スタイル変換のためのスタイルを考慮した Seq2Seq 事前訓練

松浦 駿平　　梶原 智之　　二宮 崇

愛媛大学大学院理工学研究科

matsu@ai.cs.ehime-u.ac.jp {kajiwara, ninomiya}@cs.ehime-u.ac.jp

概要

特定のタスクやドメインに特化した事前訓練が，近年盛んに研究されている．本研究では，インフォーマルな文をフォーマルに言い換えるスタイル変換に焦点を当て，本タスクに特化した事前訓練の手法を提案する．提案手法では，事前訓練の際にスタイルを学習するために，出力文のスタイルを表現する特殊トークンを入力文の文頭に加える．英語のスタイル変換タスクにおける実験の結果，自動評価と人手評価の両方で提案手法の有効性を確認した．

1 はじめに

近年，様々な自然言語処理タスクにおいて，深層学習に基づくアプローチが主流となっている．高品質な深層学習モデルを得るためには，大規模なラベル付きコーパスやパラレルコーパスを用いた訓練が有効である．しかし，アノテーションのコストのために，多くの自然言語処理タスクは少資源問題に悩まされている．本研究で扱うスタイル変換タスク [1] も，数万文対という小規模なパラレルコーパスしか使用できず，少資源問題が深刻なタスクのひとつである．他のタスクと同様に，スタイル変換タスクにおいても，マルチタスク学習 [2] やデータ拡張 [3] などの少資源対策の手法が研究され，現在では BART [4] などの事前訓練された汎用モデルの再訓練 [5–8] がデファクトスタンダードとなっている．

特定のタスクやドメインに特化した事前訓練の有効性も，広く知られている．例えば，診療記録 [9] や科学技術論文 [10] などのドメインに特化した事前訓練モデルや，生成型要約 [11] や言い換え生成 [12] などのタスクに特化した事前訓練モデルは，対象のタスクやドメインにおいて高い性能が報告されている．しかし，事前訓練の際にテキストのスタイルについて明示的に学習するスタイル変換タスクに特化した事前訓練は，これまで研究されていない．

本研究では，事前訓練の際にスタイルに関する知識を獲得するために，出力文のスタイルを表現する特殊トークンを入力文の文頭に追加した上で，BART と同様のノイズ除去自己符号化の事前訓練を行う．GYAFC コーパス [1] を用いたインフォーマルな英文からフォーマルな英文へのスタイル変換の実験の結果，スタイル・同義性・流暢性の全ての人手評価の項目において，提案手法は通常の BART の事前訓練よりも有効であった．

2 関連研究

2.1 事前訓練の先行研究

大規模な生コーパスを用いて事前訓練された Transformer [13] が，多くの自然言語処理タスクで優れた性能を達成している．言語理解や言語生成の応用タスクにおいて，Transformer エンコーダをマスク言語モデリングのタスクで事前訓練した BERT [14] や Transformer デコーダを言語モデリングのタスクで事前訓練した GPT-2 [15] が，広く使用されている．

本研究で扱う系列変換タスクにおいても，様々な事前訓練モデル [16, 17] が提案されている．スタイル変換タスク [7, 8] では，語句の穴埋めなどのノイズ除去自己符号化のタスクで事前訓練された BART [4] が優れた性能を達成している．本研究では，BART の汎用的な事前訓練をスタイル変換タスクに特化させ，事前訓練の際にスタイルに関する知識を獲得することでスタイル変換の性能改善を目指す．

2.2 スタイル変換の先行研究

シェイクスピアの英文をモダンに変換するスタイル変換 [18] や難解な英文を平易に変換するスタイル変換 [19] など，様々なスタイルを対象とするスタイル変換が研究されている．本研究では，インフォーマルな英文をフォーマルに変換するフォーマルさに関するスタイル変換 [1] に取り組む．

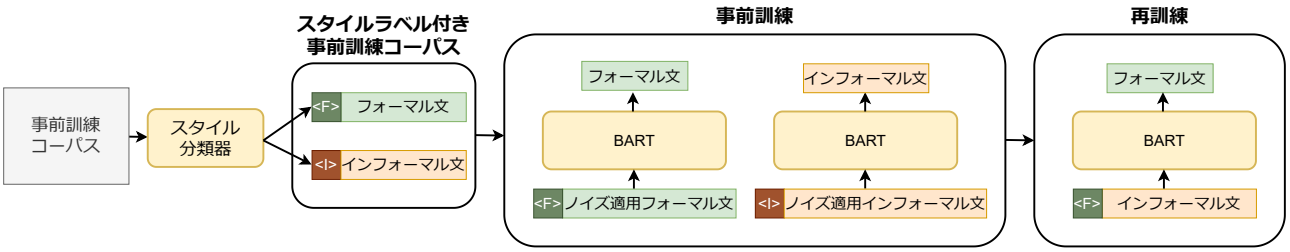

図 1　提案手法

フォーマルさに関するスタイル変換のために使用できる GYAFC [1] は 5 万文対の小規模なパラレルコーパスであるため，多くの少資源対策の手法が提案されてきた．大規模なパラレルコーパスを利用できる機械翻訳とのマルチタスク学習 [2] や，自動生成された擬似的な単言語パラレルコーパスを用いる事前訓練 [3, 12]，GPT-2 [15] の再訓練 [5, 6] などが成功を収め，BART [4] の再訓練 [7, 8] が現在の最高性能を達成している．本研究では，スタイルを明示的に訓練するように BART の事前訓練を改良する．

3　提案手法

本研究では，BART [4] の事前訓練をスタイル変換タスクのために改良する．語句の穴埋めを行う事前訓練の際に，入力文の文頭にスタイルラベルの特殊トークンを付与することにより，出力文のスタイルを制御する．これにより，従来の BART の事前訓練では考慮していないテキストのスタイルに関する知識の獲得を目指す．

提案手法の概要を図 1 に示す．本手法では，まず，各文のスタイルを表現するために，スタイルラベル付きの事前訓練用コーパスを作成する（3.1 節）．次に，作成したコーパスを用いて BART と同様の事前訓練を行う（3.2 節）．最後に，事前訓練で学習したスタイルラベルを利用し，目的のスタイル変換用パラレルコーパス上で再訓練を行う（3.3 節）．

3.1　スタイルラベル付きコーパスの作成

まず，スタイル分類器を用いて，事前訓練に用いる生コーパスの各文をソーススタイルの文とターゲットスタイルの文に分類する．その後，ソーススタイルの確率が高い文とターゲットスタイルの確率が高い文をそれぞれ同じ文数だけ抽出する．最後に，モデルがソーススタイルの文とターゲットスタイルの文を識別できるようにするために，各文の文頭にスタイルを示すラベルを特殊トークン（図 1 の <F>や<I>）として追加する．

スタイル分類器には，フォーマルさのラベル付きコーパスを用いて訓練した分類器 [20] や言語モデルに基づく分類器 [2] など，任意の分類器を適用できる．本研究では，事前訓練された RoBERTa [21] を GYAFC [1] のラベル付きコーパス上で再訓練する．

3.2　事前訓練

3.1 節のコーパスを用いて，Transformer モデル [13] を事前訓練する．事前訓練のタスクには，BART [4] と同様の語句の穴埋めを採用する．つまり，入力文の一部の語句をマスクして入力し，元の文を復元するノイズ除去自己符号化の事前訓練を行う．また，入力文の文頭のスタイルラベルはマスク処理の対象外とし，出力文にはスタイルラベルを含めない．BART と同様に，ミニバッチは約 512 トークンとなるように複数文を連結して構成する．ただし，同じスタイルの文のみでミニバッチを構成し，スタイルラベルはミニバッチの先頭に 1 つだけ付加する．

提案手法の事前訓練では，モデルはソースまたはターゲットの各スタイルラベルが付与された入力文から当該スタイルの文の生成を学習する．この事前訓練により，スタイルラベルによって出力文のスタイルを制御可能なモデルを得られると期待できる．

3.3　再訓練

再訓練には，スタイル変換用のパラレルコーパスを用いる．ただし，事前訓練時に学習したスタイルラベルを引き続き利用するために，パラレルコーパスの入力文の文頭にもスタイルラベルを付加する．

4　評価実験

4.1　実験設定

モデル　スタイル変換器は fairseq[1] [22] を用いて実装し，bart-base[2] [4] と同じ構造の Transformer [13]

1）https://github.com/pytorch/fairseq
2）https://github.com/pytorch/fairseq/tree/main/examples/bart

表 1　GYAFC コーパスの文対数

	訓練用	検証用	評価用
E&M ドメイン	52,595	2,877	1,416
F&R ドメイン	51,967	2,788	1,432

を採用した．スタイル分類器は huggingface[3] [23] を用いて実装し，事前訓練済みの RoBERTa[4] [21] を採用した．スタイル分類器の訓練には，バッチサイズを 32 文，学習率を 1e-5 とし，GYAFC コーパス [1] のフォーマル文とインフォーマル文を 2 値分類するようにクロスエントロピー損失最小化の訓練を 5 エポック行った．GYAFC の検証用データで分類器の性能を評価したところ，88% の正解率を持つ高品質なスタイル分類器を構築できた．

事前訓練　事前訓練のための生コーパスには，CC100 [24] の英語コーパスを使用した．コーパス全体にスタイル分類器を適用し，フォーマルな確率およびインフォーマルな確率の高い順に 5 千万文ずつを抽出した合計 1 億文を用いて事前訓練を行った．なお，これらの文はフォーマル確率またはインフォーマル確率が 0.9 を超えており，充分にフォーマルまたは充分にインフォーマルな文である．スタイルラベルとして，フォーマルな文の文頭には<F>，インフォーマルな文の文頭には<I>の特殊トークンをそれぞれ追加した．前処理として，Moses[5] [25] によって normalize および tokenize を行い，語彙サイズ 3 万でサブワード分割[6] [26] を行った．

事前訓練には BART [4] と同様の語句の穴埋めタスクを採用し，マスク確率およびポアソン分布の λ は 0.35 とした．バッチサイズは 512 文，Dropout 率は 0.1 とし，エンコーダとデコーダで埋込層を共有した．最適化には Adam を用い，ラベル平滑化クロスエントロピー損失最小化の訓練を 50 万ステップ行った．なお，ラベル平滑化のハイパーパラメータは $\epsilon = 0.2$ とした．学習率スケジュールには Inverse Square Root Decay を用い，最大の学習率は 5e-4，Warmup ステップは 1 万とした．

再訓練　再訓練のためのパラレルコーパスには GYAFC [1] を使用し，インフォーマルな英文をフォーマルに言い換えるスタイル変換を行った．GYAFC には，表 1 に示すように Entertainment & Music（E&M）および Family & Relationships（F&R）の 2 つのドメインのパラレルコーパスが含まれる．先行研究 [2,7,27] と同様に，2 つのドメインの訓練用データを合わせて両ドメインの訓練に用いた．前処理として，Moses によって normalize および tokenize を行い，事前訓練と同じサブワード分割を行った．

表 2　BLEU による自動評価

	E&M	F&R
Transformer	70.7	74.7
BART	72.3	76.4
提案手法	**73.5**	**77.1**

再訓練では，バッチサイズは 1,024 トークン，Dropout 率は 0.1 とし，エンコーダとデコーダで埋込層を共有した．最適化には Adam を用い，検証用データにおける perplexity が 5 エポック改善されなくなるまでラベル平滑化クロスエントロピー損失最小化の訓練を行った．なお，ラベル平滑化のハイパーパラメータは $\epsilon = 0.1$ とした．学習率スケジュールには Inverse Square Root Decay を用い，最大の学習率は 3e-5，Warmup ステップは 500 とした．

推論　モデルの評価時には，ビーム幅 5 のビーム探索を行った．また，シード値による影響を緩和するために，シード値の異なる 4 つのモデルをアンサンブルして出力文を生成した．

比較手法　スタイルラベルを用いる事前訓練の有効性を検証するために，事前訓練を行わない Transformer およびスタイルラベルを用いない BART と比較した．Transformer ベースラインは，提案手法と同じモデル構造であるが，GYAFC のパラレルコーパスのみを用いて訓練する．BART ベースラインは，提案手法と同じモデル構造であり，同じコーパスを用いて事前訓練および再訓練を行うが，事前訓練の際にスタイルラベルを使用しない．

評価　自動評価と人手評価の両方で各モデルの性能を評価した．自動評価には SacreBLEU[7] [28] を用いた．人手評価には Amazon Mechanical Turk[8] を使用し，米国在住，98% の承認率，5,000 件以上のタスク承認歴を持つアノテータを 5 人雇用[9] した．評価用データの中からドメインごとに 100 文の合計 200 文を無作為抽出し，入出力間の同義性・出力文の流暢性・出力文のスタイルのフォーマルさの 3 つの観

3）https://github.com/huggingface/transformers
4）https://huggingface.co/roberta-base
5）https://github.com/moses-smt/mosesdecoder
6）https://github.com/rsennrich/subword-nmt
7）https://github.com/mjpost/sacrebleu
8）https://www.mturk.com/
9）時給が約 8.5 ドルになるように報酬を支払った．

表 3　出力例（赤字はインフォーマルな表現、青字はフォーマルな表現）

入力文	It's a good idea to always be honest besides not many guys I know don't like being told they're hot!
Transformer	It is a good idea to always be honest, besides not many men I know that I do not like being attractive.
BART	It is a good idea to always be honest, besides not many guys I know do not like being told they are hot.
提案手法	It is a good idea to always be honest. Besides, not many men I know do not like being told they are attractive.

表 4　人手評価の平均値

	同義性	流暢性	スタイル
Transformer	5.12	4.29	1.63
BART	5.18	4.45	1.94
提案手法	**5.24**	**4.46**	**1.98**

点から人手評価を行った．先行研究 [7] に従い，同義性を [1, 6] の 6 段階，流暢性を [1, 5] の 5 段階，スタイルを [−3, 3] の 7 段階で人手評価した．

4.2　実験結果

表 2 に自動評価の結果を示す．事前訓練を行わない Transformer ベースラインと比較して，BART ベースラインおよび提案手法が顕著に高い性能を示した．また，提案手法は BART と比べて統計的に有意（$p < 0.05$）に BLEU が向上したことから，スタイルを考慮する事前訓練の有効性を確認できた．

表 4 に人手評価の結果を示す．同義性・流暢性・スタイルの全ての指標において，提案手法が比較手法たちを上回る評価を得た．これらの実験結果から，事前訓練の際に明示的にテキストのスタイルを考慮する提案手法の有効性が，自動評価と人手評価の両方で明らかになった．

5　分析

5.1　小規模なパラレルコーパスでの再訓練

より小規模なパラレルコーパスしか使用できない設定における提案手法の有効性を検証する．4 節の実験では，GYAFC に含まれる 2 つのドメインの訓練用データを合わせた約 10 万文対を用いて再訓練を行ったが，本節ではこれを 5 万文対および 1 万文対に無作為抽出して削減する．4 節の実験と同様に，シード値の異なる 4 つのモデルをアンサンブルして出力文を生成し，BLEU による自動評価を行う．

図 2 に，再訓練に用いる文対数を削減した際の性能の変化を示す．再訓練用のパラレルコーパスを 5 万文対や 1 万文対に削減した場合も，提案手法は一貫して Transformer および BART のベースライン性能を上回る．また，10 万文対のデータで再訓練した BART ベースラインと比較して，5 万文対のデータで再訓練した提案手法は両ドメインにおいて 0.1 ポイント高い BLEU を達成しており，提案手法が少資源の設定においても有効であることを確認できた．

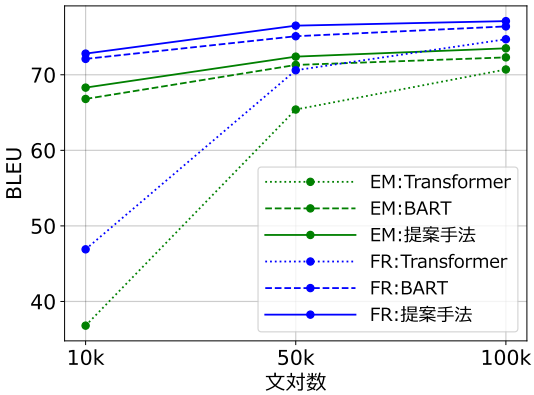

図 2　小規模なパラレルコーパスによる再訓練

5.2　生成文の定性評価

表 3 に各モデルの出力例を示す．Transformer は入力文と意味の異なる文を生成しているが，BART と提案手法は入力文の言い換えの生成に成功している．また，提案手法は BART とは異なり，guys を men，hot を attractive へと言い換えており，表 4 で示したとおり，よりフォーマルな文を生成している．

6　おわりに

本研究では，事前訓練の際にテキストのスタイルを明示的に学習するために，出力文のスタイルを表現する特殊トークンを文頭に追加する事前訓練の手法を提案した．そして，英語のフォーマルさに関するスタイル変換での実験を通して，提案手法の有効性を自動評価と人手評価の両方で示した．今後の課題として，他のスタイルや言語にも適用したい．

謝辞

本研究は JST（ACT-X，課題番号：JPMJAX1907）の支援を受けたものです．

参考文献

[1] Sudha Rao and Joel Tetreault. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In **Proc. of NACCL**, pp. 129–140, 2018.

[2] Xing Niu, Sudha Rao, and Marine Carpuat. Multi-Task Neural Models for Translating Between Styles Within and Across Languages. In **Proc. of COLING**, pp. 1008–1021, 2018.

[3] Yi Zhang, Tao Ge, and Xu Sun. Parallel Data Augmentation for Formality Style Transfer. In **Proc. of ACL**, pp. 3221–3228, 2020.

[4] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In **Proc. of ACL**, pp. 7871–7880, 2020.

[5] Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. Harnessing Pre-Trained Neural Networks with Rules for Formality Style Transfer. In **Proc. of EMNLP**, pp. 3573–3578, 2019.

[6] Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and WenHan Chao. Formality Style Transfer with Shared Latent Space. In **Proc. of COLING**, pp. 2236–2249, 2020.

[7] Kunal Chawla and Diyi Yang. Semi-supervised Formality Style Transfer using Language Model Discriminator and Mutual Information Maximization. **Findings of EMNLP**, pp. 2340–2354, 2020.

[8] Huiyuan Lai, Antonio Toral, and Malvina Nissim. Thank you BART! Rewarding Pre-Trained Models Improves Formality Style Transfer. In **Proc. of ACL**, pp. 484–494, 2021.

[9] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly Available Clinical BERT Embeddings. In **Proc. of Clinical NLP Workshop**, pp. 72–78, 2019.

[10] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. In **Proc. of EMNLP-IJCNLP**, pp. 3615–3620, 2019.

[11] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In **Proc. of ICML**, pp. 11328–11339, 2020.

[12] Tomoyuki Kajiwara, Biwa Miura, and Yuki Arase. Monolingual Transfer Learning via Bilingual Translators for Style-Sensitive Paraphrase Generation. In **Proc. of AAAI**, pp. 8042–8049, 2020.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Proc. of NIPS**, 2017.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proc. of NACCL**, pp. 4171–4186, 2019.

[15] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[16] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In **Proc. of ICML**, pp. 5926–5936, 2019.

[17] 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**, Vol. 21, No. 140, pp. 1–67, 2020.

[18] Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. Paraphrasing for Style. In **Proc. of COLING**, pp. 2899–2914, 2012.

[19] Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. Data-Driven Sentence Simplification: Survey and Benchmark. **Computational Linguistics**, Vol. 46, No. 1, pp. 135–187, 2020.

[20] Ellie Pavlick and Joel Tetreault. An Empirical Analysis of Formality in Online Communication. **Transactions of the Association for Computational Linguistics**, Vol. 4, pp. 61–74, 2016.

[21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **arXiv:1907.11692**, 2019.

[22] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In **Proc. of NAACL**, pp. 48–53, 2019.

[23] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In **Proc. of EMNLP**, pp. 38–45, 2020.

[24] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In **Proc. of LREC**, pp. 4003–4012, 2020.

[25] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In **Proc. of ACL**, pp. 177–180, 2007.

[26] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In **Proc. of ACL**, pp. 1715–1725, 2016.

[27] Tomoyuki Kajiwara. Negative Lexically Constrained Decoding for Paraphrase Generation. In **Proc. of ACL**, pp. 6047–6052, 2019.

[28] Matt Post. A Call for Clarity in Reporting BLEU Scores. In **Proc. of WMT**, pp. 186–191, 2018.