Balancing the Style-Content Trade-Off in Sentiment Transfer Using Polarity-Aware Denoising

Anonymous ACL submission

Abstract

We present a polarity-aware denoising-based sentiment transfer model, which accurately controls the sentiment attributes in generated text, preserving the content to a great extent. Though current models have shown good results, still two major issues exist: (1) target sentences still retain the sentiment of source sentences (2) content preservation in transferred sentences is insufficient. Our proposed polarity-aware enhanced denoising mechanism helps in balancing the style-content trade-off in sentiment-controlled generation. Our proposed method is structured around two key stages in the sentiment transfer process: better representation learning using a shared encoder (pre-trained on general domain) and sentiment-controlled generation using separate decoders. Our extensive experimental results show that our method achieves good results for balancing the sentiment transfer with the content preservation.

1 Introduction

Text sentiment transfer is the task of changing the sentiment properties of the text while retaining the sentiment-independent semantic content within the context (Shen et al., 2017; Prabhumoye et al., 2018; Li et al., 2018; Luo et al., 2019).

With the success of deep learning in the last decade, a variety of neural methods have been recently proposed for this task (Toshevska and Gievska, 2021). If parallel data are provided, standard sequence-to-sequence models can be directly applied (Rao and Tetreault, 2018). However, due to lack of parallel corpora (paired source data and target data), sentiment transfer represents a research challenge. The first line of research disentangles text representation into its content and attribute in a latent space and applies generative modeling (Hu et al., 2017; Shen et al., 2017; Prabhumoye et al., 2018). Another line of research is prototype editing (Li et al., 2018), which extracts a sentence template and its attribute markers to generate the text. These research lines are further advanced with the emergence of transformer-based models (Sudhakar et al., 2019; Malmi et al., 2020). These methods mainly focus on how to disentangle the content and style in the latent space. The latent representation needs to preserve the meaning of the text while abstracting away from its stylistic properties, which is not trivial (Lample et al., 2018). Theoretically, disentanglement is impossible without inductive biases or other forms of supervision (Locatello et al., 2019).

Our work addresses this problem with more supervision, which is obtained automatically by implementing polarity-aware denoising. First, we randomly delete (or mask) pivot word(s) of input sentences. Then a shared encoder pre-trained on general domain helps in preparing a latent representation, followed by separate sentiment-specific decoders that are used to change the sentiment of the original sentence. We follow back-translation for style transfer approach proposed by Prabhumoye et al. (2018) to represent the sentence meaning in the latent space. Our proposed model gets us the best performance for a style-content trade-off.
Our contributions are summarized as follows:

- We design a sentiment transfer model using an extended transformer architecture and polarity-aware denoising. Our extensions provide more control while generating outputs with changed sentiment.

- We introduce polarity-masked BLEU (MaskBLEU) and similarity score (MaskSim) for automatic evaluation of content preservation in this task. These metrics are derived from the traditional BLEU score (Papineni et al., 2002) and Sentence BERT-based cosine similarity score (Reimers and Gurevych, 2019). In our approach, we mask polarity words beforehand for sentiment-independent content evaluation.

- We develop a new non-parallel sentiment transfer dataset derived from Amazon Review Dataset (Ni et al., 2019). It is more topically diverse than earlier used datasets Yelp (Li et al., 2018) and IMDb (Lin et al., 2011), which were majorly focused on movie and restaurant/business-related reviews. We will publish our dataset with the final version of this paper.

- Both automatic and human evaluations on our dataset show that our proposed approach generally outperforms state-of-the-art (SotA) baselines. Specifically, with respect to the content preservation, our approach achieves substantially better performance than other methods.

2 Related Work

Sentiment Transfer A common method for sentiment transfer task is to separate content and style in a latent space, and then adjust the separated style. Hu et al. (2017) use the variational auto-encoder (Kingma and Welling, 2013) model to derive the disentanglement of the content between the generated sentence and the original sentence through KL divergence loss. Fu et al. (2017) compare a multi-decoder model with a setup using a single decoder and style embeddings. Shen et al. (2017) proposed a cross-aligned auto-encoder with adversarial training to learn a shared latent content distribution and a separated latent style distribution. Prabhumoye et al. (2018) propose to perform text style transfer through the back-translation method. In a recent work, He et al. (2020) present a new probabilistic graphical model for unsupervised text style transfer. Although their approach is able to successfully change the text style, it also changes the text content, which is a major problem.

Latent Representation Many previous methods (Hu et al., 2017; Shen et al., 2017; Fu et al., 2017; Prabhumoye et al., 2018) formulate the style transfer problem using the encoder-decoder framework. The encoder maps the text into a style-independent latent (vector) representation, and the decoder generates a new text with the same content but with a different style using the latent representation and a style marker. The major issue of these models is poor preservation of non-stylistic semantic content.

Content Preservation To further deal with the above problem, Li et al. (2018) first extract content words by deleting phrases, then retrieves new phrases associated with the target attribute, and finally uses a neural model to combine these into a final output. Style transformer (Dai et al., 2019) uses transformer as a basic module for training a style transfer system. Luo et al. (2019) employs a dual reinforcement learning framework with two sequence-to-sequence models in two directions, using style classifier and back-transfer reconstruction probability as rewards. Though these works have shown some improvement over the previous works, they are still not able to properly balance the objectives of preserving the content while transferring the style. Our polarity-aware denoising technique aims to solve this problem by specifically targeting and changing polarity words while preserving the rest of the content.

Evaluation Another challenge remains in the evaluation of controllable NLG models. There is no clear standard for evaluating the output of natural language generation (Novikova et al., 2017). Previous work on style transfer (Hu et al., 2017; Prabhumoye et al., 2018; Dai et al., 2019; He et al., 2020) has re-purposed metrics from other fields such as BLEU (Papineni et al., 2002) and PINC (Chen and Dolan, 2011) for evaluation. However, none of the techniques is capable of evaluating style transfer methods specifically with respect to preservation of content (Toshevska and Gievska, 2021). These metrics do not take into account the necessity of changing individual words while altering the sentence style. Intended differences between the source sentence and the transferred sentence are thus penalized. In this regard, we have introduced polarity masked BLEU score (MaskBLEU) and polarity masked similarity measure (MaskSim), where we have masked the polarity words beforehand.

3 Method

Given two datasets, $X_{pos} = \{x_1^{(pos)}, \ldots, x_m^{(pos)}\}$ and $X_{neg} = \{x_1^{(neg)}, \ldots, x_n^{(neg)}\}$ which represent two different sentiments pos and neg, respectively, our task is to generate sentences of the desired sentiment while preserving the meaning of the input sentence. Specifically, we generate samples of dataset X_{pos} such that they belong to sentiment neg and samples of X_{neg} such that they belong to sentiment pos. We denote the output of dataset X_{pos} transferred to sentiment neg as $X_{pos \rightarrow neg} = \{\hat{x}_1^{(neg)}, \dots, \hat{x}_n^{(neg)})\}$ and the output of dataset X_{neg} transferred to sentiment pos as $X_{neg \rightarrow pos} = \{\hat{x}_1^{(pos)}, \dots, \hat{x}_n^{(pos)}\}$.

In all our experiments, we train the sentiment transfer models using back-translation between English and German (Section 3.1). First, we present transformer-based baselines for sentiment transfer with style-conditioning (Section 3.2). Next, we propose an approach based on the extended transformer architecture, in which we use separate modules (either the whole transformer model, or the transformer decoder only) for the respective target sentiment (Section 3.2). We further improve upon our approach using polarity-aware denoising (Section 3.3) which we propose as a new scheme for pre-training the sentiment transfer models.

3.1 Back-translation

Back-translation for style transfer was introduced in Prabhumoye et al. (2018). Following their approach, we use back-translation for getting a latent text representation for our sentiment transfer task. We refer to this experiment as *Back-Translation*. Prior work has also shown that the process of translating a sentence from a source language to a target language retains the meaning of the sentence but does not preserve the stylistic features related to the author's traits (Rabinovich et al., 2016).

We also experimented with an auto-encoder, but we have found that the back-translation model gives better results for sentiment transfer. We hypothesise that it is due to the fact that back-translation allows to neglect word boundaries, resulting in a more abstract latent representation.

3.2 Our Base Models

We present several straight-forward baseline approaches. The first baseline is a back-translation model based on a vanilla transformer architecture (Vaswani et al., 2017) in which we add source sentiment identifiers (<*pos*> or <*neg*>) to the output. At the time of sentiment transfer we interchange the sentiment identifiers (<*pos*> $\rightarrow$ <*neg*>, <*neg*> $\rightarrow$ <*pos*>). We refer to this experiment as *Style Tok*.

We extend the first baseline by adding a sentence-style loss and a style embedding. For the style loss, we use a pre-trained transformer-based sentiment classifier's[1] (Wolf et al., 2020) polarity score as sentence-style loss and we add the same to the translation loss (from the back-translation process, Section 3.1). For better supervision during training, we also add randomly initialized style embedding along with the transformer's token and position embeddings. We refer to this experiment as *Style (Tok + Embedd + Loss)* .

We then extend the transformer's encoder-decoder architecture to have more control over the sentiment-specific generation. We train two separate transformer models for the positive and negative sentiment text generation, using only sentences of the target sentiment in training. During inference, the model is fed with inputs of the opposite sentiment, which it did not see during training. We refer to this experiment as *Two Sep. transformers*.

We further extend the above approach by using a shared encoder and separate decoders. During training, both negative and positive text is passed through the same shared encoder and the positive and negative texts are generated by the respective decoders. The sentiment transfer is achieved by decoding the shared latent representation using the decoder for the opposite sentiment. We refer to this experiment as *Shrd Enc + Two Sep Decoders*.

3.3 Polarity-Aware Denoising

We devise a task-specific pre-training (Gururangan et al., 2020) scheme for improving the style transfer abilities of the model. Our pre-training scheme—*polarity-aware denoising*—uses polarity labels for adding more supervision on the word level.

We experiment with three approaches: deleting or masking (1) *general* words (i.e., all the words uniformly), (2) *polarity* words (i.e., only high-polarity words according to a lexicon), or (3) *general* and *polarity* words together (with a different probability for each). We use a German polarity lexicon to automatically identify the pivot words. We prepared the German polarity lexicon by first translating the words from German to English using an off-the-shelf translation system, followed by labeling the words with *positive* and *negative* labels using the English NLTK Vader lexicon (Hutto and Gilbert, 2014).

We use polarity-aware denoising for pre-training the encoder, following the shared encoder and separate decoders design from Section 3.2. The encoder is further fine-tuned during the sentiment transfer training.

[1]https://github.com/huggingface/transformers

3.4 Summary of Our Method

Here we summarize the final design of our proposed method, in which we combine our proposed model and our new denoising scheme.

We translate English input text x_{en} to German text x_{de} using our translation model (Section 3.1). Next, we prepare a noisy text x_{noise} from x_{de} using the polarity-aware denoising technique (Section 3.3) as follows:

$$x_{noise} = Noise(x_{de}; \theta_N). \tag{1}$$

We provide x_{noise} to the shared encoder of the *German* $\rightarrow$ *English* back-translation model. The model converts the text to the latent representation z as follows:

$$z = Encoder(x_{noise}; \theta_E) \tag{2}$$

where, θ_E represent the parameters of the shared encoder and z is derived from a pre-trained encoder trained with general domain data (this encoder is not style specific).

During training, the latent representation z (of positive/negative text) is passed through respective decoders as follows:

$$\hat{x}_{pos} = Decoder_{pos}(z; \theta_{D_{pos}}) \tag{3}$$

$$\hat{x}_{neg} = Decoder_{neg}(z; \theta_{D_{neg}}) \tag{4}$$

Finally, the sentiment transfer is achieved by decoding the shared latent representation using the decoder for the opposite sentiment as follows:

$$\hat{x}_{neg} = Decoder_{pos}(z; \theta_{D_{pos}}) \tag{5}$$

$$\hat{x}_{pos} = Decoder_{neg}(z; \theta_{D_{neg}}) \tag{6}$$

where $\hat{x}_{neg}$, $\hat{x}_{pos}$ are the sentences with transferred sentiment conditioned on z and $\theta_{D_{pos}}$ and $\theta_{D_{neg}}$ represent the parameters of the positive and negative decoders, respectively.

Figure 1 shows the overview of our proposed architecture.

4 Experiments

4.1 Datasets

For our back-translation process and model pre-training, we have used the WMT14 English-German *(en-de)* dataset (1M sentences) from Neidert et al. (2014).

For finetuning and experimental evaluation, we built a new English sentiment dataset, based on the Amazon Review Dataset (Ni et al., 2019). We have selected Amazon Review because it is more diverse topic-wise (books, electronics, movies, fashion, etc.) than existing datasets Yelp (Li et al., 2018) and IMDb (Lin et al., 2011), which are majorly focused on movie and restaurant/business-related reviews. While the data is originally intended for recommendation, it lends itself easily to our task. We have split the reviews to sentences using NLTK (Bird et al., 2009) and then used a pre-trained transformer-based sentiment classifier (Wolf et al., 2020) to select the sentences with high polarity. Our intuition is that high-polarity sentences are more informative for the sentiment transfer task than neutral sentences.

We filter out short sentences (less than 5 words) since it is hard to evaluate content preservation for these sentences. We also ignored sentences with repetitive words (e.g., *"no no no no thanks thanks."*) because these sentences are noisy and do not serve as good examples for the sentiment transfer model. We evaluated and compared our approaches with several state-of-the-art systems (Shen et al., 2017; Prabhumoye et al., 2018; Li et al., 2018; Luo et al., 2019; Wang et al., 2019; He et al., 2020) on our dataset.

The statistics of our sentiment dataset are shown in Table 1. We aim for comparable size to existing datasets (Li et al., 2018).

Dataset	Positive	Negative
Train	100k	100k
Valid	1k	1k
Test	1k	1k
Avg sent. length (words)	13.04	

Table 1: Our sentiment dataset statistics.

4.2 Training Setup

In all our experiments, we have used a 4-layer transformer (Vaswani et al., 2017) with 8 attention heads in each layer. The hidden size, embedding size, and positional encoding size in transformer are all set to 512. During our experiments, we have tested various combinations of noise settings w.r.t. noise probability, noise type (general or polarity-aware denoising), and noise mode (deleting or masking). These parameters are selected based on our preliminary experiments with the translation model (see Section 3.1). The parameters are encoded in the

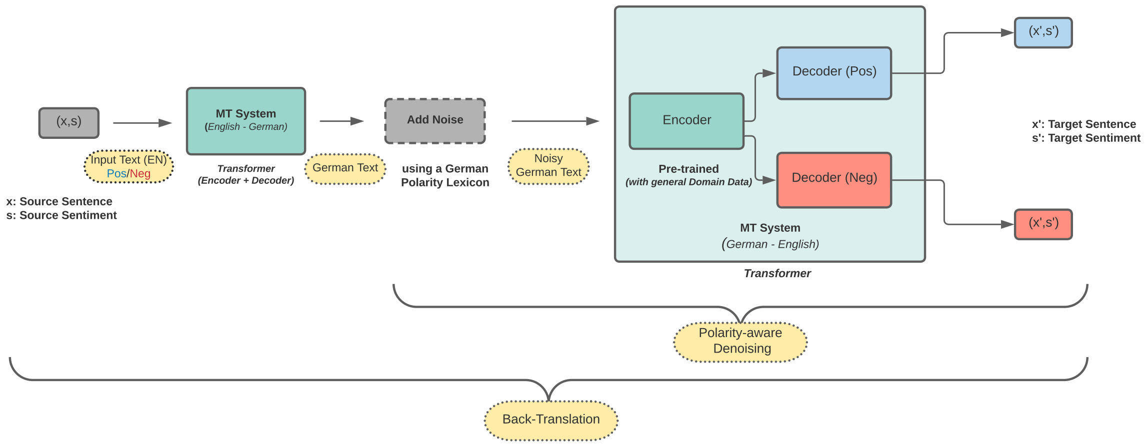

Figure 1: Our sentiment transfer pipeline. In the pipeline, we (1) *translate* the source sentence from English to German using a transformer-based machine translation (MT) system; (2) *apply noise* on the German sentence using a German polarity lexicon; (3) *encode* the German sentence to latent representation using an encoder of German-to-English translation model; (4) *decode* the shared latent representation using the decoder for the opposite sentiment.

name of the model as used in Table 2 (see the table caption for details).

4.3 Evaluation and Results

To evaluate the performance of the models, we compare the generated samples along three different dimensions using automatic metrics, following previous work: (1) style control, (2) content preservation, and (3) fluency. Furthermore, we perform human evaluation of the model outputs.

4.3.1 Automatic Evaluation

Style Accuracy We measure sentiment accuracy automatically by evaluating the target sentiment accuracy of transferred sentences. Instead of using our own data-based sentiment classifier, we use the pre-trained transformer based sentiment analysis pipeline (Wolf et al., 2020) for unbiased evaluation.

Content Preservation: Common Metrics To measure content preservation, we calculate the BLEU score (Papineni et al., 2002) between the transferred sentence and its source. Higher BLEU score indicates higher n-gram overlap between the sentences, which correlates with better content preservation. We also compute Sentence BERT (Reimers and Gurevych, 2019) based cosine similarity score to match the vector space semantic similarity between the source and the transferred sentence. None of the techniques is capable of evaluating style transfer methods specifically with respect to preservation of content in style transfer (Toshevska and Gievska, 2021). These metrics do not take into account the necessity of changing individual words while altering the sentence style. Intended differences between the source sentence and the transferred sentence are thus penalized.

Content Preservation: Newly Introduced Metrics To avoid the problems of the commonly used metrics, it makes sense in sentiment transfer to evaluate the content and similarity while ignoring any polarity tokens. Thus, we introduce MaskBLEU and MaskSim scoring methods – these are identical to BLEU and cosine similarity, but they are computed on sentences where polarity words (found by NLTK Vader (Hutto and Gilbert, 2014)) have been masked. This allows measuring content preservation while ignoring the parts of the sentences that need to be changed.

Fluency We use the negative log-likelihood score from the GPT-2 (Radford et al., 2019) language model as an indirect metric for evaluating the sentence fluency. We also calculate average sentence lengths of the sentiment-transferred sentences. We normalize the score from GPT-2 by the sentence length.

4.3.2 Human Evaluation

Automatic metrics are not sufficient to evaluate the quality of the transferred sentence (Novikova et al., 2017). Therefore, we also conduct human evalu-

Models	Accuracy	Sim	MaskSim	BLEU	MaskBLEU	LM Score	Avg-SL-Tg	Avg (AC-MS-MB)
Back-Translation Only (Section 3.1)								
Back-translation only	0.4	0.8282	0.7684	27.99	45.30	-78.61	11.90	40.85
Our Models (Vanilla) (Section 3.2)								
Style Tok	13.2	0.5356	0.5596	4.77	8.64	-52.08	7.64	25.93
Style (Tok + Embedd + Loss)	19.4	0.6719	0.6553	8.43	18.04	-116.76	20.96	34.32
Two Sep. transformers	89.3	0.3940	0.6109	6.78	19.59	-79.04	13.74	56.66
Shrd Enc + Two Sep Decoders	88.1	0.3968	0.6001	7.35	20.05	-77.98	12.50	56.03
Pre Training Enc	55.3	0.5916	0.7317	22.65	33.92	-93.34	13.40	54.13
Our Models (w/ Denoising) (Section 3.3)								
WG01-AG01-D	71.4	0.5173	0.6944	17.07	29.78	-88.73	13.71	56.87
WG01-AG01-M	68	0.5361	0.7108	19.45	31.06	-86.31	12.63	56.71
WG03-AG03-D	83	0.4466	0.6481	11.71	24.45	-82.97	13.72	57.42
WG03-AG03-M	78.8	0.4815	0.6686	14.23	28.20	-82.73	12.98	57.96
WP08-AP08-D	66.9	0.5276	0.7010	19.47	31.34	-82.81	12.38	56.12
WP08-AP08-M	64	0.5475	0.7260	21.37	33.99	-89.10	12.87	56.86
WP1-AP1-D	58.7	0.5703	0.7265	22.70	33.06	-87.21	12.23	54.81
WP1-AP1-M	58.9	0.5673	0.7156	22.25	32.97	-86.55	12.22	54.48
WG03-AG01-D	68	0.5294	0.6966	17.87	30.86	-89.50	13.26	56.17
WG03-AG01-M	80.7	0.4730	0.6649	13.95	27.47	-82.75	13.07	58.22
WG01-AG03-D	85.2	0.4411	0.6461	11.75	25.38	-79.77	13.05	58.40
WG01-AG03-M	70	0.5339	0.7111	19.66	32.26	-84.34	12.38	57.80
WP08-AP1-D	61.6	0.5778	0.7362	22.54	34.95	-94.42	13.42	56.73
WP08-AP1-M	60.9	0.5543	0.7244	21.97	33.34	-85.54	12.55	55.56
WP1-AP08-D	68.5	0.5255	0.6987	19.27	31.15	-83.99	12.42	56.51
WP1-AP08-M	61.1	0.5603	0.7142	21.46	32.88	-85.99	12.12	55.13
WG03-AP08-D	67	0.5335	0.6968	20.26	31.73	-84.31	12.54	56.13
WG03-AP08-M	65.7	0.5464	0.7249	21.21	33.49	-85.02	12.53	57.23
WP08-AG03-D	83.3	0.4360	0.6354	11.00	24.32	-80.50	13.31	57.05
WP08-AG03-M	79.6	0.4730	0.6647	13.22	26.87	-83.14	13.21	57.65
WG03P08-AG03P08-D	65.5	0.5466	0.7045	20.31	32.56	-90.43	13.17	56.17
WG03P08-AG03P08-M	82	0.4600	0.6647	13.69	27.45	-79.60	12.75	**58.64**
State-of-the-Art Models								
Shen et al. (2017)	88.6	0.3462	0.5129	3.23	18.31	-73.99	10.95	52.73
Li et al. (2018)	69.9	0.4573	0.6318	14.69	25.33	-85.13	12.19	52.80
Luo et al. (2019)	92.4	0.2786	0.4684	0.00	9.14	-42.00	7.81	49.43
Prabhumoye et al. (2018)	93.5	0.3078	0.5042	0.86	15.16	-61.05	10.28	53.03
Wang et al. (2019)	79.3	0.3850	0.5449	10.56	20.28	-116.84	15.13	51.36
He et al. (2020)	91.5	0.3516	0.5422	9.53	21.78	-65.89	8.23	55.83

Table 2: **Automatic evaluation**. *Accuracy*: Sentiment transfer accuracy. *Sim* and *BLEU*: Cosine similarity and BLEU score between input and sentiment-transferred sentence. *MaskSim* and *MaskBLEU*: Masked similarity and BLEU score (same as conventional similarity and BLEU score, but polarity words are masked beforehand). *LM Score*: Average log probability assigned by vanilla GPT-2 language model. *Avg-SL-Tg*: Average length of transferred sentences. *Avg(AC-MS-MB)*: Average score between sentiment transfer accuracy, masked similarity score and masked BLEU score. *Back-Translation Only* model is explained in Section 3.1, *Our Models (Vanilla)* are explained in Section 3.2. *Our models (w/Denoising)* involve our polarity-aware denoising technique, explained in Section 3.3. All numbers are based on a single run, with identical random seeds. Model names reflect noise settings as follows: *W* denotes WMT pretraining data, *A* denotes Amazon finetuning data, the following tokens denote noise probability values are associated with the respective data. *G/P* represents general/polarity token noising, *D/M* represents noising mode deletion/masking (e.g, *WG03P08-AG03P08-D*: noise probabilities on WMT data and Amazon data are identical. Both general and polarity token noising are applied (with probabilities 0.3 and 0.8, respectively). Deletion is applied in this specific setting.

ation experiments on same dataset. We randomly select 100 source sentences (50 for each sentiment) from each test set. For each example, the original sentence and the sentence with the changed sentiment are shown to the annotator. The annotators rate the outputs using a 1-5 Likert scale (Likert, 1932) for style control, content preservation, and fluency.

4.4 Results

Results of the automatic metrics are presented in Table 2. Compared to the state-of-the-art approaches, our model achieves better trade-off between preservation of semantic content and sentiment transfer. We also plot the correlations between the automatic metrics in Figure 2. The results clearly indicate that accuracy is negatively correlated with BLEU score, similarity measures and their corresponding masked scores.

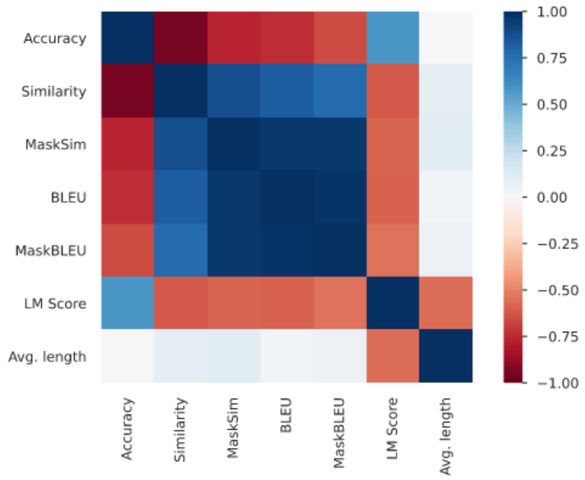

Figure 2: Correlations between all the automatic evaluation metrics. This figure indicates that accuracy is negatively correlated (value towards -1.0) with BLEU score, similarity measures and their corresponding masked scores. It also indicates LM Score negatively correlates with the average length of the sentence.

Our baseline models do not perform well in changing the sentiment even after adding style embedding and style loss. Using two separate decoders lead to major improvements on sentiment transfer over baseline methods. However, preservation of the content is very poor according to BLEU and similarity scores (and their polarity-masked equivalents). Using the pre-trained encoder has helped to improve the content preservation, but sentiment transfer accuracy degrades significantly.

The main motivation for our work was to find a denoising strategy which offers the best balance between sentiment transfer and content preservation. Our results suggest putting an emphasis on denoising high-polarity words results in the best ratio between the sentiment transfer accuracy and content preservation metrics.

Overall, our denoising approaches are able to balance well between sentiment transfer and content preservation. The models which perform the best on sentiment transfer usually achieve worse results on content preservation and similarity metrics.

For the human evaluation, we have chosen two models (*WG01-AG03-D* and *WG03P08-AG03P08-M*) which performed the best according to the average between accuracy, MaskSim and MaskBLEU score (Table 2). We have also chosen four state-of-the-art models for comparison: two of the most recent models (Wang et al., 2019; He et al., 2020), and the models with best accuracy (Prabhumoye et al., 2018) and MaskBLEU score (Li et al., 2018).

We have evaluated over 600 model outputs. Results are presented in Table 3. The human evaluation results mostly agree with our automatic evaluation results. The results also show that our models are better in content preservation than the competitor models.

Finally, to illustrate the behavior of different models, we picked one positive and one negative sentence from our sentiment dataset and the respective outputs from the models, which are shown in Table 4.

5 Conclusions and Future Work

In this paper, we proposed an approach for the text sentiment transfer task based on polarity-aware denoising. Experimental results on our sentiment dataset have shown that our method achieved a competitive or better performance compared to state-of-the-art approaches. While our extended transformer-based architecture provides more control for generating sentiment transferred outputs, at the same time polarity-aware enhanced denoising technique helps to achieve good style-content trade-off. As shown by both human evaluation scores and our manual inspection, our models still sometimes fail to preserve the meaning of the original. While we improve upon previous works in this respect, this still remains a limitation.

In the future, we plan to adapt our method to the different kind of style transfer tasks such as formality transfer or persona based text generation. We also intend to focus on better controlling content preservation with the use of semantic parsing.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Models	Sentiment	Content	Fluency
Prabhumoye et al. (2018)	3.95	1.19	3.56
Li et al. (2018)	3.35	2.3	3.34
(Wang et al., 2019)	3.48	1.67	2.54
(He et al., 2020)	3.69	1.66	3.26
Ours$_1$ (WG01-AG03-D)	**3.99**	2.56	**3.79**
Ours$_2$ (WG03P08-AG03P08-M)	3.94	**2.61**	3.73

Table 3: Human evaluation

	Negative → Positive	Positive → Negative
Source	**movie was a waste of money : this movie totally sucks .**	**my daughter loves them :)**
Prabhumoye et al. (2018)	stan is always a great place to get the food .	do n't be going here .
Li et al. (2018)	our favorite thing was a movie story : the dream class roll !	my daughter said i was still not acknowledged .
Wang et al. (2019)	movie is a delicious atmosphere of : this movie totally sucks movie !	i should not send dress after me more than she would said not ?
He et al. (2020)	this theater was a great place , we movie totally amazing .	yup daughter has left ourselves .
Ours$_1$ (WG01-AG03-D)	this movie is a great deal of money.	my daughter hated it .
Ours$_2$ (WG03P08-AG03P08-M)	movie : a great deal of money : this movie is absolutely perfect .	my daughter hates it : my daughter .
Source	**nothing truly interesting happens in this book .**	**best fit for my baby : this product is wonderful ! !**
Prabhumoye et al. (2018)	very good for the best .	bad customer service to say the food , and it is n't .
Li et al. (2018)	nothing truly interesting happens in this book .	my mom was annoyed with my health service is no notice .
Wang et al. (2019)	nothing truly interesting happens in this book make it casual and spot .	do not buy my phone : this bad crap was worst than it ?
He et al. (2020)	haha truly interesting happens in this book .	uninspired .
Ours$_1$ (WG01-AG03-D)	in this book is truly awesome .	not happy for my baby : this product is not great ! !
Ours$_2$ (WG03P08-AG03P08-M)	in this book is truly a really great book .	not good for my baby : this product is great ! ! ! ! ! ! ! !
Source	**the picture quality is horrible .**	**they love it too !**
Prabhumoye et al. (2018)	the selection of the food is delicious .	they did n't good .
Li et al. (2018)	the picture quality is superb !	horrible service .
Wang et al. (2019)	the best family always great offers delicious best enjoy always specials definitely best .	then they should n't charge leaving so it was n't gross as they ?
He et al. (2020)	picture boxes have good food .	ummm do n't bother .
Ours$_1$ (WG01-AG03-D)	the cast is awesome !	they didn ' t like this one .
Ours$_2$ (WG03P08-AG03P08-M)	picture quality quality is great .	! you feel also !

Table 4: Example outputs of different models on sentiment transfer task.

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

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *arXiv preprint arXiv:2002.03912*.

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

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2018. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

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*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Dekang Lin, Yuji Matsumoto, and Rada Mihalcea. 2011. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR.

Fuli Luo, Peng Li, Pengcheng Yang, Jie Zhou, Yutong Tan, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Towards fine-grained text sentiment transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2013–2022.

Eric Malmi, Aliaksei Severyn, and Sascha Rothe. 2020. Unsupervised text style transfer with padded masked language models. *arXiv preprint arXiv:2010.01054*.

Julia Neidert, Sebastian Schuster, Spence Green, Kenneth Heafield, and Christopher D Manning. 2014. Stanford university's submissions to the wmt 2014 translation task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 150–156.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

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 of the Association for Computational Linguistics*, pages 311–318.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. 2016. Personalized machine translation: Preserving original author traits. *arXiv preprint arXiv:1610.05461*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

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.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. *arXiv preprint arXiv:1908.09368*.

Martina Toshevska and Sonja Gievska. 2021. A review of text style transfer using deep learning. *IEEE Transactions on Artificial Intelligence*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. *Advances in Neural Information Processing Systems*, 32:11036–11046.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.