Disentangling Meaning and Style for Positive Text Reframing

Xu Sheng[1], Yoshimi Suzuki[2], Jiyi Li[2], Kentaro Go[2], and Fumiyo Fukumoto[2]
[1]Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences
[2]Interdisciplinary Graduate School
University of Yamanashi
{g22dts03,ysuzuki,jyli,go,fukumoto}@yamanashi.ac.jp

Abstract

Positive Text Reframing (PTR), as the new branch of Text Style Transfer (TST) task has attracted interest from researchers as it is crucial and extensively applicable in the NLP area. Due to the significant generalization and representation capability of the Transformer based pre-trained language model (PLM), a beneficial baseline can be easily obtained by just fine-tuning the PLM on the annotated dataset directly. However, it is a challenging problem to transfer the sentiment attributes of the source sentence into a sentence that gives a positive perspective while preserving the original sense of the context. In this paper, we disentangle positive text reframing into aspects: a sentence meaning and style and learn a model for each aspect, i.e., paraphrase generation and sentiment transfer to boost the generation performance on a positive perspective. Experimental results on Positive Psychology Frames (PPF), show that our approach outperforms the baseline by seven evaluation metrics.

1 Introduction

Text style transfer (TST), as one of the important NLP tasks, has been explored by the frame language-based systems by [1] and schema-based Natural Language Generation by [2] in the 1980s. The goal is to change the text style, such as formality, politeness, or sentiment with preserving the original sense of the source text. With a recent surge of interest in deep learning techniques, TST has had much attention and positive text reframing (PTR) has been explored as one of the sub-fields in TST research. As shown in the example sentence from Positive Psychology Frames (PPF) [3] in Figure 1, the goal of PTR is to generate a sentence with more positive sentiment and preserve the original content meaning of the given sentence.

Leveraging supervised learning with parallel data is one type of route among various approaches for the TST task. This line of approaches are, for example, Xu et al. [4] and Zhang et al. [5] proposed a multi-task learning-based method. Rao and Tetreault [6] presented data augmentation strategies for mitigating a small size of the training dataset. Another line is to utilize a non-parallel dataset. John et al. proposed a disentanglement method [7], and Shen et al. presented a cross-alignment algorithm to perform style transfer [8]. Fu et al. attempted to explore non-parallel data by using adversarial networks [9]. Lai et al. designed two rewards of target style and content for formality style transfer based on reinforcement learning paradigm [10].

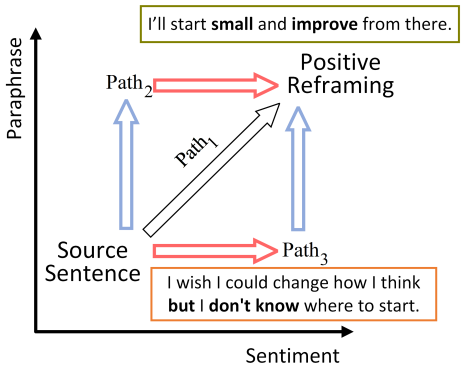

Figure 1 The Concept for Disentangling Meaning and Style

The main challenge in the PTR task is how to control diversity and extent of style transfer, i.e., the trained model, which straightforward generates target by end-to-end following the $Path_1$ in the Figure 1, and finally either simply copies most of the words that appeared in the source input to preserve the meaning, or transfers the input sentence with different sentiment polarity, causes a lack of diversity or reduce transferring sentiment quality. To transfer the source sentence into a diverse and positive target, we propose a simple approach that divides PTR into two aspects: sentence meaning and style. As shown in Figure 1, the model is trained for each aspect, i.e., paraphrase generation and sentiment transfer, and further fine-tuned

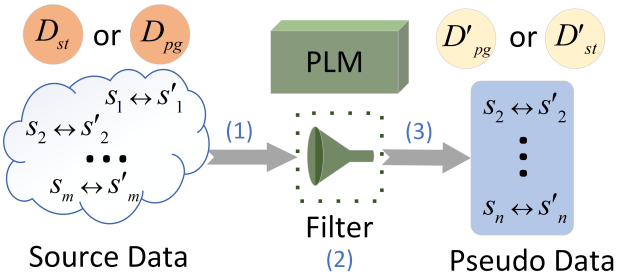

Figure 2 The procedure for creating pseudo data

to fuse the capabilities learned from these two aspects by alternative paths, Path_2 and Path_3, instead of Path_1.

The contributions of this paper can be summarized: (1) we propose two data augmentation strategies to generate pseudo-positive reframing datasets for disentangling the PTR, (2) we propose a simple but effective multi-task learning-based model to fuse the capabilities learned from the two pseudo datasets for PTR, and (3) The experimental results show that our pseudo datasets and our strategies can improve the performance compared with the baseline on PPF dataset.

2 Methodology

2.1 Creating Pseudo Data as Prior Knowledge

We utilized existing paraphrase and sentiment datasets to create two pseudo-parallel datasets instead of creating a sentence reframing dataset manually. Figure 2 illustrated the procedure for creating two synthetic datasets. The procedure consists of three steps.

(1) **Selecting annotation pairs**

For the paraphrase generation auxiliary task, we choose Microsoft Common Objects in COntext (MSCOCO) and call it D_{pg}. In contrast, for the sentiment transfer auxiliary task. Shen et al. modified the huge Yelp reviews dataset for sentence-level sentiment analysis [8]. We divided it into two sets, S_{neg} and S_{pos} consisting of sentences with negative and positive sentiment labels, respectively. We created pairs for $\forall s_i \in S_{neg}$, and $\forall s'_i \in S_{pos}$. To reduce the computation cost, for a given s_i, we randomly chose the number of $0.05 \times |S_{pos}|$ samples from the set S_{pos}. Therefore, $0.05 \times |S_{pos}| \cdot |S_{neg}|$ negative and positive sentence pairs are gained. We call the result D_{st}.

(2) **Training PLM as Filter**

To utilize two datasets, D_{pq} and D_{st} as pseudo datasets of PTR, each sentence of a pair extracted from D_{pq} should be different polarity from each other. Similarly, each sentence of a pair from D_{st} should be a similar meaning. To this end, a semantic similarity classifier is trained as a semantic filter (F_{sem}), to remove inappropriate from D_{st}. In the same manner, a sentiment classifier is trained as a sentiment filter (F_{senti}) and predicts one of the three polarities, i.e., negative, neutral, and positive. To simplify our model, two common parallel paraphrase generation and sentiment analysis datasets, for measuring semantic similarity and sentiment polarity classification can be utilized to obtain F_{sem} and F_{senti} respectively by leveraging PLM.

(3) **Filtering Two Pseudo Datasets**

We recall that the goal of PTR is to generate a sentence that gives a positive perspective with preserving the original sense of the source sentence. Therefore, the model F_{sem} predicts the semantic similarity score ranging from 0 to 5.0. The higher the score value, the more semantically similar the two sentences are. Likewise, we chose two types of sentence pairs only, i.e., ($Negative_s \rightarrow Neutral_s$, $Neutral_s \rightarrow Positive_s$ labeled by F_{senti}) from the set D_{pg}, resulting in pseudo set D'_{pg}.

2.2 Fusion Strategies

The straightforward fine-tuning of PLM is indicated in the path marked with $Path_1$ of Figure 1. The strategy requires the model to directly learn the capability to inject diversity (paraphrase generation) and improve positive perspective (sentiment transfer) for the source sentence. However, it is challenging for the model to directly capture all of the complicated features at once. We thus divide this path into two relative steps to make the problem easier i.e., paraphrase generation and sentiment transfer, which are marked with blue and red colors in Figure 1. The model further fuses these two steps by utilizing two alternative paths masked with $Path_2$ (from paraphrase generation to sentiment transfer) and $Path_3$ (from sentiment transfer to paraphrase generation). To this end, we propose an approach that consists of two fine-turning stages and four data flows illustrated in Figure 3. More specifically, after training the PLM on two pseudo datasets, D'_{pg} and D'_{st}, parallel by using multi-task learning (Stage 1), the same model is further fine-tuned on PPF dataset following four optional data flow candidates, marked as ST, PG, PG2ST, and ST2PG (Stage 2). The decoder of PLM is a shared

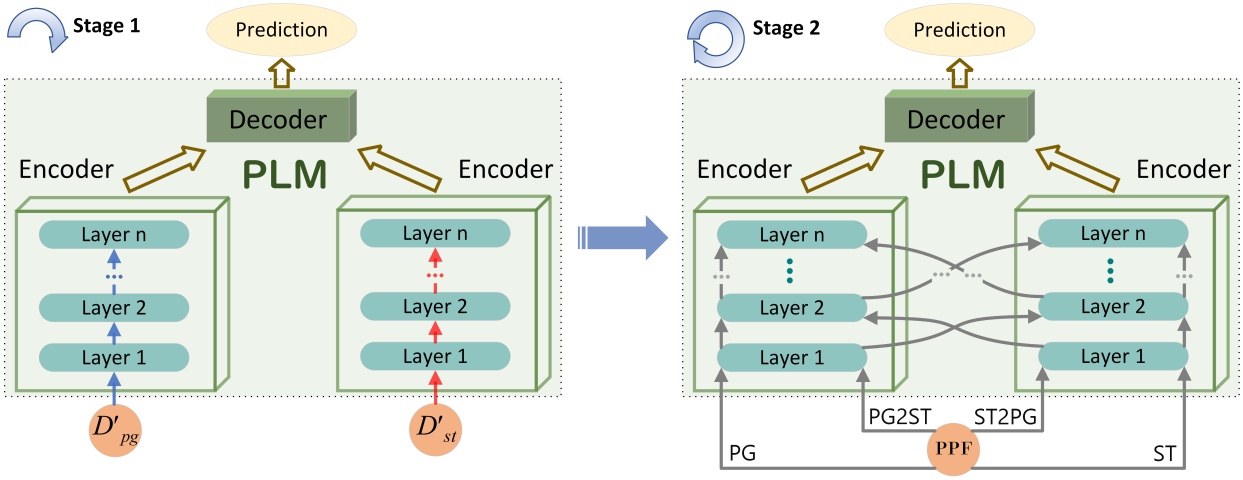

Figure 3 The model structure and data flow

part, while the two encoders are used to model to generate paraphrase, and transfer sentiment, respectively.

The first fine-tuning stage based on multi-task learning which is shown in Figure 3, can also balance the paraphrase generation and sentiment transfer. Therefore, the data flow PG, and ST, which only pass one of the two encoders, are regarded as two implicit fusion strategies. In contrast, the remaining two data flows, i.e., PG2ST and ST2PG which are classified by the order of hidden layers from two encoders are called explicit fusion strategies.

3 Experiment

Dataset	Train	Validation	Test
PPF	6,679	835	835
D'_{pg}	15,181	13.4	1,899
D'_{st}	14,807	139	215
STSB	5,749	1,500	1,379
TE-sentiment	45,615	2,000	12,284

Table 1 The statistics of Five Datasets Invovled

3.1 Experimentail Setting

Following Ziems et al. [3] setting, we chose BART pre-trained model as the PLM in our method [11]. The original BART is downloaded from the version "facebook/bart-base", Hugging Face [1)]. The five dataset statistics are summarised in Table 1. Semantic Textual Similarity Benchmark (STSB) [12] and TweetEval Sentiment (TE-sentiment) [13] are used to train the filters, F_{sem}, and F_{senti}, respectively. During the first stage in which PLM is trained with the prior knowledge to learn deriving diversity and transfer positivity, we utilized the multi-task learning algorithm proposed by Liu et al. [14] to fine-tune the PLM.

During the second stage, we utilized the PPF dataset to evaluate our method. It consists of 8,349 sentence pairs with manual annotation. The same BART trained in the first stage is further trained on the PPF training set. We tuned the hyperparameters as follows: the batch size is 4, 8, 16, 32, the number of epochs is from 2 to 5, and the value of the learning rate is from 1e-5 to 1e-4. The procedure of tuning hyperparameters is automatically conducted by the third-party library named "Ray Tune"[2)].

For a fair comparison with the baseline proposed by [3], we included three metrics in our evaluation metrics. The metrics are: (1) ROUGE [15], BLUE [16] and BERT-Score [17] referring to the gold reference for assessing the performance on content preservation. (2) The Δ TextBlob value [18] for assessing the positivity transfer effectiveness. (3) The Average Length and Perplexity [19], followed by [20] for measuring the fluency of the output sentences.

3.2 Results

Table 2 shows the main results on PPF test dataset. We can see from Table 2 that the results obtained by our approach improve the performance compared with the baseline. This indicates that our approach contributes to give a positive perspective while preserving the original contents. Four out of five content preservation metrics show that our method can keep the meaning of source sentences better as the improvements on ROUGE-1, ROUGE-2, and ROUGE-L are 5.2%, 2.7%, and 3.0% respectively. Although there is no significant improvement on Δ TB, the

1) https://huggingface.co/facebook/bart-base

2) https://docs.ray.io/en/latest/tune/index.html

Method	R-1	R-2	R-L	BLUE	BScore	Δ TB	Avg.Len	PPL
BART (baseline)	27.7	10.8	24.3	10.3	**89.3**	0.23	24.4	24.2⋆
ST (ours)	32.7	13.2	26.8	9.8	89.2	**0.24**	**27.7**	**22.8**
PG (ours)	**32.9**	**13.7**	**27.3**	10.9	89.1	0.17	25.9	25.9
PG2ST (ours)	32.5	13.4	27.1	**11.0**	89.0	0.12	24.6	26.7
ST2PG (ours)	32.4	13.4	27.0	10.9	88.9	0.12	24.7	26.8

Table 2 Main results Against the baseline on a test set on PPF. Δ TextBlob (Δ TB) refers to the evaluation metric about the positivity improvement. Average Length (Avg.Len) and Perplexity (PPL) denote the fluency metrics. Bold font shows the best result with each line. The result marked with ⋆, which has not been reported in the source paper, is obtained by following the same hyperparameter setting of the baseline, BART

orignial sentence 1	Been a hell of a week. Tired is not the word. Anxieties on top of anxieties. Overworked. Over tired. Emergency vet trips. Dog mum stress. No words down yday. Today, rest. Tomorrow we start again. WritingCommnunity WritingCommunity.
output of ST	I hope I am able to get through this week better than last week.
output of Baseline	Been a hell of a week. Anxieties on top of anxieties. Overworked. Over tired. Emergency vet trips. Dog mum stress. Today, rest. Tomorrow we start again.
orignial sentence 2	Trying to remind myself that bombing this audition is a good thing. Stupid Jazz stupid scholarships.
output of ST	I hope I get into a better frame of mind going forward so that I can do my best to win this audition.
output of Baseline	I'm trying to remind myself that bombing this audition is a good thing.
orignial sentence 3	1 test down, 1 test to go. usually i dont hate fridays but this one sucks.
output of ST	I have 1 test down, 1 test to go. I hope this one goes well.
output of Baseline	1 test down, 1 test to go. Usually I don't hate fridays but this one sucks. But I'm sure next time I'll be better.

Table 3 Some examples from test set of PPF, and their correspond reframe from our ST varient and Baseline. The parts marked with pink color are critical for positivity increase and content preservation.

decrease of perplexity by 1.4 scores demonstrates that the sentence output from our ST variant is more fluent than that from the baseline method.

Table 3 illustrates some example sentences obtained by our approach, ST, and the baseline. As shown in words/phrases highlighted by the green color, the output sentences generated by our model are more proper than that by the baseline. For instance, in the first example, "hope" and "better" are more positive expressions and the rest part keeps the meaning and topic of the original sentence, while the output of the baseline is duplicated with the source input. Likewise, "best to win this audition" in the output obtained by our method increases the positivity of sentiment style and preserver the content properly for the second input sentence. Although "But I'm sure next time I'll be better" in the output by baseline method for the last example transfers the sentiment from negative to positive, the first short sentence blindly copies the counterpart from the input. Obviously, the entire output of our method is more fluent and diverse, such as "I hope this one goes well.", compared with the baseline.

4 Conclusion

In this paper, we proposed a method for positive text reframing by leveraging two pseudo-datasets, paraphrase pairs with sentiment polarities, and sentiment pairs with paraphrases created by utilizing existing sentiment and paraphrase datasets. The experimental results on the PPF dataset showed that our simple approach attained good performance compared with the baseline, especially, we found that it is effective for generating fluent sentences. Future work will include: (1) Exploring more effective fusion strategies by leveraging multi-task learning techniques, (2) Applying our approach to other tasks on TST, and (3) Incorporating sentiment transfer knowledge during pre-training stage by leveraging a huge number of non-parallel datasets.

Acknowledgement

This work was supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2117, JKA, and Grant-in-aid for JSPS, Grant Number 20K11904.

References

[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**, pp. 187–193, 1985.

[2] Eduard Hovy. Generating natural language under pragmatic constraints. **Journal of Pragmatics**, Vol. 11, No. 6, pp. 689–719, 1987.

[3] Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. Inducing positive perspectives with text reframing. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3682–3700, 2022.

[4] Ruochen Xu, Tao Ge, and Furu Wei. Formality style transfer with hybrid textual annotations, 2019. https://arxiv.org/abs/1903.06353.

[5] Yi Zhang, Tao Ge, and Xu Sun. Parallel data augmentation for formality style transfer. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3221–3228, 2020.

[6] 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, 2018.

[7] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 424–434, 2019.

[8] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In **Advances in Neural Information Processing Systems**, Vol. 30, 2017.

[9] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. **Proceedings of the AAAI Conference on Artificial Intelligence**, Vol. 32, No. 1, pp. 663–670, 2018.

[10] Huiyuan Lai, Antonio Toral, and Malvina Nissim. Thank you BART! rewarding pre-trained models improves formality style transfer. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)**, pp. 484–494, 2021.

[11] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880, 2020.

[12] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**, pp. 1–14, 2017.

[13] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In **Findings of the Association for Computational Linguistics: EMNLP 2020**, pp. 1644–1650, 2020.

[14] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4487–4496, 2019.

[15] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, 2004.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In **Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics**, pp. 311–318, 2002.

[17] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In **International Conference on Learning Representations**, pp. 74–81, 2020.

[18] Steven Loria. textblob documentation. **Release 0.16**, Vol. 2, , 2018.

[19] Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. Unsupervised text style transfer using language models as discriminators. In **Advances in Neural Information Processing Systems**, Vol. 31, 2018.

[20] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. **Computational Linguistics**, Vol. 48, No. 1, pp. 155–205, 2022.