緊急事態宣言
5	# 新型コロナウィルス

5.2 ハッシュタグ選択タスク

投稿文とハッシュタグの対応関係を学習できているかを評価するため，4.2 節で設計したハッシュタグ選択タスクを用いて性能比較を行った．本タスクは，2 値分類問題として設計されており，ランダムに選択した場合の精度は 0.5 となる．

表 3 に，各モデルのハッシュタグ選択タスクの精度を示す．提案手法は，ベースモデルや OpenAI の埋め込み特化モデルと比較して性能が大幅に高い精度を示した．

この結果は，(i) 共起グラフに基づくハッシュタグの埋め込み獲得と，(ii) 投稿文とハッシュタグの対照学習による対応付けを組み合わせることで，「近い＝対応する」という関係が成立した共通埋め込み空間を構築できたことを示唆する．

投稿文と付与されたハッシュタグのように，対応関係がペアとして観測可能であり，かつ対応が一定程度一貫している場合，本手法のような共通の埋め込み空間を構築する手法は，有効に機能すると考えられる．

表 3 ハッシュタグ選択タスクの精度比較

モデル	精度
ベースモデル	
tohoku-nlp/bert-base-japanese-v3	0.564
***OpenAI* 埋め込みモデル**	
text-embedding-3-small	0.581
提案手法	**0.849**

6 おわりに

本研究では，SNS 投稿文とハッシュタグの対応関係を捉えるため，投稿文とハッシュタグを共通の埋め込み空間に写像し，類似度に基づいてハッシュタグを選択する手法を提案した．提案手法は，(i) ハッシュタグ共起グラフからタグ埋め込みを学習し，(ii) 投稿文とハッシュタグの対照学習により埋め込み空間上で対応付ける，2 段階の学習から構成される．二択のハッシュタグ選択タスクにおいて，提案手法は既存の文埋め込みに基づく比較モデルより高い精度で正例タグを選択でき，提案した共通埋め込み空間が投稿文とハッシュタグの対応を反映することを確認した．

今後の課題は，より実運用に近い Top-K 推薦としての評価や，より難易度の高い負例（hard negative）を含む設定での検証が挙げられる．また，流行語や新規タグの出現に伴う時間変化を考慮した拡張や，多言語対応なども今後の研究課題である．

5） text-embedding-3-small

謝辞

本研究は, JSPS 科研費（JP23K22076, JP25K21648）, JST さきがけ（JPMJPR2461）, およびセコム科学技術財団特定領域研究助成の助成を受けたものです.

参考文献

[1] Haoran Huang, Qi Zhang, Yeyun Gong, and Xuanjing Huang. Hashtag recommendation using end-to-end memory networks with hierarchical attention. In **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**, pp. 943–952, December 2016.

[2] Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In **Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining**, KDD '23, pp. 5597–5607, 2023.

[3] Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. **Knowledge-Based Systems**, Vol. 151, pp. 78–94, 2018.

[4] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In **Advances in Neural Information Processing Systems**, Vol. 30, 2017.

[5] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. **Technologies**, Vol. 9, No. 1, p. 22, 2020.

[6] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910, November 2021.

[7] 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, June 2019.

[8] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas A. Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David P. Schnurr, Felipe Petroski Such, Kenny Sai-Kin Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training. **arXiv preprint abs/2201.10005**, 2022.

A　付録

A.1　訓練時の各種パラメータ設定

本研究で使用した各モデルの学習時の各種パラメータ設定は以下の通りである．

ハッシュタグ埋め込みモデル　ハッシュタグ埋め込みの学習には GraphSAGE を用いた．ハッシュタグ共起グラフを入力とし，リンク予測タスクとして学習を行った．

- グラフ埋め込み手法: GraphSAGE（mean aggregator）
- 層数: 2
- エポック数: 1000
- 学習率: 1e-3
- 出力次元数: 768

投稿文とハッシュタグの対照学習　投稿文エンコーダには事前学習済みの日本語 BERT モデルを用い，ハッシュタグエンコーダには 3.1 節で学習したハッシュタグ埋め込みモデルを初期値として用いた．

- エポック数: 10
- バッチサイズ: 128
- 学習率: 5e-5
- 温度パラメータ: 0.05
- 出力次元数: 768

A.2　投稿文から近傍となるハッシュタグの検索

投稿文から近傍となるハッシュタグを検索するために，提案手法を用いて投稿文とハッシュタグの埋め込みを計算し，コサイン類似度に基づいて近傍ハッシュタグを抽出した．表 4 に，投稿文に対する近傍ハッシュタグの例を示す．検索文に対し，意味的に関連するハッシュタグが近傍に配置されていることが確認できる．

表 4　投稿文から近傍ハッシュタグの検索結果 (Top5).

検索文：開会式めちゃ面白かった!!

1	# 東京 2020
2	# 卓球混合ダブルス
3	# nhk2020
4	# オリンピック開会式
5	# tokyoolympics

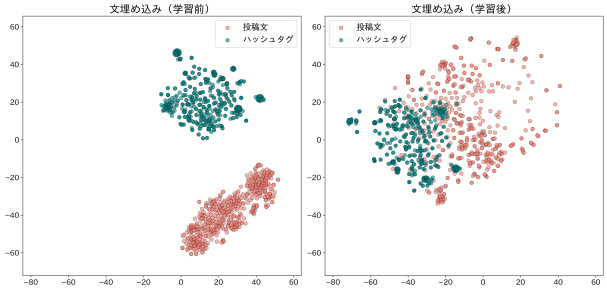

図 2　投稿文（赤）とハッシュタグ（緑）の埋め込みの 2 次元可視化（左：対照学習前，右：対照学習後）．

A.3　共通埋め込み空間の分析

3.2 節の対照学習により，投稿文側の埋め込みとハッシュタグ側の埋め込みが埋め込み空間上で対応付けされることを確認するため，学習前後の投稿文-ハッシュタグの埋め込みを 2 次元に可視化した．評価データセットからランダムに抽出した 500 ペアを対象とし，t-SNE により次元削減を行った．図 2 に，投稿文とハッシュタグの埋め込みの 2 次元可視化結果を示す．左図は対照学習前，右図は対照学習後の可視化結果である．

学習前は投稿文（赤）とハッシュタグ（緑）がほぼ別領域に分離しており，両者の対応関係が捉えられていないことがわかる．一方，学習後には両者が同一領域に分布し，混在が生じている．これは対照学習によって，投稿文側の埋め込みとハッシュタグ側の埋め込みが同一空間上で対応付けられ，投稿文近傍に「対応しうるタグ」が配置されるようになったことを示唆する．ただし，図 2 は t-SNE による次元削減に基づく可視化であり，解釈には限界があることに注意が必要である．

A.4　時系列情報を考慮したハッシュタグの多義性への対応

SNS 上のハッシュタグは，時系列に伴い意味が変化したり，多義的に用いられたりすることがある．例えば，「# コロナ」というハッシュタグは，2020 年以前には「コロナビール」などを指すことが多かったが，2020 年以降は「新型コロナウイルス感染症」を指すことが圧倒的に多くなった．

このようなハッシュタグの多義性は，ハッシュタグの埋め込み表現の揺れを引き起こす可能性がある．今後は，この問題に対応するために，ハッシュタグの埋め込みモデルを学習する際に，特徴量として時系列情報を組み込むことが考えられる．