Style SimSCE: SNS ユーザ同一性に基づく対照学習による スタイル類似性を捉えた文ベクトルの獲得

仲田明良 [1]　狩野芳伸 [1]
[1] 静岡大学 情報学部
{anakada,kano}@kanolab.net

概要

自然言語には，意味的内容だけでなくどのように表現するかという書き手特有のスタイルが含まれる．X（旧：Twitter）をはじめとする SNS では特にその傾向が強く，ユーザの投稿文におけるスタイルはユーザ自身の個性や社会的背景を反映しており，多種多様であると考えられる．本研究では，同一アカウントの投稿が大量に取得可能なことに着目し，投稿文ペアが同じユーザが書いたものか異なるユーザのものかユーザ同一性に基づいた対照学習を用いて，SNS 投稿文のスタイル類似性を捉えた文ベクトルを学習させることを提案する．この手法によって得られる文ベクトルは，単純なコサイン類似度の比較のみで高精度に同一ユーザによる投稿文かを識別することを可能にする．さらに人手評価により，提案手法の出力する文ベクトルの類似度は既存の文埋め込みモデルよりもスタイルの近さを反映していることを示した．

1　はじめに

現代社会に普及したソーシャルメディア（以下，SNS）では，異なる地域・文化・世代・言語背景を持つ人々が日々多様な情報を発信している．SNS の投稿文には，そのような多様な人々の個別の言語的スタイル [1] が反映されている．例えば，語尾に「だぜ」と付けるユーザと「です」と付けるユーザとでは，同じ内容の投稿文でも，そのスタイルは大きく異なる．

我々は，このような言語的スタイルは，ユーザの各投稿に一貫性を持って現れると考えた．この仮定に基づき，本研究では，ユーザの投稿中に現れるスタイルの類似性を捉えた文ベクトルを非常にシンプルな手法で学習させた **Style SimCSE** を提案する．Style SimCSE の学習では，同じユーザの投稿文ペアを学習済みエンコーダに入力して得られたベクトル表現を正例ペアとし，異なるユーザの投稿文ペアのベクトル表現を負例ペアとした対照学習を行う．この手法によって得られるベクトル表現により，単純なコサイン類似度の比較のみで高い精度でユーザを識別することができた．また，クエリ検索に用いることで，既存の文埋め込みモデルと比べて，文のスタイルをより効果的に捉えられていることを人手評価により示した．

2　関連研究

2.1　スタイルの類似性を捉えたベクトル

赤間ら [2] は，「同一発話内に含まれる単語は同一のスタイルを持つ」という仮定のもと，スタイルの類似性を捉えた単語ベクトル空間を獲得する手法を提案した．しかし，この手法では，単語単位でのスタイルの類似性は捉えることが出来るが，文全体のスタイルの類似性を捉えることは出来ない．

文全体のスタイルの類似性を評価可能な文ベクトルの学習手法として，銭元ら [3] は，小説内に登場する同一人物の発話を正例のペアとした対照学習を提案した．しかし，この手法では，小説内で登場する人物のスタイルにしか対応できないという問題があると共に，このようにして得られた文ベクトルは現実世界の人間のスタイルを必ずしも反映しておらず，適用対象が限定される可能性がある．

2.2　教師あり対照学習

対照学習（Constrastive Learning）[4] は，ラベル付けされていないデータの中から，互いに類似した事例のペア（正例）と異なる事例のペア（負例）を取り出し比較を行うことでモデルの学習を行う手法である．自然言語処理においては，対照学習を用いて文埋め込みの訓練に利用することがある．

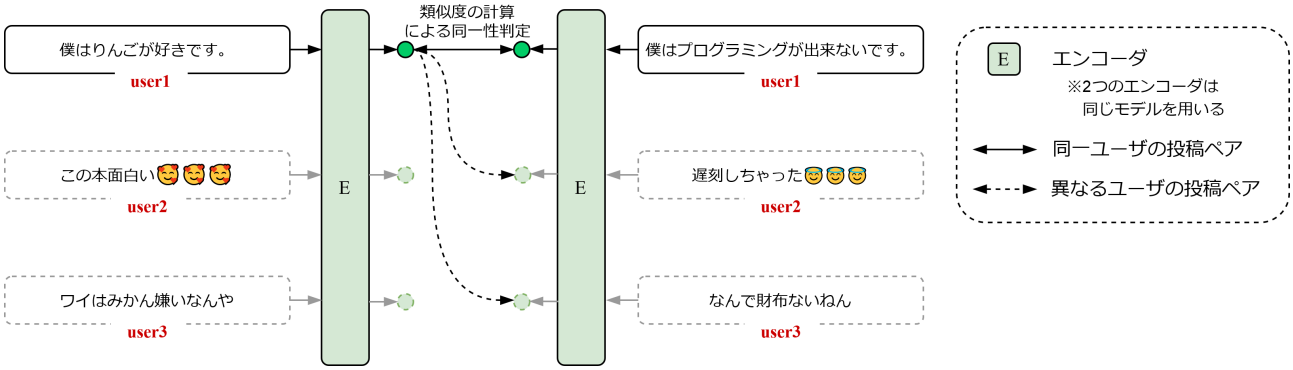

図 1 Style SimCSE の学習イメージ．ペアの各文を左右に示している．図中の各ツイート例は，筆者による作例．

Gao ら [5] が提案した教師あり SimCSE は，対照学習を用いた文埋め込みの学習に教師ありの設定を適用したモデルである．自然言語推論（Natual Language Inference，NLI）データセットの前提文に対して，「含意」のラベルが付いた仮説文とのペアを正例のペアとし，ミニバッチ内の正例以外の事例を全てバッチ内負例とし，対照学習を行う．また，「矛盾」のラベルが付いた仮説文をハード負例（hard negative）として定義することでさらなる性能の向上が見られたと報告している．

3 ユーザ同一性に基づく対照学習

本研究では，ユーザの同一性に基づく教師あり対照学習を用いて，スタイルの類似性を捉えた文ベクトルを学習する．図 1 に，提案手法である Style SimCSE での対照学習のイメージを示す．

まず，アンカーとなる投稿文 x_i を抽出する．次に，アンカーと同じユーザの投稿文を正例のペアとして定義する (x_i^+)．そして，モデルに入力するミニバッチ内の正例以外の投稿文はすべて異なるユーザの投稿文で構成し，これらをバッチ内負例として対照学習を行う．また，過度に類似している 2 つの投稿文は，キャンペーン投稿や bot による投稿などの可能性があり，ユーザのスタイルを反映していないと考えられるため，正例，バッチ内負例ともに含めない．類似度の計算には，編集距離（Levenshtein 距離）[6] を用い[1)]，類似度が 0.7 を超える投稿文は除外する．

アンカーとなる投稿文 x_i，正例となる投稿文 x_i^+ から成るデータセット $D = \{(x_i, x_i^+)\}_{i=1}^m$ が与えられた時，h_i，h_i^+ がそれぞれのベクトル表現だとすると，対照学習の損失関数は以下のようになる．

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

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 SNS 投稿の収集と前処理

本研究で使用するデータは，Twitter API V2 の Academic Research アクセス[2)]を用いて取得し，ランダムに抽出した 15,000 ユーザの 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用いた．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．また，他のアカウントに対するリプライ及びメンションを含む投稿は，相手との関係性の違いによりユーザごとのスタイルの一貫性が損なわれる可能性があると考え除外した．さらに文字フィルタとして”RT”という単語を含む投稿を除外し，URL とハッシュタグは削除した．絵文字や顔文字に関しては，投稿のスタイルに関係すると考え，残すこととした．これらの前処理を行った投稿のうち，文字数が 15 文字以上，60 文字以下の投稿を抽出した．

4.2 学習用データセット

前章で述べた 15,000 ユーザの投稿文を，アカウント単位で訓練: 検証: テストが 7:2:1 になるよう分割した．結果，10,500 ユーザが訓練，1,500 ユーザが検証，3,000 ユーザが評価データセットとなった．なお，様々なユーザのスタイルを反映するため，投

1) 2 つの文字列間の編集距離を計算し，最大文字列長で正規化する．

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

稿数が非常に多いユーザに偏らないように各ユーザの最大投稿数は 20,000 件とし，これを超える投稿は除外した．各データセットの投稿数を表 1 に示す．

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

訓練データ 訓練には，アンカーとなる投稿文，正例となる投稿文の 2 つ組を 1,000,000 ペア抽出して用いた．

検証データ 検証では，文ペアを入力した時に訓練モデルが出力する類似度スコアを用いた．アンカーとなる投稿文とランダムに抽出した投稿文の 2 つ組を 20,000 ペア抽出し，検証データとして使用した．この時，ラベルとして，アンカーと同じユーザの投稿文であれば 1，異なるユーザの投稿文であれば 0 を付与した．そして，モデルが出力する 2 文の類似度スコアとラベルの間の Peason 相関係数を評価指標として用いた．

ユーザ同一性分類の自動評価 評価として，ユーザ同一性分類タスクを設計し，データセットを構築した．アンカーとなる投稿文，投稿文 1，投稿文 2 の 3 つ組が与えられた時，アンカーと同じユーザの投稿がどちらかを予測する 2 値分類タスクをユーザ同一性分類タスクとして設計した．予測ラベルは，モデルが出力したベクトル表現のコサイン類似度によって決定さした．評価データセットのために割り当てた 3,000 ユーザの投稿から 40,000 ペアを抽出し，評価に用いた．

スタイル的類似の人手評価 ユーザ同一性の類似が必ずしもスタイルの類似であるとは限らないため，クエリ検索を用いた評価を行った．検索文に対し，評価データセットのために割り当てられた 3,000 ユーザの 4,140,965 件の投稿文の中から，コサイン類似度が最も高い投稿文を出力させ，スタイルの類似性を捉えられていることを確認する．

また，人手評価として，提案モデルと比較モデルのどちらがより検索文のスタイルに近いかを 3 人の評価者に選択させた（付録 A.2）．検索文は，評価データセットよりランダムに 100 件抽出した．

表 2 ユーザ同一性分類タスクの精度比較

Model	Accuracy
Publicly Released Pretrained Model	
$\text{BERT}_{\text{base}}$	0.6075
Publicly Released Embedding Model	
Unsupervised SimCSE-$\text{BERT}_{\text{base}}$	0.6078
Supervised SimCSE-$\text{BERT}_{\text{base}}$	0.5997
Open AI Embedding Model	
text-embedding-ada-002	0.6203
Proposed Model	
Style SimCSE-$\text{BERT}_{\text{base}}$	**0.7728**

4.3 モデルの学習

提案モデル 本研究では，東北大学が公開している日本語 $\text{BERT}_{\text{base}}$ モデル[3]をベースモデルとして対照学習に用いた [7]．訓練時の各種パラメータ設定は，付録 A.1 に記載する．出力されるベクトル表現の次元数は，ベースモデルと同じ 768 次元である．

比較モデル 提案手法である Style SimCSE の比較対象として，下記の四つのモデルを用意した．一つ目として，StyleSimCSE のベースモデルである日本語 $\text{BERT}_{\text{base}}$ モデルの [CLS] トークンに該当するベクトル表現を用いた．二つ目と三つ目として，名古屋大学が公開している教師あり SimCSE モデル[4]と教師なし SimCSE モデル[5]を用いた [8]．これらは，Style SimCSE のベースモデルと同じ日本語 $\text{BERT}_{\text{base}}$ モデルに対して対照学習を行ったモデルであり，出力されるベクトル表現の次元数は，768 次元である．

さらに，OpenAI が公開している埋め込み特化モデル[6]も四つ目として比較対象として用いた [9]．このモデルが出力するベクトル表現の次元数は，1,536 次元である．

5 実験結果

ユーザ同一性分類の評価 表 2 に，モデルごとのユーザ同一性分類タスクの結果を示す．既存の事前学習済みモデルや SimCSE モデルと比較して，提案手法である Style SimCSE は，得られるベクトル表現の単純な類似度比較のみで高い精度でユーザを識別

3) https://huggingface.co/cl-tohoku/bert-base-japanese-v3
4) https://huggingface.co/cl-nagoya/sup-simcse-ja-base
5) https://huggingface.co/cl-nagoya/unsup-simcse-ja-base
6) text-embedding-ada-002

表 3　クエリ検索結果 (Top3)．表中の各ツイートは，匿名化加工済み

検索文：ウチは { 食べ物 (1)} 好きやねん	
比較モデル：Supervised SimCSE-BERT$_{\text{base}}$	
#1	めっちゃ { 食べ物 (1)} 食いたいやん。それは。でもちゃんとチチしてるの凄い!
#2	{ 食べ物 (1)} 行きたくなってきちゃったわー。
#3	{ 食べ物 (1)} を食べに行くぎゆたんたまらん好きなんですが…
提案モデル: Style SimCSE-BERT$_{\text{base}}$	
#1	コレあかんヤツやんか { 状況 } やんか
#2	ウチ、{ 食べ物 (2)} はご飯に掛ける派やねん。
#3	塩抜き { 食べ物 (3)} 食べたいホンマなんやろか？

することが可能であった．さらに，OpenAI の埋め込み特化モデルと比較しても，提案手法の方が高い精度でユーザを識別できた．

スタイルの類似性評価　表 3 に，検索文に対する類似度上位 3 件の例を示す．比較モデルには，前章で述べた教師あり SimCSE モデル[7)]を用いた．既存の埋め込みモデルは，文章の意味に関わる単語である「食べ物 (1)」を含む投稿文が上位に出力されていた．一方，提案手法である Style SimCSE は，検索文の「ウチ」や「やねん」に反応していると考えられる投稿文が上位に出力されていた．

人手評価の結果を表 4 に示す．結果より，提案手法である Style SimCSE は，既存の埋め込みモデルと比較し，文章のスタイルの類似を効果的に捉えられていた．

表 4　スタイル的類似の人手評価

評価者	A	B	C	平均
比較モデル	19%	11%	21%	17% (51 / 300)
提案モデル	**81%**	**89%**	**79%**	**83%** (249 / 300)

6　考察

表 2 に示したように，提案手法である Style SimCSE は，既存の文埋め込みモデルと比較し，高い精度でユーザを識別できた．意味的内容を捉えるように学習が行われる既存の文埋め込みモデルでは精度が低かったことから，ユーザを識別する上では，意味的な内容は必ずしも重要ではない可能性がある．本手法でユーザが識別できるという事は，同一ユーザの投稿に一貫性を持って現れる意味的内容ではない「何か」を捉えることが出来た可能性があり，これがユーザ特有のスタイルであると考える．

Style SimCSE と既存の埋め込みモデルを用いたクエリ検索の結果を比較すると（表 3），Style SimCSE では意味的内容でなく文章に現れるスタイルの類似性を捉えられていることが分かる．この例では，関西地方の方言である「やねん」を入力として与えたときに，関西地方の方言を含む投稿文が上位に出力されている．方言はユーザの地域背景を反映するものであり，ユーザの各投稿に一貫性を持ってスタイルとして現れると考えられる．また，一人称に関しても同様にユーザ特有のスタイルとして現れると考えられる．提案手法では，「ウチ」という一人称を含む投稿文が上位に出力されている．方言や一人称は一例であり，提案手法ではこのようなユーザの個性や社会的背景を反映した様々なスタイルを捉えることが可能であると考えられる．（付録 A.3）

7　おわりに

本研究では，SNS 投稿文のスタイルの類似性を捉えた文ベクトルを，ユーザの同一性に基づいた対照学習を用いて学習させることを提案した．提案手法である Style SimCSE を用いて得られるベクトル表現は，文章に含まれるスタイルの類似性を捉え，高い精度でユーザを識別できた．この手法は，ユーザの数だけスタイルが存在しうる SNS 投稿文のスタイルを捉えるための有効な手法であると考えられる．

今後は，得られたベクトル表現がユーザの個性や社会的背景を反映しているか定量的に評価することや，ユーザ投稿文のスタイルの変化を捉えることが可能であるかを検証したい．また，生成モデルと組み合わせることで，ユーザのスタイルをより反映した文章生成を試みたい．

7）　https://huggingface.co/cl-nagoya/sup-simcse-ja-base

謝辞

本研究は，JSPS 科研費 JP22H00804，およびセコム科学技術財団特定領域研究の助成を受けたものです.

参考文献

[1] Kalaivani Sundararajan and Damon Woodard. What represents "style" in authorship attribution? In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 2814–2822, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[2] 赤間怜奈, 渡邉研斗, 横井祥, 小林颯介, 乾健太郎. スタイルの類似性を捉えた単語ベクトルの教師なし学習. 人工知能学会全国大会論文集, Vol. JSAI2018, pp. 1N203–1N203, 2018.

[3] 銭本友樹, 古俣慎山, 宇津呂武仁. 対照学習による口調の類似性評価のための文ベクトルの獲得. 人工知能学会全国大会論文集, Vol. JSAI2023, pp. 4A2GS605–4A2GS605, 2023.

[4] 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.

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. **Soviet physics. Doklady**, Vol. 10, pp. 707–710, 1965.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, **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.

[8] Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. Japanese SimCSE Technical Report. **arXiv:2310.19349**, 2023.

[9] 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**, Vol. abs/2201.10005, , 2022.

A　付録

A.1　訓練時の各種パラメータ設定

訓練時の各種パラメータ設定は以下の通りである．

- バッチサイズ：128
- エポック数：3
- 検証ステップ：250
- 学習率：$5e-5$
- 温度パラメータ：0.05

A.2　スタイル的類似の人手評価

スタイル的類似を測るために行った人手評価の質問内容を図 2 に示す．

質問内容

文 1・文 2 について，基準文とスタイル（文体）が似ている方を答えてください．
スタイル（文体）とは，「その作者にみられる特有な文章表現上の特色」の事です．

- 文 1 の方が似ている場合：1
- 文 2 の方が似ている場合：2

図 2　人手評価の質問内容

A.3　様々なスタイルの類似性を捉えた文ベクトル

表 3 に示した検索文と意味的内容は同じだが，スタイルが異なる検索文を用いてクエリ検索を行った結果を表 5 に示す．既存の文埋め込みモデルでは，表 3 に示した検索結果と似たような結果が得られていることが分かる．しかし，提案手法である Style SimCSE では，検索文のスタイルに反応していると考えられる「僕」や「です。」を含む投稿文が上位に出力されており，表 3 に示した検索結果とは異なる結果が得られていることが分かる．

表 5　クエリ検索結果（Top3）表中の各ツイートは，匿名化加工済み

検索文：僕は { 食べ物 (1)} 好きです。	
比較モデル：Supervised SimCSE-BERT$_{\text{base}}$	
#1	{ 食べ物 (1)} はちょっとだけ焼きたい派です。
#2	{ 食べ物 (1)} が食いたい。ただし、うまいものに限る。
#3	{ 食べ物 (1)} 食べ放題食べてきました^^
提案モデル：Style SimCSE-BERT$_{\text{base}}$	
#1	僕は { 食べ物 (4)} が好きです。
#2	僕は肉じゃがより { 食べ物 (5)} 派です。
#3	僕は { 履物 } が嫌いです。理由はない。