Interspeech 2022
18-22 September 2022, Incheon, Korea

Spoken-Text-Style Transfer with Conditional Variational Autoencoder and Content Word Storage

Daiki Yoshioka[1], *Yusuke Yasuda*[1], *Noriyuki Matsunaga*[2], *Yamato Ohtani*[2], *Tomoki Toda*[1]

[1]Nagoya University, Japan
[2]AI Inc., Japan

yoshioka.daiki@g.sp.m.is.nagoya-u.ac.jp

Abstract

Text style transfer is the task of converting textual style while preserving content. Content preservation is still challenging in text style transfer under the training condition with non-parallel data. We improve the content preservation performance of text style transfer using a labeled non-parallel corpus, targeting interest styles for text-to-speech synthesis. We propose a content word storage mechanism to preserve "content words", particularly for improving content preservation, and incorporate it in the conditional variational autoencoder to capture the style information from the labeled non-parallel corpus. We have conducted a bi-directional transfer experiment of Japanese texts about "disfluency removal/insertion" and "standard/Kansai dialect conversion" as target styles. From the results of automatic and human evaluations, we found that 1) the proposed method improved the content preservation without compromising other performances and 2) the proposed method had different performances depending on the direction of style transfer.

Index Terms: spoken style, spoken text style, text style transfer, text style conversion, dialect, disfluency

1. Introduction

Speech communication is an essential part of human social life. In recent years, there have been increasing opportunities for technology-assisted speech communication with the advent of voice assistants. This trend has been driving active research on text-to-speech (TTS) synthesis, and its quality of synthetic speech is now as natural as human speech [1]. TTS research has been focusing not only on sound quality but also on rendering "spoken style", which defines how to speak such as emotional representation and persona characteristics, as the next challenge [2, 3]. What TTS research has been mainly working on regarding spoken style is "paralinguistic information" such as intonation, inflection, speaking speed, and voice pitch. On the other hand, spoken style is also strongly dependent on the linguistic information, in other words, the text itself that is uttered. Hence, we believe that it is beneficial to control the style of texts in coordination with the style of speech to realize more advanced style control. There is a study focusing on spoken-text-style conversion to improve the readability of automatically transcribed speech by disfluency deletion [4], but more advanced spoken style conversion involving word insertion and replace is relatively unexplored, which must be remedied to enable its application to TTS.

The task of "text style transfer", in which only the style of a text is transformed to another target style while preserving its meaning, has attracted considerable attention [5]. Using rule-based methods [6] and deep learning methods with parallel corpora [7, 8], it is relatively easy to give a desired style to a text. However, much manual work is required to create a large number of conversion rules and a large parallel corpus. This is very costly in terms of both time and labor, and is impractical for application to various scenarios. In contrast, "text with a specific style but no pair data" is relatively easy to obtain from various media. Therefore, we can create a non-parallel corpus that can be constructed using these data at a lower cost than the creation of a parallel corpus.

To achieve text style transfer without much time and effort, a method that can be learned using a non-parallel corpus is required. The most popular text style transfer methods for non-parallel corpora are those that use adversarial learning [9]. In particular, several research groups [10, 11, 12] have improved the performance by introducing cyclic loss functions [13]. In addition, autoencoder (AE) and variational autoencoder (VAE)-[14] based methods, which can provide disentangled latent representation, have been proposed [15, 16]. The VAE-based methods model the content and style of text separately, and convert texts by manipulating only the learned style. These models realize style transfer without a parallel corpus by learning reconstruction rather than conversion.

Recently, many research groups have been looking for ways to effectively convert styles. In contrast, in terms of content preservation, even state-of-the-art methods [17] using the large pre-trained model T5 [18] have not performed well. Therefore, there is significant room in research to improve content preservation. The purpose of our study is to improve content preservation in text style transfer methods using labeled non-parallel corpora, targeting interesting styles for application to TTS. We propose a method that combines conditional VAE with a content word preservation mechanism to improve the performance of content preservation during style transfer. The main contributions of this paper are summarized as follows:

- We propose content word storage to improve the content preservation of the text style transfer framework using non-parallel corpora.
- The proposed method improves content preservation significantly in both disfluency and dialect transfer, as determined by automatic and human evaluation.

2. CVAE-based Text Style Transfer

Conditional VAE (CVAE) is one of the probabilistic models used for text style transfer [15]. Figure 1 shows a schematic diagram of CVAE. CVAE models textual contents and styles using latent variable $\boldsymbol{z}$ and class label y. Latent variable $\boldsymbol{z}$ captures textual content in a compressed form of input text $\boldsymbol{x}_{1:M}$, which is an M-length word sequence $\boldsymbol{x}_{1:M} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_M)$. The class label indicates the explicit style of utterances such as idiolect, dialect, or disfluency level. The text style transfer using CVAE can be conducted by encoding texts with a source style

10.21437/Interspeech.2022-10118

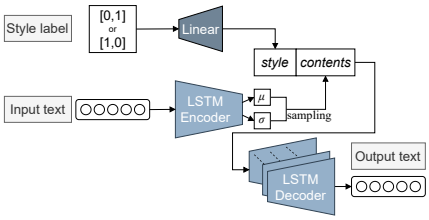

Figure 1: *schematic diagram of CVAE.*

$s \in y$ while conditioning a model with a class label of target style $t \in y$ to decode output text $\hat{\boldsymbol{x}}_{1:N}$.

CVAE is based on VAE [14], of which marginal distribution $p(\hat{\boldsymbol{x}}|y)$ can be optimized by maximizing variational lower bound L,

$$L = \mathbb{E}[\log p(\hat{\boldsymbol{x}}_{1:N}|\boldsymbol{z}, y)] - \mathrm{KL}[q(\boldsymbol{z}|\boldsymbol{x}_{1:M})\|p(\boldsymbol{z})],$$

where KL is Kullback–Leibler divergence. The approximate posterior $q(\boldsymbol{z}|\boldsymbol{x}_{1:M})$ can be implemented as an encoder to encode input text $\boldsymbol{x}_{1:M}$ into content representation $\boldsymbol{z}$, and the output probability $p(\hat{\boldsymbol{x}}_{1:N}|\boldsymbol{z}, y)$ can be implemented as a decoder to decode output text $\hat{\boldsymbol{x}}_{1:N}$ from content representation $\boldsymbol{z}$ and style label y. During training, CVAE is trained to reconstruct input texts given style label y, so its training does not require parallel texts of source and target styles. The style labels must be known for the training and prediction of CVAE.

Content representation $\boldsymbol{z}$ is commonly modeled as a Gaussian distribution. The mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ of the Gaussian distribution are modeled by the encoder. In text style transfer, bi-directional LSTM [19] is commonly used to implement the encoder. The outputs of bidirectional LSTM at the final steps can be used to compute the parameters of latent distribution. In this case, a single latent vector represents a whole input text sequence. The latent content variable can be sampled by a reparametrization trick to enable backpropagation [14].

The output probability of the decoder can be factorized into the product of word probabilities at each time step as

$$p(\hat{\boldsymbol{x}}|\boldsymbol{z}, y) = p(\hat{\boldsymbol{x}}_1|\boldsymbol{z}, y) \prod_{m=2}^{M} p(\hat{\boldsymbol{x}}_m|\hat{\boldsymbol{x}}_{m-1}).$$

The probability distribution of the decoder is a softmax distribution to model output word probability. The sampled latent representation $\boldsymbol{z}$ and embedding vector of style label y are used to derive the decoder hidden state at the initial step. Thus, they are shown in the condition term of the first word probability. The decoder can be implemented with LSTM [20].

3. CVAE with Content Word Storage

In CVAE, all words are embedded in a fixed-dimensional latent representation. They include words that should be retained before and after the transfer. However, it is desirable to exclude retained word information from the latent representation to make effective use of the limited capacity.

Therefore, we propose Content Word Storage (CWS), a method of explicitly defining the information to be retained during style transfer as "content words" and handling them separately from the latent content feature $\boldsymbol{z}$. CWS is a method for extracting information to be retained from text input and transmitting it directly to the decoder.

We propose two CWS to extend CVAE. The first CWS is bag-of-words and the second CWS is attention mechanism [21].

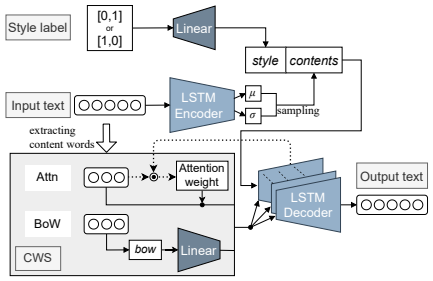

Figure 2: *Schematic diagram of CVAE + CWS.*

3.1. Bag-of-words as CWS

Bag-of-words (BoW) is one of the classical methods of representing text features. It is expressed as the sum of one-hot vectors representing the number of times a word appears in the text. BoW has a dimension equal to the vocabulary size.

A schematic diagram of using BoW as CWS is shown in Figure 2. To utilize BoW as CWS, we build a lexicon of content words from training data to define each dimension of BoW. Thus, our BoW has a dimension equal to the vocabulary size of training data. We refer to this method as CWS-BoW.

CWS-BoW can be incorporated into CVAE in the following manner. CWS-BoW first extracts only the content words from the input word sequence to obtain content word sequence $\boldsymbol{x}_{\mathrm{CW}}$. The BoW feature $\boldsymbol{z}_{\mathrm{CWS}}$ is derived from the content word sequence by summing one-hot content word vectors followed by linear transformation. The output probability when using CWS-BoW can be obtained as

$$p(\hat{\boldsymbol{x}}|\boldsymbol{z}, y) = p(\hat{\boldsymbol{x}}_1|\boldsymbol{z}, \boldsymbol{z}_{\mathrm{CWS}}, y) \prod_{m=2}^{M} p(\hat{\boldsymbol{x}}_m|\hat{\boldsymbol{x}}_{m-1}, \boldsymbol{z}_{\mathrm{CWS}}).$$

3.2. Attention mechanism as CWS

Attention mechanism [21] is a method in the sequence-to-sequence conversion framework to align input and output sequences. Attention enables the accessing of the entire input sequence to derive the hidden representation at each decoding time step. This is in contrast to the original sequence-to-sequence framework [22] in that the latter uses a fixed encoded representation to decode an entire sequence.

We propose a CWS method using attention. We expect that the hidden representation in attention can function as CWS, assuming that attention selectively attends content words in input texts. We refer to this method as CWS-Attn.

A schematic diagram of using attention as CWS is shown in Figure 2. The general workflow is as follows. Content words are extracted from the input based on the lexicon built for the BoW. The content words are converted to a content feature vector by attention at each decoding step. The content feature vector is fed to decoder to the output texts.

With CWS-Attn, the output probability is calculated as

$$p(\hat{\boldsymbol{x}}|\boldsymbol{z}, y) = p(\hat{\boldsymbol{x}}_1|\boldsymbol{z}, \boldsymbol{c}_1, y) \prod_{m=2}^{M} p(\hat{\boldsymbol{x}}_m|\hat{\boldsymbol{x}}_{m-1}, \boldsymbol{c}_m),$$

where $\boldsymbol{c}_m$ is the context vector from attention at time step m. The context vector is the weighted sum of content word vectors. CWS-Attn extracts the content words from the input word sequence to obtain content word sequence $\boldsymbol{x}_{\mathrm{CW}}$. Context vector $\boldsymbol{c}_m$ at time step m can be obtained by multiplying context word vectors and attention weights $\boldsymbol{W}_m^{\mathrm{Attn}}$; the attention weights are

calculated by taking the inner product embedded vector and hidden layer of the decoder as

$$\boldsymbol{W}_m^{\mathrm{Attn}} = \mathrm{Softmax}(\boldsymbol{e}^{\mathrm{CW}} \odot \boldsymbol{h}_m^{(D)}), \boldsymbol{c}_m = \boldsymbol{e}^{\mathrm{CW}} \cdot \boldsymbol{W}_m^{\mathrm{Attn}},$$

where $\boldsymbol{e}^{\mathrm{CW}}$ is the embedded vector of context words and $\boldsymbol{h}_m^{(D)}$ is the latent vector of the decoders at time m.

The attention mechanism has advantages over BoW in that it can consider contexts in a sentence and its capacity is proportional to the output sequence length. Note that CWS-BoW uses fixed representation $\boldsymbol{z}_{\mathrm{CWS}}$ at all decoding time steps, whereas CWS-Attn uses a different representation $\boldsymbol{c}_m$ at each time step.

4. Experimental Evaluations

We conducted two experiments on spoken style transfer: (1) disfluency and (2) dialect, to evaluate the effect of CWS on content preservation.

4.1. Datasets

For the disfluency transfer experiment, we used the corpus of the spontaneous Japanese (CSJ) dataset [23], which is a large-scale lecture speech corpus. We obtained transcriptions with and without disfluency, such as fillers (ah, er etc.), based on disfluency annotation in CSJ. The transcription without disfluency was derived by removing words labeled as disfluency. The total number of the transcripts obtained was 458k sentences. We split the transcripts into 426,400, 22,926, and 9,170 sentences to construct training, validation, and test sets, respectively, at a ratio of 93:5:2. The vocabulary size of the training set is 40,738.

For the dialect transfer experiment, we used the corpus of Kansai Vernacular Japanese (KVJ) [24] in addition to CSJ. KVJ is a collection of sociolinguistic interviews of university students with family members who were born and raised around Osaka and those who moved to Osaka during adulthood. The Kansai region has a dialect distinct from the Tokyo dialect (standard Japanese), and Osaka is the largest city in the Kansai region. We used KVJ as Kansai dialect data and CSJ without disfluency as standard dialect data. The transcripts of the CSJ portion comprised 229k sentences and those of the KVJ portion comprised 133k sentences, with a total of 362k sentences. Note that some texts in KVJ contained disfluency words such as fillers, but their amount was negligible. We split the transcripts into 337,258, 18,133, and 7,253 sentences as training, validation, and test sets, respectively, at a ratio of 93:5:2. The vocabulary size of the training set is 43,921.

We defined nouns, verbs, and adjectives as content words to build a lexicon of content words. Vocabulary sizes of each lexicon are 33,495 and 41,176. We excluded adverbs from our definition of content words because we consider adverbs in Japanese to generally not have substantive meaning nor be independent elements.

4.2. Systems and evaluation

We constructed two CVAE systems using CWS: CVAE + CWS-BoW and CVAE + CWS-Attn. We also constructed a plain CVAE system without CWS as a baseline. We included CP-VAE [25] as a conventional VAE-based method in the experiment on disfluency transfer. CP-VAE is a fully unsupervised method independent of style labels. The CP-VAE system was used as a low anchor.

The parameters common to all systems are as follows: Max epoch of the training step of 100 and learning rates of the encoder and decoder of 0.001 and 1.0, input and hidden layer dimensions of the encoder of 256 and 1024, input and hidden layer dimensions of the decoder of 128 and 1024, and dimensions of latent representation $\boldsymbol{z}$ and embedding vector of style label y of 64 and 16 respectively.

For the objective evaluation, we calculated the automatic scores using three metrics: accuracy (AC), BLEU [26], and content word error rate (CWER). AC was used to evaluate the performance of spoken style transfer, and BLEU was used to evaluate content preservation [15, 16, 25]. To calculate AC, we used pre-trained style classifier based TextCNN [27], which had 97% AC on the test set of disfluency transfer and 96% AC on the test set of dialect transfer. We also computed AC excluding content words (AC w/o CW) with the classifier to measure the AC of purer style transfer. We considered that this metrics was preferable to AC particularly for dialect transfer, because the classifier could classify styles depending on particular words that did not represent a style but merely appeared in only one of the corpora, when the transcripts of the two styles originated from different corpora, which resulted in poor accuracy. The pre-trained style classifier for AC w/o CW had 94% AC on both test sets of disfluency transfer and dialect transfer. We used the following two BLEU scores: ref-BLEU, calculated by comparison with the reference, and self-BLEU, calculated by comparison with an input. We did not evaluate ref-BLEU for dialect transfer owing to a lack of reference transcripts. CWER is the word error rate of only a content word sequence for evaluating the content word preservation. We measured these metrics for reference transcripts as well when reference transcripts were available.

For a subjective evaluation, we conducted the web-based human evaluation test. We included CVAE + CWS-Attn as the proposed system and CVAE as the baseline in the human evaluation test. The subjects were presented with a pair of original and transformed texts, and asked to evaluate them in three aspects: degree of style transfer (ST), content preservation (CP) and naturalness (Nat). ST was rated on a scale of four. CP and Nat were rated on a scale of five. We defined successful samples of style transfer on the basis of the threshold of subjective evaluation scores: we considered style transfer to be successful when ST, CP and Nat scores were greater than or equal to three. We measured the percentage of successfully style-transformed samples (Suc) as a summary of style transfer performance. We recruited ten and sixteen participants who were fluent Japanese speakers including native speakers, in the disfluency and dialect transfer experiments, respectively. Each text sample was evaluated in three or four times. We obtained 600 and 640 evaluations in total from the respective tests. We checked statistical significance using the pairwise t-test.

4.3. Results

The results of the objective evaluation in disfluency transform are shown on the left side of Table 1. Both proposed methods showed clear improvements in CWER, s-BLEU, and r-BLEU. This suggested that the introduction of CWS contributed to the preservation of content words. CWS-Attn had consistently higher scores of CWER, s-BLEU, and r-BLEU than CWS-BoW. This indicated that CWS-Attn was a more powerful content preservation method than CWS-BoW. Both BoW and Attn showed comparable AC and AC w/o CW, indicating that the introduction of CWS did not cause the degradation of style transfer performance. The poor performance of CP-VAE suggested that learning spoken styles in an unsupervised manner were difficult in disfluency transform.

Table 1: *Automatic evaluation results.*

Method	Disfluency					Dialect			
	AC↑	**AC w/o CW↑**	**s-BLEU↑**	**r-BLEU↑**	**CWER↓**	**AC↑**	**AC w/o CW↑**	**s-BLEU↑**	**CWER↓**
reference	98.35	–	79.40	100.00	0.00	–	–	–	–
CP-VAE	7.78	–	23.86	18.63	143.14	–	–	–	–
CVAE	**51.44**	49.81	38.59	35.72	54.47	**56.00**	51.86	30.85	60.99
+ CWS-BoW	49.55	**50.10**	56.40	52.05	31.37	40.77	52.27	39.47	47.78
+ CWS-Attn	49.80	50.03	**60.91**	**57.50**	**20.80**	38.98	**56.47**	**51.85**	**25.02**

Table 2: *Automatic evaluation results by transfer direction.*

Method	direction	AC↑	r-BLEU↑	direction	AC↑	AC w/o CW↑	s-BLEU↑
CVAE	fluent → disfluent	20.31	36.44	Standard → Kansai	49.90	53.57	31.94
	disfluent → fluent	82.57	33.87	Kansai → Standard	66.49	48.93	28.77
+ CWS-BoW	fluent → disfluent	19.28	51.86	Standard → Kansai	36.27	51.95	38.89
	disfluent → fluent	80.17	50.42	Kansai → Standard	48.50	52.82	39.04
+ CWS-Attn	fluent → disfluent	15.83	54.39	Standard → Kansai	35.07	52.13	53.95
	disfluent → fluent	83.77	52.32	Kansai → Standard	45.69	63.91	47.72

Table 3: *Human evaluation results. Scores are shown with 95% confidence interval.*

Method	ST↑	CP↑	Nat↑	Suc↑
	Disfluency			
CVAE	2.40 ± 0.15	2.46 ± 0.18	4.01 ± 0.24	11.67
+ CWS-Attn	2.39 ± 0.08	**3.93** ± 0.20	3.96 ± 0.20	**35.00**
	Dialect			
CVAE	2.56 ± 0.33	2.00 ± 0.15	3.61 ± 0.33	15.00
+ CWS-Attn	2.38 ± 0.24	**3.11** ± 0.23	3.39 ± 0.27	**23.44**

The results of objective evaluation in dialect transfer are shown on the right side of Table 1. The proposed methods outperform the baseline in terms of content preservation metrics such as s-BLEU and CWER in this task as well. CWS-Attn had higher scores of s-BLEU and CWER than CWS-BoW in this task as well, indicating that CWS-Attn was more effective than CWS-BoW. The two proposed systems showed lower AC values than the baseline. We suspected that this was caused by the problem of AC, as described in Section 4.2. The proposed systems, however, had comparable scores to the baseline in AC w/o CW. We therefore concluded that the proposed systems did not cause the degradation of the style transfer performance in this task as well.

Table 2 shows the objective evaluation results arranged by transfer direction. In the disfluency transfer, a huge difference in AC between the transfer directions was observed: the direction from fluent to disfluent showed greatly lower scores than the other direction. We inferred that the transfer from disfluent to fluent was easier than in the opposite direction because the former simply required the deletion of a disfluent phrase that exists in the input text. In contrast, the transfer from fluent to disfluent required decision on what and where to insert disfluent words in the original text, which was a nondeterministic task in the nature of disfluency. In the dialect transfer task, the difference was not as marked as in the disfluency task. This was due to the fact that in the dialect transfer task, insertions and deletions can occur in both directions.

The results of the human evaluation of disfluency transfer are shown at the top of Table 3. CWS-Attn improved CP significantly. ST and Nat of CWS-Attn were comparable to the baseline. This indicated CWS-Attn improved content preservation without compromising the style transfer or naturalness performance, resulting in a significant increase in the number of successful samples, shown by the high Suc percentage.

The results of the subjective human evaluation of dialect transform are shown at the bottom of Table 3. CWS-Attn improved CP significantly, and ST and Nat of CWS-Attn were not significantly degraded compared with CVAE without CWS. This showed that CWS provided a clear improvement in content preservation while minimizing the degradation of style transfer performance and naturalness in this task as well.

The results of BLEU and CWER were consistent with the trend of manually evaluated CP, indicating that content preservation evaluation works well both subjectively and objectively. In contrast, when comparing AC (w/o CW) and ST, the trends were similar for disfluency, but different for dialect. This suggests that the style learned by the model may not always match the style expected by human, and it is dangerous to simply use AC as an indicator of style transfer performance.

5. Conclusion

In this paper, we proposed a method of improving content preservation in non-parallel spoken style transfer. We extended conditional VAE (CVAE) with content word storage (CWS) to directly transmit content word information to the decoder. We proposed two CWS: bag-of-words (BoW) and attention (Attn). We evaluated them in two spoken styles: disfluency and dialect. In both experiments, CWS improved content preservation, and Attn was more effective than BoW.

This paper focused on only content preservation. We believe that any modeling of style can be combined with CWS to improve the style transfer performance. This will be a topic of our future work.

6. Acknowledgements

This paper was partly supported by a project, JPNP20006, commissioned by NEDO.

7. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[2] S. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H. Kang, "Emotional speech synthesis with rich and granularized control," in *Proc. ICASSP*, 2020, pp. 7254–7258.

[3] W. Nakata, T. Koriyama, S. Takamichi, N. Tanji, Y. Ijima, R. Masumura, and H. Saruwatari, "Audiobook speech synthesis conditioned by cross-sentence context-aware word embeddings," in *Proc. SSW*, 2021, pp. 211–215.

[4] M. Ihori, N. Makishima, T. Tanaka, A. Takashima, S. Orihashi, and R. Masumura, "Zero-shot joint modeling of multiple spoken-text-style conversion tasks using switching tokens," in *Proc. Interspeech*, 2021, pp. 776–780.

[5] M. Toshevska and S. Gievska, "A review of text style transfer using deep learning," *IEEE Transactions on Artificial Intelligence*, p. 1, 2021.

[6] M. A. Walker, G. I. Lin, and J. Sawyer, "An annotated corpus of film dialogue for learning and characterizing character style," in *Proc. LREC*, N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds., 2012, pp. 1373–1378.

[7] H. Jhamtani, V. Gangal, E. H. Hovy, and E. Nyberg, "Shakespearizing modern language using copy-enriched sequence-to-sequence models," in *Proc. the Workshop on Stylistic Variation*, 2017, pp. 10–19.

[8] K. Carlson, A. Riddell, and D. Rockmore, "Evaluating prose style transfer with the bible," *Royal Society Open Science*, vol. 5, no. 10, p. 171920, 2018.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 2672–2680.

[10] V. John, L. Mou, H. Bahuleyan, and O. Vechtomova, "Disentangled representation learning for non-parallel text style transfer," in *Proc. ACL*, A. Korhonen, D. R. Traum, and L. Màrquez, Eds., 2019, pp. 424–434.

[11] N. Dai, J. Liang, X. Qiu, and X. Huang, "Style transformer: Unpaired text style transfer without disentangled latent representation," in *Proc. ACL*, A. Korhonen, D. R. Traum, and L. Màrquez, Eds., vol. 1, 2019, pp. 5997–6007.

[12] Y. Huang, W. Zhu, D. Xiong, Y. Zhang, C. Hu, and F. Xu, "Cycle-consistent adversarial autoencoders for unsupervised text style transfer," in *Proc. COLING*, 2020, pp. 2213–2223.

[13] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV*, 2017, pp. 2242–2251.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc ICLR*, Y. Bengio and Y. LeCun, Eds., 2014.

[15] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing, "Toward controlled generation of text," in *Proc. ICML*, D. Precup and Y. W. Teh, Eds., vol. 70, 2017, pp. 1587–1596.

[16] G. Lample, S. Subramanian, E. M. Smith, L. Denoyer, M. Ranzato, and Y. Boureau, "Multiple-attribute text rewriting," in *Proc. ICLR*, 2019.

[17] L. Laugier, J. Pavlopoulos, J. Sorensen, and L. Dixon, "Civil rephrases of toxic texts with self-supervised transformers," in *Proc. EACL*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds., 2021, pp. 1442–1461.

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.

[19] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[20] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[21] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. EMNLP*, L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, Eds., 2015, pp. 1412–1421.

[22] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. NeurIPS*, 2014, pp. 3104–3112.

[23] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous Speech Corpus of Japanese," in *Proc. LREC*, 2000, pp. 947–952.

[24] H. Kevin, "An introduction to the Kansai dialect corpus," *Journal of Policy Studies*, no. 41, pp. 157–164, 2012.

[25] P. Xu, J. C. K. Cheung, and Y. Cao, "On variational learning of controllable representations for text without supervision," in *Proc. ICML*, vol. 119, 2020, pp. 10 534–10 543.

[26] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.

[27] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.