中間文生成によるスタイル変換のためのパラレルデータ拡張

大澤功記　　ルパージュ・イヴ
早稲田大学 大学院情報生産システム研究科
koki.osawa@akane.waseda.jp　yves.lepage@waseda.jp

概要

自然言語処理におけるスタイル変換とは，文書を特定のスタイルを持つ別の文書に意味を保ちながら変換するタスクである．近年の研究では，教師あり学習を活用したスタイル変換が一般的であるが，データセットが不足している．本研究では，特定のスタイルに依存しない中間文の生成によるパラレルデータ拡張法を提案し，その有効性及び一般性を検証した．実験では，Grammarly Yahoo Answers Formality Corpus と FlickrStyle10K に対してデータ拡張を行い，元データ及び拡張済みデータでスタイル変換モデルを学習し，BLEU による評価と統計的有意性の検証を行った．実験より，平均編集距離が小さい文対において提案法の有効性が示された．

1　背景

自然言語処理におけるスタイル変換とは，与えられた文書を特定のスタイルを持つ別の文書に意味を保ちながら変換するタスクである．スタイルは，個人や集団の持つ意味を表現する方法という直感的な概念であるとされ (Mcdonald et al., 1985) [1]，例として，丁寧さ (formality) や平易さ (simplicity) がある．

近年の研究では，深層学習を活用したスタイル変換が一般的であるが，利用可能なパラレルデータが少ないためデータセットの不足が問題となっている．深層学習によるスタイル変換は，主に教師あり学習と教師なし学習からなり，教師あり学習によるスタイル変換では，ニューラル機械翻訳に使用されていた時系列モデルが一般的に活用され (Niu et al., 2018; Xu et al., 2012) [2] [3]，これを踏まえてマルチタスキングやデータ拡張の研究が行われている．教師なし学習によるスタイル変換では，強化学習や敵対的生成ネットワークが活用されているが (Gong et al., 2019; Yang et al., 2018) [4] [5]，現在でも教師あり学習がより高いスコアを上げている．Zhang ら (2020) [6] は，丁寧さのスタイル変換における機械翻訳モデルの逆翻訳を活用したパラレルデータ拡張法を提案した．当手法では，機械翻訳モデルが訳として丁寧な文章を出力することを利用する．パラレルでないカジュアルな文に対して機械翻訳による逆翻訳を行い，対応する丁寧な文を取得する．実験により，丁寧さのスタイル変換におけるデータ変換において最先端の結果を得たことが示された．しかし，データ拡張のためのモデルの学習コストが高いことや，丁寧さという特定のスタイルにしか手法を適用できないことが問題点として挙げられる．

本研究では，特定のスタイルに依存しない大規模な拡張文対の生成が可能なパラレルデータ拡張法を提案する．提案法は，中間文生成モデルを用いたパラレルデータ拡張法であり，スタイル変換におけるパラレルデータにおいて対応する文章が類似した意味を持つことを利用する．実験では，Grammarly Yahoo Answer Formality Corpus (GYAFC) (Rao et al., 2018) [7] と FlickrStyle10K (Chuang et al., 2017) [8] に対してデータ拡張を適用し，元データのみ及び拡張済みデータで学習させたスタイル変換モデルを BLEU によって比較評価することで提案法の有効性を検証した．結果として，適切な数の拡張文対を加えた元データにより学習されたモデルで BLEU が向上し，統計的有意性が示された．また，考察より中間文生成に用いる文対のペアの平均編集距離が小さい場合に有効な拡張文対が生成されることがわかり，平均編集距離が 2 以下の文対において提案法が特に有効であることが示された．

2　先行研究

2.1　テキストモーフィング

テキストモーフィング (Huang et al., 2018) [9] とは，2 入力文より意味的にそれらの間に存在する文を生成するタスクである．このタスクは，文章の意味を制御した滑らかな変化を伴う文章の生成を目的とする．テキストモーフィングの例を表 1 に示す．

表 1 テキストモーフィングの例 ([9] より引用)

S_{start}:	The noodles and pork belly was my favourite .
S_1:	The pork belly was my favourite .
S_2:	The pork was very good .
S_3:	The stuff was very good .
S_3:	The stuff is very friendly .
S_{end}:	Love how friendly the stuff is !

2.2 中間文生成によるモーフィング

Wang ら (2019) [10] は，中間文の生成によるテキストモーフィングを提案した．ここで中間文とは，2 入力文に対して中間の意味を持つ文章を意味する．midgenerator はその中間文を生成するモデルであり，学習済みのオートエンコーダから生成された中間文を用いて微調整された GPT-2 (Radford et al., 2019) [11] である．当論文では，中間文の評価のために BERTScore (Zhang et al., 2020) [12] を用いた BERTScore Distance が定義された．BERTScore Distance は式 1 により計算される．

$$d_B(A, B) = 1 - BERTScore(A, B) \quad (1)$$

また，微調整にはパープレキシティが 30 以下かつ GECToR (Omelianchuk et al., 2020) [13] よって訂正されなかったもので式 2 を満たす 5000 の開始文，終端文及び中間文が使用された．

$$|d_B(S_{start}, S_{middle}) - d_B(S_{middle}, S_{end})| < 0.05 \quad (2)$$

実験により，従来手法と比較して当手法が意味を考慮したより直線的な中間文生成を行えることが示された．直線的とは埋め込み空間上で中間文が開始文と終端文がなす直線に近いことを意味する．

3 提案手法

3.1 提案手法

本研究では，スタイル変換における新たなパラレルデータ拡張法として，中間文の生成を活用する方法を提案する．提案手法の概要を図 1 に示す．

提案手法は，以下の 3 つの処理を行いパラレルデータを拡張する．

1. 一方のスタイルから 2 文を取得してそれらの中間文を生成する
2. もう一方のスタイルから対応する 2 文を取得してそれらの中間文を生成する
3. 生成された 2 つの中間文を対応付けて拡張データとする

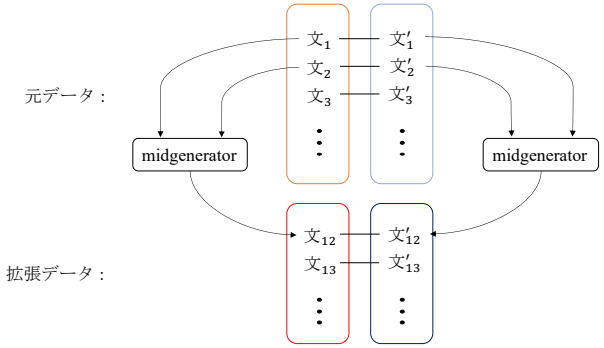

図 1 提案手法の概要

本手法は，データセットから文対のペアを取得し，中間文の生成によりデータ拡張を行うが，n 個の文対を持つパラレルデータにおいて取得される文対のペアは ${}_nC_2$ であり，元データに対して 2 次に比例する量の拡張データが取得可能である．

3.2 提案手法における類推関係

本手法は，スタイル変換のパラレルデータが対応する文対が類似する意味を持つことを利用する．以下の式に提案手法における類推関係を示す．

$$文_i \leftrightarrow 文'_i \quad (3)$$

$$文_i : 文_{ij} : 文_j \leftrightarrow 文'_i : 文'_{ij} : 文'_j \quad (4)$$

$$文_{ij} \leftrightarrow 文'_{ij} \quad (5)$$

ここで，$文_i$ と $文'_i$ が与えられた対応する文対であり，$文_{ij}$ は $文_i$ と $文_j$ の，$文'_{ij}$ は $文'_i$ と $文'_j$ の中間文を表す．このとき，$文_i$ と $文'_i$ と $文_j$ と $文'_j$ は類似した意味を持つため，それぞれの意味的な中間文である $文_{ij}$ と $文'_{ij}$ は類似する意味を持つ．

3.3 新規性・有効性

本手法の新規性は，パラレルデータ拡張に類推関係を用いることに加え，文書生成モデルが生成した文でデータ拡張を行うことである．また，有効性として学習コストが低いこと，特定のスタイルに依存しない方法であることと二次に比例する数の拡張文対が取得できることが挙げられる．

4 実験

4.1 データセット

実験には，GYAFC (Rao et al., 2018) [7] と FlickrStyle10k (Chuang et al., 2017) [14] を使用した．

GYAFC　GYAFC は，1 文あたりの単語数が 5 語から 25 語までの意味が対応するカジュアルな文と丁寧な文の対からなる，丁寧さのスタイル変換のためのパラレルコーパスである．このコーパスは，データセットは，エンターテインメントと音楽 (Entertainment&Music) 及び家族と人間関係 (Family&Relationship) の二つのカテゴリからなる．なお，このデータセットは，学習用データ，検証データとテストデータに予め分けられている (Rao et al., 2018)．GYAFC に含まれる文対の例を表 2 に，GYAFC の各カテゴリの文対の数を表 3 に示す．

表 2　GYAFC の EM カテゴリに含まれる文対の実例

カジュアル	Hope that helps or am I entirely off here?
丁寧	I am entirely off or does that help.

表 3　GYAFC の各カテゴリの文対の数

カテゴリ	E&M	F&R
学習データ	52,595	51,967
検証データ	2,877	2,788
テストデータ	1,416	1,332

FlickrStyle10K　FlickrStyle10k (Chuang et al., 2017) [14] は，Flickr30K (Hodosh et al., 2013) [8] を基に構築された，スタイルを明示的に制御したキャプション生成モデルによるデータセットである．このデータセットには，合計で 10,000 枚の画像が含まれており，各画像にはロマンチックなキャプションとユーモラスなキャプション対が付属する．現在では，7,000 枚分のみが利用可能である．Li ら (2018) [15] は，スタイル変換に本データセットを使用することを提案した．Guo ら (2019) [16] に従って，本研究の実験では，6,000 の文対を学習，500 の文対を検証，500 の文対をテストに使用した．FlickrStyle10K に含まれる文対の例を表 4 に示す．

表 4　FlickrStyle10k に含まれる文対の実例

ロマンチック	A man uses rock climbing to conquer the high.
ユーモラス	A man is climbing the rock like a lizard.

4.2　実験方法

実験は，データ拡張部と評価部から構成される．

データ拡張部　元データから取得した全ての文対の組合せのうち，単語レベルの平均編集距離が小さい組合せに対して中間文を生成し，拡張文対を取得した．実験のデータ拡張部の概要を図 2 に示す．

なお，生成する中間文対の数は使用するデータの規模を基準に設定し，GYAFC の場合は 50,000，Flickr の場合は 5,000 の中間文対を拡張した．

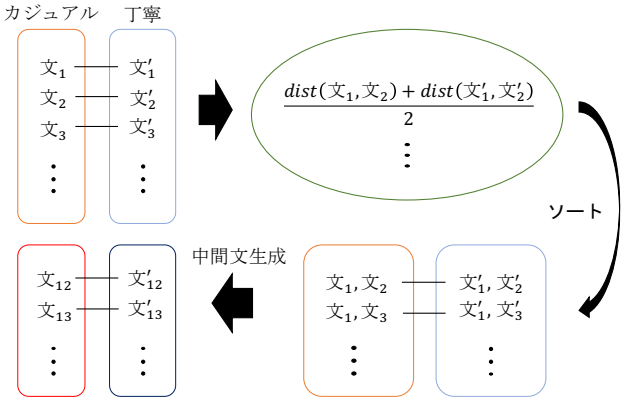

図 2　データ拡張部の概要

評価部　元データに一定量の拡張データを加えた複数の場合でスタイル変換モデルを学習し，BLEU (Paninenі et al., 2002) [17] による評価を行なった．

GYAFC では，元データのみ及び 10,000，20,000，30,000，50,000 の拡張文対を元データに加えた場合で Transformer (Vaswani et al., 2017) [18] にカジュアルから丁寧へのスタイル変換を学習させた．なお，モデルは 8 つの注意ヘッドと 4 層の 512 次元の隠れ層を持ち，512 次元の単語ベクトルにより 32 の文対をバッチとして 0.0005 の学習率で学習された．

FlickrStyle10K では元データのみ及び 1,000，2,000，3,000，5,000 の拡張文対を元データに加えた場合で GRU (Chung et al., 2014) [19] にロマンチックからユーモラスへのスタイル変換を学習させた．なお，モデルは 2 層の 512 次元の隠れ層を持ち，128 次元の単語ベクトルにより 8 の文対をバッチとして 0.00005 の学習率で学習された．

なお，モデルは OpenNMT-py[1] (Klein et al., 2017) [20] 上のものを用いた．また，BLEU の測定には Sacrebleu[2] (Post et al., 2018) [21] を使用し，MosesDecoder[3] (Koehn et al., 2007) [22] 上のツールによりモデルの統計的有意性を検証した．

5　実験結果と考察

5.1　実験結果

GYAFC の各カテゴリにおけるデータ拡張前及び拡張後の BLEU を表 5 に，FlickrStyle10K におけるデータ拡張前及び拡張後の BLEU を表 6 に示す．なお，有意性の検証により，$p < 0.05$ となり有意と認められたモデルを以下の表において太字で示す．

1）https://github.com/OpenNMT/OpenNMT-py
2）https://github.com/mjpost/sacrebleu
3）https://github.com/moses-smt/mosesdecoder

表 5 GYAFC における拡張前及び拡張後の BLEU

拡張文対数	0	10,000	20,000	30,000	50,000
E&M	24.1	24.5	24.9	24.2	23.7
F&R	30.7	30.6	**32.0**	30.1	31.0

表 6 FlickrStyle10K における拡張前及び拡張後の BLEU

拡張文対数	0	1,000	2,000	3,000	5,000
FlickerStyle10K	5.6	6.1	4.8	4.4	4.5

実験より，GYAFC の両カテゴリにおいて BLEU が向上し，提案手法が有効であることが示された．しかし，拡張文対を増やしすぎた場合には BLEU が低下した．FlickrStyle10K でも同様に BLEU が向上したが，拡張文対を増やしすぎた場合には BLEU が低下した．また，有意性の検証では，F&R カテゴリの元データと 20,000 の拡張文対で学習したモデルが $p < 0.05$ となり有意性が示された．

5.2 考察

BLEU が低下した要因として，拡張文対数が増えたことにより開始文対と終端文対の平均編集距離が大きくなり，生成された中間文の意味が対応しなかったことが考えられる．GYAFC の E&M カテゴリに含まれる文対の組合せと，平均編集距離ごとに生成された中間文対の例を表 7, 8 に示す．

表より，平均編集距離が大きい中間文対が小さい中間文対に比べて意味が対応していないと直感的に評価できる．生成された中間文対の定量的評価のために，GYAFC の E&M カテゴリにおける実験で生成した中間文対に対して，SentenceBERT (Reimers et al., 2019) [23] を用いて意味を考慮した文章ベクトルを取得し，平均編集距離ごとにコサイン類似度を算出した．編集距離ごとの類似度を図 3 に示す．

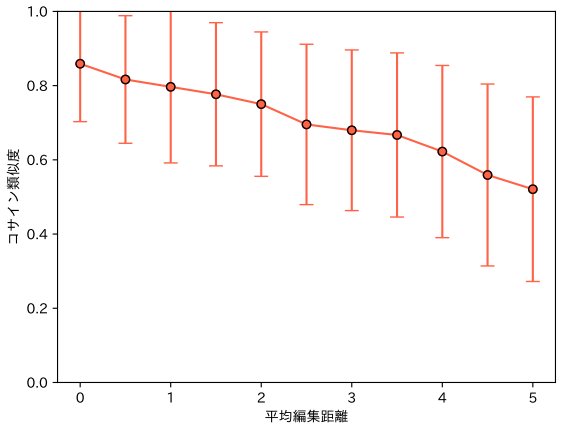

図 3 各平均編集距離における類似度

表 7 GYAFC の E&M カテゴリにおいて平均編集距離が 5 の文対の組合せから生成された中間文対の実例

カジュアル	$文_i$	verbal abuse is just as bad.
	$文_{ij}$	she is just as bad.
	$文_j$	if so she is shallow.
丁寧	$文'_i$	Verbal abuse is bad.
	$文'_{ij}$	She is bad.
	$文'_j$	She is very shallow.

表 8 GYAFC の E&M カテゴリにおいて平均編集距離が 20 の文対の組合せから生成された中間文対の実例

カジュアル	$文_i$	Sure, it's ok, but I always have let the guy ask me.
	$文_{ij}$	you can talk to her but not the guy!
	$文_j$	sit down tell her you need to talk and no yelling mom and explain it to her!!
丁寧	$文'_i$	I prefer to let the guy ask me.
	$文'_{ij}$	I can talk to her to explain it.
	$文'_j$	Sit down and tell her you need to talk, no yelling, Mom, and explain it to her!

図より，中間文生成に使用する開始文対と終端文対の平均編集距離が近い場合において生成された中間文対の意味が類似することが示された．また，平均編集距離が大きくなるにつれて中間文対の類似度の平均が下がり，分散が大きくなることがわかった．GYAFC の E&M カテゴリの元データでは，対応する文対のコサイン類似度の平均値は 0.787 となったため，平均編集距離が 1.5 程度の文対ペアで提案法が有効だと考えられる．

6 おわりに

本研究では，特定のスタイルに依存しない大規模なデータ拡張法として，中間文生成を活用したパラレルコーパス拡張法の提案を行った．

実験により，GYAFC では適切な数の拡張文を加えた元データにより学習されたモデルで BLEU が向上し，拡張文対が多すぎる場合には BLEU が低下した．FlickrStyle10K でも同様に BLEU は向上したが，拡張文対が多すぎる場合には BLEU が低下した．また，有意性の検証では GYAFC の F&R カテゴリにおいて 20,000 の拡張文対を元データに加えた場合で有意性が示された．考察より，中間文生成に用いる文対のペアの平均編集距離が小さい場合に有効な拡張文対が生成されることがわかり，平均編集距離が 2 以下の文対において提案法が特に有効であることが示された．今後の展望として，拡張文対のより良い抽出法の検討や機械翻訳での応用が考えられる．

参考文献

[1] David D. McDonald and James D. Pustejovsky. A computational theory of prose style for natural language generation. In **Second Conference of the European Chapter of the Association for Computational Linguistics**, Geneva, Switzerland, March 1985. Association for Computational Linguistics.

[2] Xing Niu, Sudha Rao, and Marine Carpuat. Multi-task neural models for translating between styles within and across languages. In **Proceedings of the 27th International Conference on Computational Linguistics**, pp. 1008–1021, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[3] Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. Paraphrasing for style. In **Proceedings of COLING 2012**, pp. 2899–2914, Mumbai, India, December 2012. The COLING 2012 Organizing Committee.

[4] Hongyu Gong, S. Bhat, Lingfei Wu, Jinjun Xiong, and Wen mei W. Hwu. Reinforcement learning based text style transfer without parallel training corpus. In **NAACL**, 2019.

[5] Zichao Yang, Zhiting Hu, Chris Dyer, Eric P. Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. In **NeurIPS**, 2018.

[6] Yi Zhang, Tao Ge, and Xu Sun. Parallel data augmentation for formality style transfer. In **Proceedings of ACL 2020**, pp. 3221–3228, Online, July 2020. Association for Computational Linguistics.

[7] Sudha Rao and Joel Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 129–140, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[8] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. **Journal of Artificial Intelligence Research**, Vol. 47, pp. 853–899, 2013.

[9] Shaohan Huang, Yuehua Wu, Furu Wei, and M. Zhou. Text morphing. **ArXiv**, Vol. abs/1810.00341, , 2018.

[10] Pengjie WANG, Liyan WANG, and Yves LEPAGE. Generating the middle sentence of two sentences using pretrained models: a first step for text morphing. 言語処理学会第 27 回年次大会発表論文集, p. 1481–1485, 2021.

[11] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.

[12] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, 2020.

[13] Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR – grammatical error correction: Tag, not rewrite. In **Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 163–170, Seattle, WA, USA → Online, July 2020. Association for Computational Linguistics.

[14] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. Stylenet: Generating attractive visual captions with styles. **2017 IEEE Conference on CVPR**, pp. 955–964, 2017.

[15] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 1865–1874, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[16] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. Mscap: Multi-style image captioning with unpaired stylized text. In **2019 IEEE/CVF Conference on CVPR**, pp. 4199–4208, 2019.

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In **Proceedings of ACL 2002**, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[18] 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 I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, **Advances in NIPS 2017**, Vol. 30. Curran Associates, Inc., 2017.

[19] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In **NIPS 2014 Workshop on Deep Learning**, p. no page number, 2014.

[20] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In **Proceedings of ACL 2017, System Demonstrations**, pp. 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[21] Matt Post. A call for clarity in reporting BLEU scores. In **Proceedings of the Third Conference on Machine Translation: Research Papers**, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics.

[22] 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 **Proceedings of ACL 2007**, pp. 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[23] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In **Proceedings of EMNLP-IJCNLP 2019**, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.