Text Style Transfer Dissection: Deeper understanding on Text Style Transfer

Nyungwoo Lee
School of Computing, KAIST
leenw2@kaist.ac.kr

Jungsoo Lee
Graduate School of AI, KAIST
bebeto@kaist.ac.kr

Donghyun Jeon
NAVER
donghyeon.jeon@navercorp.com

Nakyil Kim
Graduate School of AI, KAIST
nakyilkim@kaist.ac.kr

Abstract

Text style transfer, changing the style of an input sentence while preserving the content, has been explored extensively recently. However, due to the vague definition of 'style', previous studies limited their research on changing the formality or sentiment. In this work, we conduct an extensive experiments on text style transfer by manipulating hyper-parameters of an existing text style transfer model, Domain Adaptive Text Style Transfer. We further expand the definition of 'style' by conducting text style transfer on a newly crawled lyric datasets. Additionally, we show how the performance of existing model can improve by applying early stopping based on geometric mean between BLEU score and style transfer accuracy.

1 Introduction

Vanilla sequence-to-sequence model is a groundbreaking language modeling for response generation that can generate sentences of different lengths. However, this model often generates bland and generic responses (1). Beam search (2; 3) or re-ranking algorithm solves this problem, but it is hard to use in real-time applications due to the low diversity of generated sentences or time-consuming property. In addition, CVAE(Conditional Variational Autoencoder) models using Gaussian distribution can also generate a variety of sentences (4), but the generated sentences are still boring or generic.

To address this issue, some studies focus on changing the style of sentences or diversifying sentences. Text style transfer refers to the study of changing the 'style' of an input sentence while preserving the 'content', overall information included in the sentence. It can transfer a general sentence into varying sentences. The term 'style' was limited in the previous studies (e.g., changing the formal sentences to informal sentences or positive sentences to negative sentences). Our research team think the style of a sentence can be extended to more abstract areas. For example, in the lyrics of a song, style can be the standard for distinguishing whether a song is a love song or a hip-hop song. Style can be even extended to human personality or political orientation. Thus, we aim to generate a stylized sentence which has a more abstract style area.

We planned to take the motif of 'Domain Adaptive Text Style Transfer (DAST)' (5) which was proposed by Li *et al*. They proposed 'DAST' that can transfer the style while preserving the content of an input sentence through a lightweight model. Based on this work, we try to have better understanding of text style transfer through extensive experiments. Our experiments split into three folds: 1) manipulating the pre-trained epochs in which the model is trained to reconstruct sentences without stylizing, 2) changing the coefficient of the generator loss which drives the model to generate

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

stylized sentences, and 3) applying early stopping based on the geometric mean between BLEU score and style transfer accuracy.

This work provides the following contributions:

- We conduct extensive experiments on text style transfer and promote better understanding of it.
- We adopt text style transfer on more abstract datasets and extend the definition of 'style' which was limited in the previous studies.
- We show how text style transfer can improve by applying early stopping based on geometric mean between BLEU score and style transfer accuracy.

2 Related Work

Neural Response Generation has been predominantly evolved in recent years with the development of language model (6; 7; 8). Although these models achieved remarkable performances in word representations, the input data used for the models are general which often lead to bland and general responses when the models are applied real world applications (1). In order to mitigate the issue, numerous researchers have dived into the exploration of generating diverse or stylized sentence outputs. The rudimentary consideration in text style transfer is that the style of an input sentence should be changed while the preserving the content.

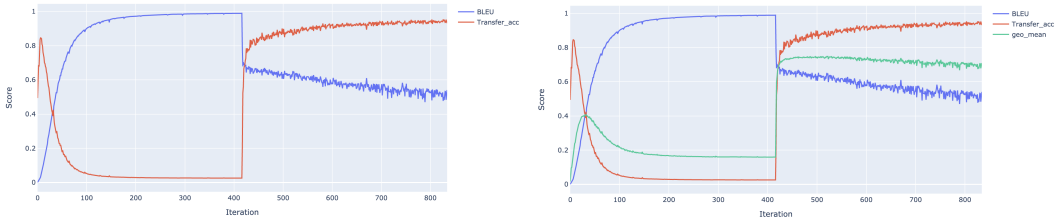
Gao *et al* designed 'StyleFusion' which maps the sentences of the general dataset(e.g. Reddit) and the stylized dataset(e.g. script of drama character) into the same latent space and stylizes an input sentence with the closest style sentence. Their model increased the relevance of responses while maintaining some style intensity by using the combination of sequence to sequence (S2S) encoder and auto encoder to consider the relevance of the answers through a generated sentence. Also, Li *et al*. proposed 'Domain Adaptive Text Style Transfer', a light-weighted model which stylizes an input sentence while preserving the content. They used autoencoder to maintain content resiliency of the source domain, and apply style classifier to force meaningful style information into the model. The model trained to restore the sentences of the source corpus and target corpus, change the style of target corpus, and then proceed training by distinguishing the changed style with a style classifier. As a result, they proved that their models can learn content resiliency in the source domain and style information in the target domain.

3 Approach

In this work, we plan to disclose two important questions: 1) the range of style in a text and 2) which part should the model consider for preserving the content. In NLP domain, style refers mainly to the explicit attributes of text such as positive/negative sentiment or formality. However, in the real world, what can be referred as the style of text is not limited to the aforementioned attributes. Additionally, since the definition of content and style are rather vague, we want to find how the model recognizes content and style while training.

We need to produce sentences by adjusting the content and style independently to clarify the above questions. So we conduct experiments based on the model proposed by Li *et al*. We leveraged this model because it uses data from source domain for content preservation and target domain for style transfer which well aligns with our intention. In this paper, domain refers to the data type such as movie review or restaurant review and style refers to the feeling or nuance of the sentence such as positive/negative.

Li *et al* trained their model in two different circumstances: 1) when style of source data is unknown or unavailable and 2) when style of source data is same as that of target data. The former model refers to 'DASTC' while the latter one refers to 'DAST'. They apply style classifier to autoencoder to help autoencoder better receive style information of the target domain. Since style classifier can learn well with less data than autoencoder, target data with less amount increases the dependence on style classifier. This situation leads the model to ignore the content and rely only on the style when generating a sentence.



(a) BLEU and Style Transfer Accuracy

(b) BLEU, Style Transfer Accuracy and Geometric mean

Figure 1: Preliminary results of the BLEU, Style Transfer Accuracy and Geometric Mean

Model	Input (negative)	Transferred sentence (negative-to-positive)
DASTC	service was slow, disinterested	service was delicious, satisfying
	skip this place and stay on the strip	delicious this place and stay on the strip
	I would give this review negative stars	I definitely give perfect review terrific stars
	service was lacking, waitress didn't know the menu	service was delicious, waitress did blast know the menu
	swimming in the grease from the beef topping	refreshing in the inside from the beef topping
Model	Input (positive)	Transferred sentence (positive-to-negative)
DASTC	just tried their food, it was amazing!	just tried their food, it was terrible!
	prices are great and the food was excellent	prices are awful and the food was disappointing
	very clean and attentive and also very quick	very clean and outdated and also very disappointing
	very good lunch, service, and atmosphere	very low lunch, service, and atmosphere
	we really enjoy everything we've had here	we really left everything we've had here

Table 1: Results of text style transfer with Yelp dataset of the preliminary, where red denotes negative term and blue denotes positive term

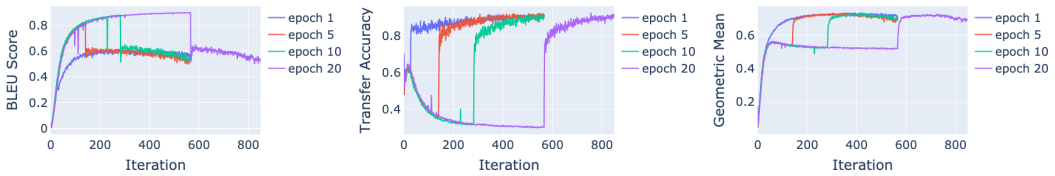
To solve the above problem, the author proposed DASTC, which tries to increase content preservation by using the massive source data, even if it did not know the style label of the source domain. This model uses jointly training for source data and autoencoder of target data. It can the learn general content information from the massive source data and style information from the target data, allowing the model to transfer the style while preserving the content better. In case the source domain and the target domain shared the same style, the authors build a model named DAST. We aim to reveal 1) the range of styles within text and 2) how a model divides content and style. We used DAST and DASTC in appropriate cases and crawled a new dataset.

4 Preliminary

We first try to understand how DASTC works by replicating the model of the paper. Preliminary experiments use IMDB movie review corpus (9) as source data, and Yelp restaurant review dataset (10) as target data. The experiment was carried out according to the default setting of the paper and the original dataset, setting pre-trained epochs to 10 and total epochs to 20. The results of the experiment are as follows.

In the pre-training phase, the model is trained to restore source sentences through reconstruction loss. During this training phase, the model learns to preserve the content of the massive source sentence. After this phase, the model is trained to change the style into the opposite style by adding the loss related to style intensity. Fig. 1 shows that the style accuracy increases as the model after the pre-training stage. Although the text style transfer task is aimed at increasing the accuracy of style transfer, the original meaning of the original sentence should also be preserved. As we continue the training, the style accuracy will continue to increase, but the content preservation of the original sentence will decrease. Thus, a baseline is needed for training, and we concluded that we needed to adopt early stopping based on the geometric mean between BLEU score and style accuracy.

Table. 1 shows the translated sentences from IMDB dataset to Yelp dataset. We report the examples of positive transfer of negative reviews, and negative transfer of positive reviews. The model appears



(a) BLEU Score (b) Style Transfer Accuracy (c) Geometric mean

Figure 2: BLEU score, style transfer accuracy and geometric mean graphs of models with variation of pre-trained epochs. 'Epoch' in the figures refer to the pre-trained epochs.

Pre-trained Epochs	Domain Acc.	Style Transfer Acc.	BLEU Score	Geometric Mean
1	0.9952	0.8676	0.5976	0.7201
5	0.9933	0.9043	0.5826	0.7258
10	0.9934	0.8613	0.6114	0.7257
20	0.9899	0.8567	0.6000	0.7169

Table 2: Quantitative analysis results of models with variation of pre-trained epochs

to adjust style mainly by changing modifiers. Therefore, we think the model tends to translate words that describe other words such as adjectives while preserving the original words. Also, we found that the model fairly translated the metaphor or indirect expressions well. For example, in the fifth example in the negative-to-positive transfer, the negative review starts with an expression 'swimming in the grease' which means that the food was greasy. The translated sentence translated this phrase by starting with 'refreshing' which means that it grasped the meaning of 'swimming in the grease'. After the preliminary study, we further wanted to explore if the model changes the text style with different datasets.

5 Experiment

In order to perform text style transfer with a different dataset, we used song lyrics. There are two genres in our lyric dataset: 1) love and 2) hiphop. Text style transfer with songs lyrics transfers love songs to hiphop songs and hiphop songs to love songs. For the love songs, we crawled the titles of songs which were under the criteria named 'Most Loved' from Spotify [1]. After collecting the titles of the love songs, we crawled the lyrics of these songs from Google. Similarly, we crawled the hiphop songs from 'hip-hop-rap' genre form Songlyrics website [2]. We pre-processed the newly collected lyrics by discarding irrelevant parts such as "[Chorus]" or "[Verse 1]" and adopted text style transfer on this dataset.

We conducted three main experiments on DASTC with lyric datasets: observing the difference by 1) adopting early stopping based on the geometric mean, 2) manipulating weight of the style loss, and 3) changing pre-trained epochs. For the first experiment, we applied early stopping on the DASTC based on the geometric mean of BLEU score and style transfer accuracy. Since text style transfer needs to change the style and preserve the content simultaneously, the model needs to take both measurements into consideration. Regarding this feature, we calculated the geometric mean of the two measurements and applied early stopping with the patience of three thousand iterations. We observe how early stopping helps text style transfer in both content preservation and style modification in this experiment through qualitative analysis. For the second experiment, we differentiated the number of pre-trained epochs to observe the difference of the stylized outputs. DASTC is first trained to reconstruct the original input sentence during the pre-trained epochs. Thus, we expect the model would have less capability to perform text style transfer as the number of pre-trained epochs increases. For the last experiment, we differentiated the weight of the style loss. When training DASTC, we multiply the coefficient on the style loss to manipulate the amount of gradient that the model should

[1] https://www.spotify.com/

[2] http://www.songlyrics.com/news/top-genres/hip-hop-rap/

	Style	Lyrics
Input	**love**	and i swear by the **moon** and the stars in the sky ill be there
early stopping	hip-hop	and i swear by the **mc** and the stars in the am ill not stan
Last epoch	hip-hop	and i swear by the **kick** and the stars in the **police** ill not stan
Input	**love**	every **inch** of your **skin** is a holy **gray** I have got to find
early stopping	hip-hop	every **hip** of your **back** is a holy **hip** I have got to find
Last epoch	hip-hop	every **mcs** of your back is a **gangsta hip** I have got to **joint**
Input	**love**	i never thought that you would be the one to **hold my heart**
early stopping	hip-hop	i never thought that you would be the one to **ask joint plug**
Last epoch	hip-hop	i never thought that you would be the one to **jock swift swift**
Input	**love**	if she changes her mind this is the first place she will **go**
early stopping	hip-hop	if she changes her mind this is the first place she can **properly**
Last epoch	hip-hop	if she **because** her mind this is the first place she **police uhhuh**

Table 3: Results of text style transfer with song lyrics with/without early stopping

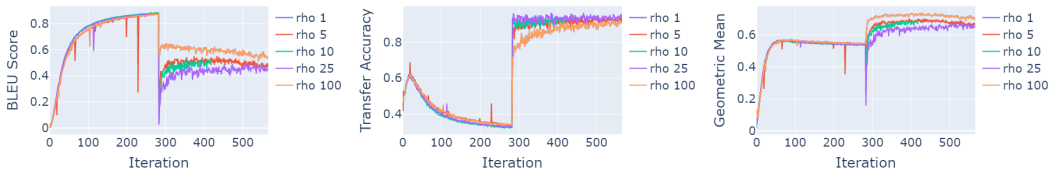
receive when stylizing. We expect that transferred sentences are more stylized when the coefficient of the style loss is high.

5.1 Effects of early stopping

We conducted qualitative analysis for evaluating how early stopping improved the performance of text style transfer. Table. 3 illustrates the two different sentences style transferred from a model which adopted early stopping and from a model which was trained until the maximum epoch. We selected several sentences from love songs to analyze. The transferred results show how the lyrics from love songs are transferred into lyrics of hiphop songs. Regarding the first sentence, the model with the early stopping changed 'moon' to 'mc' while preserving most of the words in the sentence. In hiphop, mc refers to the 'microphone checker' and it symbols the self-ego of a rapper. The model stylized the sentence by enhancing the willness to swear and convey the meaning of the sentences overall. However, the model which was trained until the last epoch changed 'moon' to 'kick' and 'sky' to 'police' which degrades the content preservation of the original sentence. In the second example, the model with early stopping changed 'inch' to 'hip', 'skin' to 'back', and 'gray' to 'hip'. While the original sentence tried to convey the complement to a lover, the stylized sentence amplified the complement in a hiphop style. However, the sentence generated from a model which was trained until the maximum epoch focused on generating words from hiphop songs without preserving the meaning of the original sentence. Similar tendency can be found in the third and fourth example in the table. Through these examples, our research team concluded that early stopping is essential for content preservation and training more epochs rather hampers the content preservation.

5.2 Effects of pre-trained epoch

We manipulated the number of pre-trained epochs and checked the effect of pre-trained epochs on text style transfer. The tendency of BLEU score and style transfer accuracy well-aligned with that of preliminary study. In this experiment, we kept track of geometric mean this time to consider the trade-off between the two evaluation metrics. Table. 2 shows the domain accuracy, style transfer accuracy, BLEU score and geometric mean score when the model finished training due to early stopping. This table also shows that the number of pre-trained epoch does not influence the evaluation metrics. Fig. 2 shows the BLEU score, style transfer accuracy score, geometric mean score of models by changing the pre-trained epochs. We observed that there is no clear difference between the evaluation metrics as the pre-trained epochs changed. We originally expected that increasing the number of pre-trained epochs would improve the capability of the model to preserve content well was wrong. However, we observed that evaluation metrics converged to a certain value after same number of iteration for training style transfer.



(a) BLEU Score (b) Style Transfer Accuracy (c) Geometric mean

Figure 3: BLEU score, style transfer accuracy and geometric mean graphs of models with variation of of Style Loss weight

ρ	Domain Acc.	Style Transfer Acc.	BLEU Score	Geometric Mean
1	0.9923	0.8491	0.6119	0.7207
5	0.9950	0.9031	0.5101	0.6785
25	0.9925	0.9327	0.4274	0.6309
100	0.9585	0.9585	0.3871	0.6090

Table 4: Quantitative analysis results of models with variation of ρ

5.3 Effects of weights in style loss

We observed how the weight of losses affect the text style transfer. The loss function of DAST is composed of reconstruction loss and style loss. A weight coefficient ρ is multiplied to the style loss and we observed how it affects the output of our model quantitatively and qualitatively. ρ controls the amount of gradient to flow in the style loss. BLEU score measures the content preservation and transfer accuracy measures how the model successfully transferred the style of output which means that there is a trade-off between two measurements. Fig. 3 well-aligns with our expectation we had after the preliminary study.

Table. 5 demonstrates the generated sentences as we manipulated ρ. Like our original expectation, as we increased ρ, the model tried to transfer the original sentence with words that are frequent in hip hop songs. However, we also found that increasing the value of ρ severely (e.g., increasing ρ to 100) rather failed to convey the original meaning of the input sentence.

5.4 Further experiment

We extended our research in order to explore whether our model can transfer sentences by reflecting a person's characteristic. However, collecting massive amount of text datasets that contain individual's characteristics was a demanding work for our research team, so we opted to leverage the dataset of Tweet data(Democrats and Republican Dataset [3] and Trump and Hillary Dataset [4]). We used the Democrats and Republican Dataset for the source dataset and Trump and Hillary Dataset for the target dataset.

We removed non formal text such as emoticon, links and hash tags using Preprocesser API [5]. We additionally removed non-alphabetic words including punctuation lowered-cased the text.

The overall quality of the transferred sentences are inferior when compared to the previous datasets used in this work. However, it was possible to observe that the transferred sentences included nouns which were not transferred in the previous datasets. With restaurant review datasets, the model mostly transferred the adjectives not nouns. However, Democrats and Republic Dataset, nouns such as election pledges, supporters, constituencies also affect the style of the sentence. Due to the characteristic of the dataset, we observed that the model also transferred nouns also. For example, Hillary frequently used words such as 'woman', 'child care'. 'democratic' while Trump used 'presidential', 'job care', 'republic' frequently. However, this unique characteristic of this

[3] https://www.kaggle.com/kapastor/democratvsrepublicantweets

[4] https://www.kaggle.com/erikbruin/text-mining-the-clinton-and-trump-election-tweets

[5] https://github.com/s/preprocessor

ρ		Lyrics
	Input	and i swear by the **moon** and the stars in the sky ill be there
1	Early stop	and i swear by the moon and the **short** in the style ill be there
5	Early stop	and i swear by the **rhythm** and the stars in the **suckers** ill be there
25	Early stop	and i swear **as** the **role** and the stars in the **cell not phase** there
100	Early stop	and i swear by the moon and the stars **in the jungle be street heard**
	Input	every **inch** of your **skin** is a **holy gray** I have got to find
1	Early stop	every **mcs** of your **brothers** is a **dandy uhh** I have got to find
5	Early stop	every **slick** of your **vocal** is a **few uncle** I have got to find
25	Early stop	every **stain** of your **skin** is a **sayin off** I am got to find
100	Early stop	**visiting livin** of your **skin** is a **old soul** I have got to find
	Input	oh her eyes her eyes make the **stars** look like they are not shinin
1	Early stop	oh her eyes her eyes make the **stars** look like they are not
5	Early stop	oh she eyes her eyes make the **o** look like **people** are not
25	Early stop	o her can to **bear** make the **stars** look like they are not
100	Early stop	oh her eyes her eyes make the **school** look like they are not

Table 5: Results of text style transfer with song lyrics with different ρ

	Style	Sentence
Input	**Hillary**	it took years but will be the year of the first **woman** POTUS great day to be an American **woman**
	Trump	it took years but will be the winner of the very **presidential person** my speech for very many the
Input	**Hillary**	I want you to know that I see you and I am with you Hillary to the Latino community at
	Trump	I want you to know that I am me today to me with I hope to the amazing amp
Input	**Hillary**	if fighting for **affordable child care** and paid family leave is playing the woman card, then deal me
	Trump	if fighting for **jobs care** than jobs is setting the opposite card then, will join me
Input	**Trump**	the ratings for the **republican national convention** were very good but for the final night my speech great thank you
	Hillary	the ratings for the **democratic congressional convention** is also happening on my heart for this speech my democratic friend
Input	**Trump**	this is why I would think the **unions** would support
	Hillary	this is why I would think the **women** would support
Input	**Trump**	we are going to bring steel and **manufacturing** back to Indiana
	Hillary	we are going to bring the **community** and build back to Indiana

Table 6: Results of text style transfer with Twitter data of Trump and Hillary

dataset hampers the model to recognize the difference between content and style. The transferred sentences with this dataset actually failed to preserve the content of the original sentence.

6 Discussion

6.1 Improvement of the style classifier

We observed that the performance of style classifier was robust when we used datasets which included distinct styles (e.g., lyric dataset with love and hiphop). However, the style classifier showed poor performance with datasets which did not have a clear distinct styles (e.g., Twitter dataset with Trump and Hillary). Since the performance of style classifier is directly related to the quality of text style transfer, improving the accuracy of the style classifier would enhance the quality of the transferred sentences. One possible method would be using a pre-trained BERT (6) and fine tune it for improving the style classifier.

6.2 Differences in model styles and content recognition by data configuration

By conducting experiments with diverse datasets, we also observed that the model learned the difference between content and style according to the characteristic of the train dataset. For example, when the model is trained with a restaurant review dataset (e.g., Yelp dataset), the content of target datasets remain similar such as type of food while the style includes adjectives which describe the positive or negative reviews.

6.3 Cycle consistency loss

Another future work is leveraging the cycle consistency loss which was originally proposed in CycleGAN (11). Cycle consistency loss forces $F(G(X) \approx X$ where F and G indicates the generator from source domain to target domain and vice versa, respectively. This loss has been used widely in NLP recently. Adding cycle consistency loss on our work would promote the quality of the stylized sentences. One issue is that this loss is non-differentiable in NLP. Thus, circumventing this problem by leveraging techniques widely used in reinforcement learning such as gumbel softmax would be one applicable method.

7 Conclusion

In this work, we conducted an extensive experiment on text style transfer. To be more specific, we leveraged the model of DAST (5) and observed the differences when we changed the hyper-parameters that influenced the text style transfer. Additionally, we crawled datasets of lyrics which included songs of love or hiphop. With the newly crawled dataset, we reported the quantitative analysis and qualitative analysis to analyze text style transfer. The limitation of our work is that we did not improve the architecture of the original model by adopting recently proposed models such as BERT (6). Adopting more sophisticated architectures and developing the performance of DAST would be one future work for exploring text style transfer. We believe that our findings give a clear blue print for future researchers who plan to dive into text style transfer.

References

[1] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 110–119. [Online]. Available: https://www.aclweb.org/anthology/N16-1014

[2] J. Li, W. Monroe, and D. Jurafsky, "A simple, fast diverse decoding algorithm for neural generation," *arXiv preprint arXiv:1611.08562*, 2016.

[3] I. Kulikov, A. Miller, K. Cho, and J. Weston, "Importance of search and evaluation strategies in neural dialogue modeling," in *Proceedings of the 12th International Conference on Natural Language Generation*, 2019, pp. 76–87.

[4] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 654–664. [Online]. Available: https://www.aclweb.org/anthology/P17-1061

[5] L. Dianqi, Z. Yizhe, G. Zhe, C. Yu, B. Chris, S. Ming-Ting, and D. Bill, "Domain adaptive text style transfer," in *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[8] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *ICLR*, 2020. [Online]. Available: https://openreview.net/pdf?id=r1xMH1BtvB

[9] Q. Diao, M. Qiu, C.-Y. Wu, A. J. Smola, J. Jiang, and C. Wang, "Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars)," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 193–202.

[10] J. Li, R. Jia, H. He, and P. Liang, "Delete, retrieve, generate: A simple approach to sentiment and style transfer," *arXiv preprint arXiv:1804.06437*, 2018.

[11] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networkss," in *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[12] J. Gao, M. Galley, and L. Li, "Neural approaches to conversational ai," in *The 41st International ACM SIGIR Conference on Research Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 1371–1374. [Online]. Available: https://doi.org/10.1145/3209978.3210183

[13] L. Zhou, J. Gao, D. Li, and H. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *CoRR*, vol. abs/1812.08989, 2018. [Online]. Available: http://arxiv.org/abs/1812.08989

[14] X. Gao, Y. Zhang, S. Lee, M. Galley, C. Brockett, J. Gao, and B. Dolan, "Structuring latent spaces for stylized response generation," *arXiv preprint arXiv:1909.05361*, 2019.

[15] P. Michel and G. Neubig, "Extreme adaptation for personalized neural machine translation," *arXiv preprint arXiv:1805.01817*, 2018.

[16] S. Mirkin, S. Nowson, C. Brun, and J. Perez, "Motivating personality-aware machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1102–1108.