Faiss を用いたデータ拡張による ポジティブテキストリフレーミングの精度向上

松田 龍之介 [1] 徐 勝 [2] 福本 文代 [3] 鈴木 良弥 [3]
山梨大学工学部 [1] 山梨大学大学院 [2] 山梨大学総合研究部工学域 [3]
{t21cs047,g22dts03,fukumoto,ysuzuki}@yamanashi.ac.jp

概要

入力文の意味を保持したまま否定的な感情を肯定的な感情へ変換するポジティブテキストリフレーミング (PTR) タスクは，大規模言語モデルの進展により広く研究されている．PTR タスクでは，事前学習済のモデルを利用し，対象データに併せてファインチューニングする手法が多く提案されている．しかし，PTR タスクにおけるファインチューニングでは，ソース文とターゲット文は同義であり，ソース文は否定を，ターゲット文は肯定を表す大量のパラレルデータが必要であることから，データ不足が問題となることが多い．本研究は，データ拡張手法により生成したパラレルデータを用いることにより，PTR の精度向上を目指す．

1 はじめに

テキストスタイル変換 (TST) は，入力文の意味内容を保ちながら，文のスタイルを変換するタスクのことを指す．例えば感情スタイル変換タスクは，文の意味内容を保持しつつ，肯定的な文と否定的な文を双方向に変換するタスクであり，この分野で近年注目されているタスクの一つである．代表的な論文として，Shen ら [1] の非パラレルデータを用いたスタイル変換のための，クロスアライメントを活用したアプローチがある．TST のサブタスクとして，ポジティブテキストリフレーミング (PTR) タスクが挙げられる．PTR タスクは，入力文の感情表現を除いた主要な内容を保持したまま，肯定的な視点を与える文を生成するタスクである．最近の研究には，Jia らの肯定的な表現への変換を多次元で最適化する手法がある [2]．例えば，”This week has prob been the worst week I have had with school in a while.”という入力文に対し，その出力は”This week was not the best, but it can get better.”となる．

近年の自然言語処理の進展により，事前学習済みモデルが多数提案されている．例えば，Vaswani ら [3] が提案した Transformer を用いた BERT [4]，及び Raffel らが提案した T5[5] などの事前学習済み言語モデルが挙げられる．事前学習済みモデルは，対象タスクのドメインに併せたラベル付きデータやパラレルデータを用い，これらをファインチューニングすることにより，質の高いモデルを作成することができる．PTR タスクで用いるパラレルデータは，意味が互いに類似している否定的な文と肯定的な文の組である．近年，感情分析に関する研究，例えば，Pang らが提案した機械学習を使用した感情分析 [6] や Go らの SNS でのポストされたテキストを使用して感情分類を行う手法 [7] など多数行われていることから，感情分析に必要となる肯定的，及び否定的なテキストデータが多く存在するものの，これらはパラレルデータではないため，PTR タスクにおいてそのまま利用することはできない．

本研究は，テキストの意味を表す内容とスタイルとを分離し，スタイルを変換した後，意味を保持したまま表現を変換する (ST2PG)，または，意味を保持したまま表現を変換した後，スタイルを変換する (PG2ST) 先行研究 [8] に対し，密ベクトル検索のためのライブラリである Faiss を使用しパラレルデータを増強することにより，ポジティブテキストリフレーミングの精度向上を目指す．

2 関連研究

PTR の先駆的な研究として Ziems らの研究がある [9]．Ziems らは，自己肯定 (Self-Affirmation) や感謝 (Thankfulness) などのような心理学的に動機づけられた 6 つのリフレーミング戦略を使用し，8,349 組のパラレルデータを含む POSITIVE PSYCHOLOGY FRAMES (PPF) データセットを構築した．このデータセットは，意味保持が求められる高度なタスク

のベンチマークとして利用されている．実験では，Transformer ベースの生成モデル（BART や T5）が特に効果的である一方，生成品質には改善の余地があることが報告されている．Lai らは，目標とするスタイルとコンテンツに対する 2 種類の報酬を用いることにより，入力文から肯定的な文を生成する手法を提案した [10]．しかし，これらの研究はいずれもスタイルと内容とが複雑に絡み合っているような入力文には十分に対応できないため，精度面で課題が残されている．

Xu らはこれら問題に対処するため，PTR タスクにおいて，文の内容とスタイルを分離し，スタイルを肯定的に変更後，文の内容を保持したまま別の表現に変換すること，あるいはその逆を行い，統合する新しいフレームワークを提案した [8]．また，PTR タスクにおいてパラレルデータセット数が少ないという問題に対し，擬似的なデータセットを生成する 2 つのデータ拡張手法を用い，マルチタスク学習によりそれらを学習することで，文の内容を保持したまま，肯定的な文へ変換する手法を提案した．PPF データセット，及び BART や T5 等の言語モデルを用いた実験を行った結果，提案手法がベースラインを超えることが示された．さらに，生成テキストの多様性や流暢さの向上にも貢献することが報告されている．しかし，感情ラベルである肯定，否定が付与された Yelp データはそれぞれ約 26 万，及び約 17 万文存在するため，これらのデータの一部を用い，意味的類似性を計算することでパラレルデータを作成した．したがって，高精度なデータ増強には至っていない．

本研究は，Xu らの手法に対し，近似最近傍探索ライブラリ Faiss によるデータ拡張手法を用いることにより，PTR の精度向上を目指す．

3 提案手法

3.1 文の内容とスタイルとの分離

本研究は，PTR として Xu らの手法を用いる．手法の流れを図 1 に示す．図 1 で示すように，Xu らの手法は 2 段階のファインチューニングを行う．ステージ 1 では，モデルに 2 つのエンコーダを用意し，それぞれパラフレーズ生成 (PG) のための MSCOCO データセットと，感情変換 (ST) 用の Yelp データセットを用い，マルチタスク学習によりそれぞれファインチューニングを行う．

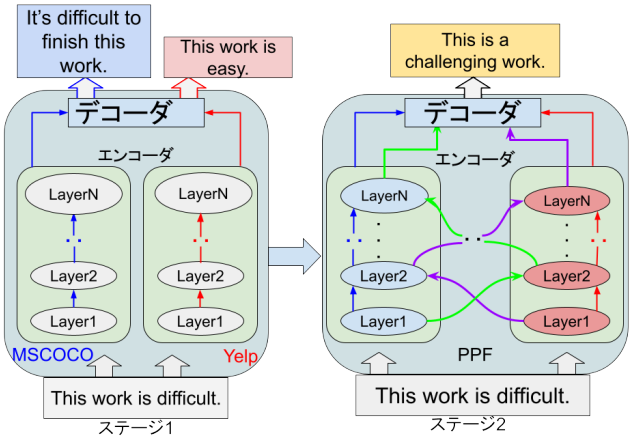

図 1 文内容とスタイルとの分離による PTR タスク [8]

ステージ 2 では，PPF データセットを用い，4 種類のフロー，すなわち (1) ステージ 1 で学習した PG のエンコーダを用いて PPF を学習する PG，(2) ST のエンコーダを用いて PPF を学習する ST，(3) エンコーダの PG 及び ST の各 Layer を交互に重ね，PPF で学習する PG2ST (PG から ST)，及び (4) ST2PG (ST から PG) によりそれぞれ学習する．

データの作成方法は以下の通りである．MSCOCO は入力文と正解文で構成されている．また，入力文が否定的な文のとき，正解文は中立的な文と肯定的な文，入力文が中立的な文のとき，正解文は肯定的な文となるような組を作成している．また，Yelp データセットのソースデータは，非パラレルデータであるため，一部のデータを使用して，類似度計算を用いることにより組を作成し，データセット STSB と BERT モデルを使用して，テキストの意味的類似性が，3.0 以上のスコアになるよう抽出した．その際の問題点として，データ量の少なさ，同じテキストを複数回使うことによるデータの偏り，及び肯定的な文への変換精度の低さが挙げられる．

3.2 Faiss

Faiss (Facebook AI Similarity Search) は，大規模な埋め込みベクトルの効率的な管理を目的としたオープンソースライブラリである．Johnson らが提案した類似性検索 [11] を大規模データセットで効率的に実行するための GPU 最適化手法を基にしており，近似最近傍検索アルゴリズムを中心に，インデックス作成，クラスタリング，圧縮，及び変換などの機能を提供する．検索精度，速度，及びメモリ使用量のバランス調整が柔軟で，GPU 対応により高速な処

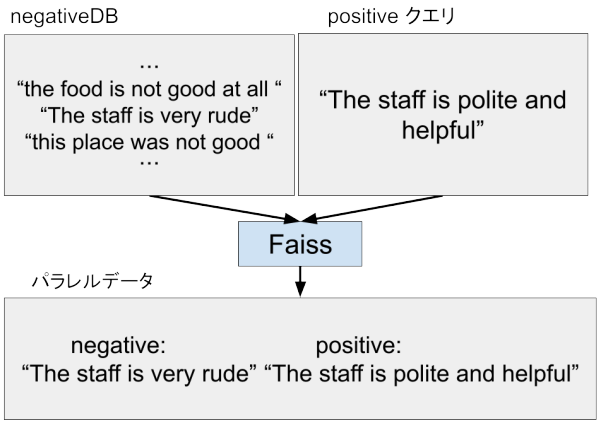

図 2　パラレルデータの作成

表 1　データセット数

	MSCOCO	Yelp	PPF	STSB
ソースデータ数	900,000	(Pos) 267,314 (Neg) 176,787	-	-
Xu らの手法	18,976	18,508	8,349	8,628
本手法	851,246	387,651	8,349	8,628

理も実現している．転置ファイルやグラフベースのインデックスを活用し，大規模データセットにおける効率的な検索が可能となっている．本研究では，Faiss ライブラリを使用し，非パラレルデータからパラレルデータを作成する．

3.3　パラレルデータ作成

Faiss を用いたパラレルデータの作成の概要図を図 2 に示す．ソースデータとして，Shen らが利用した Yelp の非パラレルデータセットを用いた [1]．本データセットは，否定的な文と肯定的な文それぞれ 176,787 文，及び 267,314 文で構成されている．これらを用い Faiss ライブラリにより類似度計算を行い最も意味的に近い文を抽出することにより組を作成した．具体的には，先ず否定的な文をデータベース化し，それを Faiss のインデックスに格納する．その後，肯定的な文をクエリとし，Faiss によりネガティブなテキストデータベースを検索し，意味的に一番近い否定的な文を返すことにより，組を作成する．作成されたパラレルデータの例を，以下に示す．

positive: ”The staff is polite and helpful.”
negative: ”The staff is very rude.”

作成された組の中には，重複した組が存在するため，それらを削除した．さらに組の中には意味に類似していない組も存在するため Xu らと同様，データセット STSB とモデル BERT を使い，意味的類似度を計算し，閾値 3.0 以上の組を抽出した．

4　実験

実験は 2 回行い，実験 1，Yelp データのみを増やした場合と実験 2，MSCOCO と Yelp の両方を増やした場合での実験を行った．

4.1　データセット

実験では，MSCOCO [12], Yelp [1], PPF [9], 及び STSB [13] を用いた．各データ数を表 1 に示す．表 1 において，ソースデータは，先行研究である Xu ら [8] と同じデータセットを使用した．本手法における MSCOCO は，ソースデータから重複を除いたデータである．また，Yelp は Faiss を利用しデータ拡張を行ったデータ数を示す．

4.2　評価指標，及び学習モデル

評価には以下で示す 8 種類の評価尺度を用いた．ROUGE は，生成されたテキストと正解テキストがどれだけ一致しているかを評価する指標であり，正解テキストと比較して共通する 1 単語の割合を示す R-1，共通する連続した 2 単語の割合を示す R-2，テキストの共通する部分の割合を示す R-LCS の 3 つで評価する [14]．BLEU は，意味的なコンテンツ保存を測定するための標準的な指標であり，正解テキストとの比較を基に評価を行う [15]．BERTScore は，正解テキストと比較した際の意味的な類似度を評価する指標である [16]．ΔTB は，入力テキストと生成されたテキストの感情スコアの差を示す [17]．値が大きいほど肯定的な文に変換されていることを意味する．AVG.LEN は，生成された各テキストのトークン数の平均を計算し，生成されたテキスト長を評価する．PPL は，生成されたテキストの流暢さを測定するための指標であり，事前訓練された言語モデル（LM）である GPT-2 を使用し，平均トークンパープレキシティ（PPL）を求めた．学習で用いたモデルは，Bart-base を使用した．

4.3　実験結果

実験結果を表 2 に示す．表 2 において，黒字は各評価尺度において 3 種の手法のうちの最高精度を示す．本手法における ΔTB，及び PPL は最高精度が得られていることから，肯定的な文生成，及び流暢性については Faiss によるデータ拡張手法が有効で

表 2　実験結果

手法		R-1 ↑	R-2 ↑	R-LCS ↑	BLEU ↑	BScore ↑	ΔTB ↑	AVG.LEN ↑	PPL ↓
先行研究 [8]	ST	32.71	13.45	26.98	10.4	89.19	0.220	54.82	106.76
	PG	33.10	**13.81**	**27.34**	**10.6**	89.16	0.220	**57.43**	116.06
	PG2ST	32.87	13.57	27.13	10.2	**89.20**	0.224	49.60	91.85
	ST2PG	**33.17**	13.77	27.31	10.5	**89.20**	0.209	50.86	93.34
本手法　実験 1	ST	31.61	12.25	26.21	9.7	89.13	0.284	41.37	79.05
	PG	32.65	13.03	26.77	9.9	**89.20**	0.259	49.15	74.03
	PG2ST	32.45	12.73	26.66	9.8	89.18	0.272	53.68	98.32
	ST2PG	32.53	12.87	26.70	9.8	89.14	0.264	46.90	76.55
本手法　実験 2	ST	29.21	10.08	24.73	8.0	88.95	0.372	37.46	**41.99**
	PG	28.05	9.75	23.24	7.5	88.80	0.375	42.55	43.02
	PG2ST	29.09	10.14	24.06	7.7	88.93	0.384	37.32	42.45
	ST2PG	28.82	10.29	23.84	8.2	88.87	**0.387**	38.17	49.77

あると言える．一方，R-1, R-2, R-LCS, BLUE, BScore, 及び AVG.LEN はいずれも先行研究の結果が本手法よりも上回っている．これは先行研究と本手法における 4 種の各フローを比較した場合にも同じ傾向であった．原因として，ランダムに抽出した 50 文からなるテストデータを調査したところ，本研究は 4 文に対し，先行研究は 18 文が入力文に表出する内容を示す単語をそのままコピーし生成しているため，フローに関係なく，先行研究における各尺度の精度が本手法よりも高いと考えられる．

4.4　人手による評価

本手法の有効性を検証するため，生成文について (1) 肯定的な文に変換できているか (Style)，(2) 入力文の意味を保持しているか (Content)，及び (3) 流暢な文が生成できているか (Fluency) について，人手による評価を実施した．評価者は，自然言語処理に詳しく，かつ英語に堪能な 3 名により実施した．ST2PG から無作為に 50 文を抽出し，関連研究と本手法に関する評価を行った．評価結果を表 3 に示す．表 3 における各数値は，3 名の評価者の平均値を示す．

表 3 より，Content，及び Fluency 共に先行研究の結果が若干上回っている．これは，表 2 において，本手法よりも先行研究のほうが，表出する内容を示す単語をそのままコピーし生成しているため，先行研究の各尺度の精度が本手法よりも高いという理由と一致している．今後は Style の精度を低下させることなく，多様な表現で生成可能な手法を検討する必要がある．

表 3　人手による評価

	Style ↑	Content ↑	Fluency ↑
先行研究	2.86	4.24	4.50
本手法	3.69	3.67	4.33

5　おわりに

本研究は，テキストの意味を表す内容とスタイルとを分離し，スタイルを変換した後，内容を保持しまま表現を変換する，あるいはその逆を行う先行研究 [8] に対し，Faiss を用いることにより PTR に必要となるパラレルデータを増強し，PTR の精度向上を目指した．実験の結果，先行研究と比較し ΔTB，及び PPL は先行研究よりも高い精度が得られていることから，肯定的な文生成，及び流暢性については Faiss によるデータ拡張手法が有効であることが明らかになった．一方，R-1, R-2, R-LCS, BLUE, BScore, 及び AVG.LEN はいずれも先行研究の結果が本手法よりも上回っていること，人手による評価においても同じ傾向が観察できたことから，今後は肯定的な文生成精度を低下させることなく，多様な表現で生成可能な手法を検討する必要がある．

謝辞

本研究は，鹿島学術振興財団 (2023 共同新 04)，及び JKA(2024M-557) の助成を受けたものです．

参考文献

[1] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017), pp. 6833–6844, 2017.

[2] Shutong Jia, Biwei Cao, Qingqing Gao, Jiuxin Cao, and Bo Liu. Positive text reframing under multi-strategy optimization. In Submitted to ACL Rolling Review - April 2024, 2024. under review.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. Proceedings of the 30th International Conference on Neural Information Processing Systems (NeurIPS 2017), pp. 5998–6008, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), pp. 4171–4186, 2019.

[5] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, pp. 1–67, 2020.

[6] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pp. 79–86, 2002.

[7] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. 2009.

[8] Sheng Xu, Fumiyo Fukumoto, Jiyi Li, Kentaro Go, and Yoshimi Suzuki. Learning disentangled meaning and style representation for positive text reframing. Proceedings of the 16th International Natural Language Generation Conference (INLG 2023), pp. 424–430, 2023.

[9] Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. Inducing positive perspectives with text reframing. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), pp. 3682–3700, 2022.

[10] Lai Huiyuan, Toral Antonio, and Nissim Malvina. Thank you BART! rewarding pre-trained models improves formality style transfer. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021.

[11] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. IEEE Transactions on Big Data, pp. 535–547, 2017.

[12] Bandel Elron, Aharonov Ranit, Shmueli-Scheuer Michal, Shnayderman Ilya, Slonim Noam, and Ein-Dor Liat. Quality controlled paraphrase generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 596–609. Association for Computational Linguistics, 2022.

[13] Cer Daniel, Diab Mona, Agirre Eneko, Lopez-Gazpio I nigo, and Specia Lucia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp. 1–14, Vancouver, Canada, 2017. Association for Computational Linguistics.

[14] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out: Proceedings of the ACL Workshop, pp. 74–81, 2004.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 311–318, 2002.

[16] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. Proceedings of the International Conference on Learning Representations (ICLR 2020), 2020.

[17] Steven Loria. textblob documentation. Release 0.16, 2.