Multiple Text Style Transfer by using Word-level Conditional Generative Adversarial Network with Two-Phase Training

Chih-Te Lai[1], **Yi-Te Hong**[1,2], **Hong-You Chen**[1], **Chi-Jen Lu**[1], **Shou-De Lin**[2]
[1]Academina Sinica, Taipei, Taiwan
[2]National Taiwan University, Taipei, Taiwan
[1]{cloudylai, ted0504, cbbjames, cjlu}@iis.sinica.edu.tw
[2]sdlin@ntu.edu.tw

Abstract

The objective of non-parallel text style transfer is to alter specific attributes (e.g. sentiment, mood, tense, politeness, etc) of a given text while preserving unrelated content. Adversarial training is a popular method to ensure the transferred sentences have the desired target styles. However, previous works often suffer from content leaking problem. In this paper, we propose a new adversarial training model with a word-level conditional architecture and a two-phase training procedure. By using a style-related condition architecture before generating a word, our model is able to maintain style-unrelated words while changing the others. By separating the training procedure into reconstruction and transfer phases, our model is able to balance the reconstruction and adversarial losses. We test our model on polarity sentiment transfer and multiple-attribute transfer tasks. The empirical results show that our model achieves comparable evaluation scores in both transfer accuracy and fluency but significantly outperforms other state-of-the-art models in content compatibility on three real-world datasets.

1 Introduction

Text style transfer is a challenging problem in natural language generation, whose objective is to alter specific attributes (e.g. sentiment, mood, tense, voice, politeness, etc (Hu et al., 2017; Shen et al., 2017; Sennrich et al., 2016; Logeswaran et al., 2018; Prabhumoye et al., 2018)) of a given text while preserving its remaining attributes and contents. This task has potential applications such as paraphrasing, summarizing articles, author obfuscation (Reddy and Knight, 2016), poems/lyrics rewriting, and scenario-adaptive machine translation (Michel and Neubig, 2018).

One major challenge for text style transfer is that parallel data across different text attributes is difficult to collect and label. Without parallel data, supervised deep-learning models are not applicable and the transfer rules among styles unclear. Unlike image style transfer, another main challenge for text style transfer is the difficulty to identify and disentangle neural feature representations for texts.

In previous models (Shen et al., 2017; Hu et al., 2017) the core idea for non-parallel text style transfer, is to training an auto-encoder with additional adversarial loss (Goodfellow et al., 2014), (or a VAE with a classifier), for the discriminator, (or classifier), to guide the decoder generate sentences to have a specific target style.

Despite previous success in attribute-conditioned text generation, several research questions remain, regarding to previous models, including: 1) previous modeling limitations to transformations between only a few attributes. 2) the issue of trade-off among text fluency, content preservation, and the accurate transfer with the desired attributes and 3) the unstability of adversarial training.

To address these issues, we make contribution by adopting an encoder-decoder framework and propose a novel conditional adversarial training, including several improvements as follows: (1) a word-level attribute condition architecture in both the decoder and discriminators to capture relations between styles and words; (2) we employ a seq-to-seq attention mechanism and (3) a two-phase training procedure (reconstruction/transfer phases) for better content preservation. It is trained with standard adversarial learning approach. We test our conditional adversarial training on two tasks: (a) polarity style transfer, and (b) multiple-attribute style transfer.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3577–3582, Hong Kong, China, November 3–7, 2019.

2 Related Works

Text style transfer without parallel data is an active research topic. Mueller et al. (2017) designed a variational auto-encoder (VAE) framework; Hu et al. (2017) used VAE with controllable attributes; Shen et al. (2017) proposed to adversarially train a Cross-Aligned Auto-Encoder (CAAE) to align two different styles. To improve performances, several works including, (Fu et al., 2017; Yang et al., 2018; dos Santos et al., 2018; Logeswaran et al., 2018) were proposed. Fu et al. (2017) suggested a multi-head decoder to generate sentences with different styles; Yang et al. (2018) utilized language models as discriminators to stabilize training; dos Santos et al. (2018) used a classifier to aid style transfer; Logeswaran et al. (2018) also made use of a conditional discriminator for multiple style transfer.

On the other hand, a few works including, Li et al. (2018), Xu et al. (2018) adopt an *erase-and-replace* approach and design their methods to erase the style-related words first and then fill in words of different style attributes. Non-parallel text style transfer is also relevant to unsupervised machine translation. Prabhumoye et al. (2018), Subramanian et al. (2018), Logeswaran et al. (2018) and dos Santos et al. (2018) apply *back-translation* technique from unsupervised machine translation for style transfer task. Our work follows the framework of CAAE, and we propose several adjustments to improve the performance.

3 Methodology

3.1 Model Architecture

As shown in Figure 1, our model contains an encoder-decoder (E, G) and two discriminators D_{cnn}, D_{rnn}. We describe each model architecture in details.

3.1.1 Encoder-Decoder

Following prior works (Hu et al., 2017; Shen et al., 2017; Yang et al., 2018; Logeswaran et al., 2018), we use a seq-to-seq de-noising encoder-decoder (E, G) with attention mechanism (Bahdanau et al., 2014). For each input sentence x and attribute $\tilde{y}$, the encoder E encodes x to a latent code $z = E(x)$, and the decoder G decodes transferred sentence $\tilde{x} = G(z, \tilde{y})$, which can be further back-translated to reconstruct the original sentence. Based on this framework, we design *word-level condition* model for better results.

Word-level Condition The most unique component between our encoder-decoder and previous works is the condition architecture of attributes. Most of style transfer works (Hu et al., 2017; Shen et al., 2017; Logeswaran et al., 2018) treat attributes y as part of the initial vector fed into the RNN cell in the decoder, and we argue that the conditioning structure is important to the model performance. In our decoder G, we embed y to a vector and concatenate the vector with the output h_t of GRU cell at each time step t. More formally, at each time step t, the hidden state h_t and output probabilities o_t are generated as follows:

$$h_t = GRU(h_{t-1}, x_t, c_t|z)$$
$$o_t = \sigma(W_p([y, h_t]) + b_p)$$

where GRU denotes a Gated Recurrent Unit (Chung et al., 2015) in decoder G, and c_t is the content vector from the attention mechanism; W_p, b_p denote the projection matrix and bias to map a hidden state to an output vocabulary distribution; σ is the softmax function.

3.1.2 Discriminator

Our discriminators take output probability distributions o as inputs along with attribute labels y. Formally, for each discriminator $D \in \{D_{cnn}, D_{rnn}\}$, the function of D can be expressed as:

$$D(o, y) = f_{disc}(f_{trans}(f_{cond}(o, y))).$$

f_{cond} is a multi-layer perceptron. f_{trans} is either a bi-directional GRU or a CNN to perform a global feature transform in D_{cnn}, D_{rnn} respectively. Finally, f_{disc} is a fully-connected layer with the sigmoid function to output decisions. For simplicity, we substitute x for o as the input to discriminators in the following description.

3.2 Loss functions

We train our model with reconstruction loss L_{rec}, back-translation loss L_{bt}, discrimination loss L_{disc}, and adversarial loss L_{adv}. L_{rec} and L_{bt} force the decoder G to reconstruct the input sentence x. L_{rec}, L_{bt} are expressed as:
$L_{rec} = -\log p(x|z, y)$, $L_{bt} = -\log p(x|\tilde{z}, y)$.
L_{disc} and L_{adv} force the decoder G to output the transferred sentence $\tilde{x}$ with correct attributes $\tilde{y}$. For $D \in \{D_{cnn}, D_{rnn}\}$, L_{disc}, L_{adv} are listed as:
$L_{disc} = -\mathbb{E}_{x,y,\tilde{x},\tilde{y}}[\log D(x, y) + \log(1 - D(\tilde{x}, \tilde{y}))]$
$L_{adv} = -\sum_{D \in \{D_{cnn}, D_{rnn}\}} \{\mathbb{E}_{\tilde{x},\tilde{y}}[\log D(\tilde{x}, \tilde{y})]\}$.

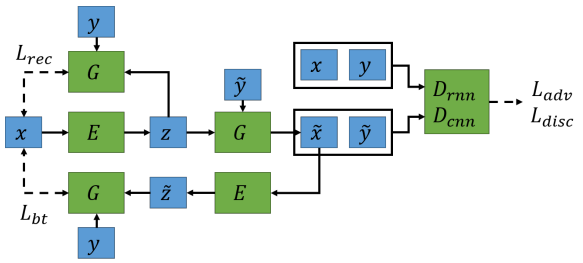

Figure 1: Overview of our model framework. Our model contains an encoder-decoder (E, G), and two discriminators D_{rnn} and D_{cnn}. We train our model with reconstruction loss L_{rec}, back-translation loss L_{bt}, adversarial loss L_{adv}, and discrimination loss L_{disc}. The dashed arrows describe where the losses are computed.

The total loss L_{tot} for our encoder-decoder (E, G) is a weighted sum of L_{rec}, L_{bt}, and L_{adv}.

3.3 Training

We split our training procedure into two phases: *reconstruction* and *transfer* phase. We first train our encoder-decoder (E, G) with loss L_{rec} in reconstruction phase, and then in transfer phase, we use total loss L_{tot} and L_{disc} to train our encoder-decoder (E, G) and the discriminators D_{cnn}, D_{rnn}, respectively. Our experiments show this approach improves content preservation significantly. A diagram about this two-phase training is provided in Appendix A.1.

4 Experiments Setup

4.1 Datasets

YelpSent Dataset The preprocessed Yelp dataset (Shen et al., 2017) consists of sentences with length limit of 15 words, labeled with either positive or negative sentiment as attributes.

AmaProd and AmaSent Dataset Amazon product dataset (He and McAuley, 2016): consists of product reviews associated with ratings of the products. We select 4 product types: (books/ movies/ electronics/ CDs), following the approach in (Kim et al., 2017) to select relevant sentences. [1]

YelpTense Dataset We use similar approach to label sentences with sentiment and past/present tense from Yelp Dataset Challenge[2]. If a sentence contains at least one verb in past tense, we label it as 'past tense'; otherwise as 'present tense'.

All the data statistics are presented in Table 5 in Appendix A.2.

4.2 Evaluation Metrics

We follow previous works and apply three automatic evaluation metrics and one human evaluation for three indicative aspects: *attribute compatibility*, *content preservation*, and *fluency*.

Attribute Compatibility To measure how well our model transfers sentence attributes, we pre-train a CNN classifier (Kim, 2014) for each attribute category (i.e. product types, sentiments, tenses) on our training data, and use the classifiers to measure the accuracy of transferred sentences associated with the desired attribute. We report the accuracy of each attribute category separately.

Content Preservation Measuring content preservation is still an open research problem. Following the previous works, we compute the BLEU score (Papineni et al., 2002) used in machine translation. In our experiments, we use self BLEU to evaluate on transferred sentences as a measurement of content preservation.

Fluency To test fluency of the generated sentences, we train a bi-directional LSTM language model on our training data for each dataset. We regard the perplexity of generated sentences as a measure for fluency.

Human Evaluation We also evaluate the transferred sentences with human assessments. In the evaluation, 8 people are asked to rate sentences based on criteria associated with three dif: attribute compatibility, content preservation, and fluency. Each aspect is rated on a 5-point Likert scale. 20 sentences on *YelpSent* with corresponding transferred sentences are randomly selected as testing examples. More details about our human evaluation are provided in Appendix A.3.

4.3 Comparison with State-Of-The-Arts

We compare with four different State-of-the-Art models: **CAAE**, **DAR**, **MultiAttr**, **ContPrev**. **CAAE** (Shen et al., 2017) consists of an auto-encoder with discriminator networks to guide text generation. **DAR** (Li et al., 2018) uses a delete-and-retrieve approach. [3] **MultiAttr** (Subramanian et al., 2018) performs multiple-attributes style transfer using back-translation (Lample et al.,

[1] We refer to dataset labeled with both sentiment and product attributes as *AmaProd Dataset*, and dataset labeled only with sentiment attribute as *AmaSent Dataset*, respectively.

[2] https://ww.yelp.com/dataset/challenge

[3] We train DAR with the source code from https://github.com/rpryzant/delete_retrieve_generate

2018, 2017; Artetxe et al., 2017) and latent representation pooling. [4] **ContPrev** (Logeswaran et al., 2018) is an auto-encoder model with a conditional discriminator for multiple attributes transfer. Among these models, in our experiments, CAAE and DAR are compared only on polarity sentiment style transfer tasks. More details about our model and training settings are provided in Appendix A.4.

5 Experimental Results

5.1 Quantitative Result

Automatic Evaluation Results Table 1 and Table 2 show the automatic evaluation results. Our model achieves higher BLEU scores and comparable transfer accuracy. We also notice that our model generates sentences with higher perplexity, while other models produce sentences with perplexity lower than the real data.

Human Evaluation Results Table 3 exhibits the human evaluation results on *YelpSent*. The results show that our sentences are evaluated higher on content preservation, and share comparable scores with other models on attribute compatibility.

Ablation test We conduct ablation experiments on our model on *YelpSent* dataset. As the results shown at Table 4, removing the word-level condition architecture decreases transfer accuracy and BLEU scores. The two-phase training procedure can also ensure a much higher BLEU scores. Using both CNN and RNN discriminator sightly improves the performance on all metrics.

Evaluation Curve We also plot the curves of transfer accuracy, BLEU score and perplexity, respectively, on the validation set, as the number of training epochs increases, across different models in Appendix A.5. According to the curves, our model achieves higher BLEU scores and relatively stable performance on all three metrics.

5.2 Qualitative Results

Our model exhibits a tendency to follow the original sentence surface structure. With the help of word-level conditional architecture, the decoder learns to make word adjustments. Sample sentences are shown in Tables 7 and 8 in Appendix A.6. For frequent occurring sentence structures across different attribute domains, our

[4] We re-implement this model with comparable results to the original publication.

Table 1: Polarity sentiment transfer on *YelpSent* & *AmaSent*

YelpSent	Sent.	BLEU	PPL
Real data	96.9%	–	20.3
Our model	87.8%	**35.5**	34.5
CAAE	83.5%	11.5	19.0
DAR	**96.3%**	0.1	20.6
MultiAttr	86.0%	12.9	**8.4**
ContPrev	91.3%	14.7	10.5
AmaSent	Sent.	BLEU	PPL
Real data	94.4%	–	34.1
Our model	77.5%	**24.5**	57.4
CAAE	75.4%	5.3	23.7
DAR	**83.4%**	0.1	37.6
MultiAttr	76.6%	12.7	**6.0**
ContPrev	72.1%	12.7	17.0

Sent. represents the transfer accuracy measured by the pretrained sentiment-attribute classifier. **BLEU** and **PPL** stand for self BLEU and perplexity score, respectively.

Table 2: Multiple attribute transfer on *YelpTense* & *AmaProd*

YelpTense	Sent.	Tense	BLEU	PPL
Real data	95.4%	99.9%	–	24.5
Our model	**79.9%**	**96.1%**	**32.2**	40.0
MultiAttr	74.5%	91.4%	25.9	**8.2**
ContPrev	76.6%	94.9%	14.7	13.4
AmaProd	Sent.	Prod.	BLEU	PPL
Real data	94.4%	95.6%	–	34.1
Our model	76.0%	87.4%	**22.4**	26.8
MultiAttr	**79.9%**	**90.4%**	15.3	**7.7**
ContPrev	75.8%	81.4%	11.2	17.7

Sent., **Tense** and **Prod.** represent the transfer accuracy measured by the pretrained sentiment-, tense- and product-attribute classifiers, respectively.

Table 3: Human evaluation on *YelpSent*

YelpSent	Cont.	Flu.	Sent.
Our model	**4.01**	3.54	3.17
CAAE	3.14	2.99	3.19
DAR	1.17	**4.24**	3.35
MultiAttr	3.26	3.91	**3.59**
ContPrev	3.02	3.76	3.56

Columns **Cont.**, **Flu.**, **Sent.** stand for human evaluation scores for *content preservation*, *fluency* and *sentiment compatibility*, each measured in a Likert scale from 1 through 5. (1: lowest and 5: highest)

Table 4: Model ablation test on *YelpSent*

Model	Sent.	BLEU	PPL
Real data	96.9%	–	20.3
Our model	87.8%	35.5	34.5
w/o two-phase	86.4%	26.2	44.5
w/o word-level	91.5%	3.5	34.2
w/o CNN	87.3%	28.5	39.3
w/o RNN	86.1%	31.7	42.7
w/o CNN, RNN	42.4%	72.5	20.7

model shows good results. When the original sentence sometimes has a distinct sentence structure and would require multiple editing in a larger scale, our model may not find a correct similar sentence-level surface form adjustment in the target-attribute domain and the word-level local adjustments may sometimes not be coherent.

6 Conclusion

In this paper, we conduct non-parallel style transfer among multiple attributes. We propose a seq-to-seq model with word-level condition and two-phase training. The empirical results demonstrate that our model outperforms our competitors in the polarity sentiment transfer task on *YelpSent*. In multiple attribute transfer tasks, our model also achieves comparable results with the state-of-the-art MultiAttr on *YelpTense* and *AmaProd*. We also analyze our model with ablation tests.

Although our model achieves better content preservation, the general quality of our transferred sentences can be further improved. Also, designing proper evaluation metrics is still an open problem for text style transfer. We leave these two questions as the future works.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2015. Gated feedback recurrent neural networks. In *International Conference on Machine Learning*, pages 2067–2075.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2017. Style transfer in text: Exploration and evaluation. *arXiv preprint arXiv:1711.06861*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517. International World Wide Web Conferences Steering Committee.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. *arXiv preprint arXiv:1703.00955*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Yoon Kim, Kelly Zhang, Alexander M Rush, Yann LeCun, et al. 2017. Adversarially regularized autoencoders. *arXiv preprint arXiv:1706.04223*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with attribute controls. In *Advances in Neural Information Processing Systems*, pages 5108–5118.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 312–318.

Jonas Mueller, David Gifford, and Tommi Jaakkola. 2017. Sequence to better sequence: continuous revision of combinatorial structures. In *International Conference on Machine Learning*, pages 2536–2544.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 866–876.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 189–194.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6830–6841.

Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text style transfer. *arXiv preprint arXiv:1811.00552*.

Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. *arXiv preprint arXiv:1805.05181*.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. *arXiv preprint arXiv:1805.11749*.