Text Style Transfer
Thesis Proposal

Sourabrata Mukherjee
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Prague, Czech Republic
mukherjee@ufal.mff.cuni.cz

Abstract

Text style transfer (TST) is one of the most important tasks of Natural Language Generation (NLG). The aim of *TST* is to automatically control the style attributes of text while preserving the content. We aim to expand the capabilities and increase the output quality of systems for NLG by controlling style attributes. In order to do that, we aim to present: (i) *TST* models which balance the style-content trade-off well, (ii) new models for real-world downstream applications, and (iii) better evaluation metrics for this task. In this thesis proposal, we first introduce the landscape of current approaches to *TST* and the background for our work. Next, we report the methodology and results of our experiments. Regarding future work, we outline our plans to continue to combat challenges in *TST*.

1 Introduction

The main goal of NLG is to automatically produce narratives that describe, summarize and explain the input data or text in a human-like manner. Some of the popular tasks of NLG include data-to-text generation (Kasner and Dušek, 2020), response generation of dialogue systems (Fan et al., 2020), paraphrase generation (Ma et al., 2018), and text summarization (Rush et al., 2017).

However, there are subtle attributes in the text, including *style*, that aren't controlled by default in most of these applications. This led to further research on control of the output from the text generation systems (Len et al., 2020; Hu et al., 2022). There is a large body of prior work in controllable text generation (Wang et al., 2018a; Young et al., 2018; Hu et al., 2022; Len et al., 2020). The aspects of text generation that are commonly controlled include topic (Dziri et al., 2019; Feng et al., 2018; Wang et al., 2018b; Xing et al., 2017), style (Li et al., 2018; Sudhakar et al., 2019; Prabhumoye et al., 2018; Chen et al., 2018), emotion (Fu et al., 2018; Kong et al., 2019; Sun et al., 2020; Zhou et al., 2018), and user preferences (Li et al., 2016a; Luan et al., 2017; Yang et al., 2018, 2017).

Text Style Transfer (TST), a subtask of controllable text generation, is a method that aims to control certain attributes of text while preserving its content as much as possible. Style attributes of a text can range from demographic attributes of a person writing the text such as personality (Li et al., 2016b), gender (Prabhumoye et al., 2018) to sentiment (Mukherjee et al., 2022), emotion (Zhou et al., 2018), or politeness (Niu and Bansal, 2018).

TST has gained significant attention thanks to the rise of deep neural models (Jin et al., 2022). However, *TST* task still requires deeper attention to its following challenges. Firstly, disentangling content and style in texts has proven to be very hard (see Section 2.3.2). Secondly, Text Style Transfer models can be developed in a supervised way with parallel corpora, i.e. text that comes in pairs with the same content but with different styles (Hu et al., 2022). However, most use cases do not have parallel data available (see Section 2.3.1), so *TST* on non-parallel corpora has become a prolific research area (see Section 2.5). Also, research in the area of style transfer for text is currently bottlenecked by a lack of standard evaluation practices (Mir et al., 2019). A successful style-transferred output not only needs to demonstrate the correct target style, but also needs to verify that it preserves the original semantics, and maintains fluency (see Section 2.3.3).

In this thesis proposal, we address the following research questions:

i. How to perform *TST* task without direct supervision (i.e., in case of the unavailability of the parallel data).

ii. How to ensure that the text generated by *TST* models balances the style-content trade-off by accurately controlling the style attributes and

preserving the style-independent content as much as possible.

iii. How to deal with the barriers of lack of training and evaluation datasets in *TST* tasks.

iv. How to evaluate the performance of *TST* systems.

v. How to build *TST-based* downstream applications.

To address these research questions (majorly i, ii, v) we have already conducted a few experiments (Section 3) and will continue further addressing these questions (especially iii, iv) in our future work (Section 4).

The thesis proposal is structured as follows: in Section 2, we lay out the background for our work, including the distinction between style and content (Section 2.1), challenges (Section 2.3), existing datasets (Section 2.4), existing *TST* approaches (Section 2.5), evaluation measures (Section 2.6) and applications (Section 2.7). In Section 3, we present our works so far. We set a plan for future research in Section 4. Finally, we summarize our work in Section 5.

2 The Task of Text Style Transfer

Text style transfer (*TST*) is an NLG task that aims to automatically control the style attributes of a text while preserving the style-independent content. Table 1 shows some basic examples of *TST*.

We have thoroughly studied the area's literature and compiled two survey papers: (1) a basic overview of *TST* (Mukherjee and Dusek, 2023), and (2) applications and ethical discussions of *TST* (Mukherjee et al., 2023b). The brief excerpts from these papers are presented in this section as an overview and background literature.

2.1 Style vs. Content Distinction

In (McDonald and Pustejovsky, 1985), style is defined as a notion that refers to the manner in which semantics is expressed. Style has also been defined in (Hovy, 1987) by its pragmatic aspects, which can be expressed as a variety of concepts, such as sentiment, emotion, humor, similes, personality, politeness, formality, simplicity, or authorship, which is generally expressed in the *TST* research as a variety of styles (Jin et al., 2022; Hu et al., 2022).

Content can also be understood as subject matter, theme, or topics the author writes about. Some of the *TST* tasks are built upon the assumption that style is localized to certain tokens in a text, and a token has either content or style information, but not both (Lee et al., 2021).

2.2 Problem Formulation

Given a text x with source style s_1, our goal is to rephrase x to a new text $\hat{x}$ with target style s_2 ($s_2 \neq s_1$) while preserving its style-independent content.

For example, we can consider the *Negative* $\rightarrow$ *Positive* instance from Table 1. Here, x is "The food is tasteless" and s_1 is *negative*. The *style-independent content* in x is "The food is". After sentiment transfer, $\hat{x}$ is "The food is delicious" and s_2 is *positive*.

2.3 Challenges

Modeling the style of text comes with a lot of challenges in practice, which are discussed in this section.

2.3.1 No Parallel Data

TST models could be trained with respect to parallel text from a given style or on non-parallel corpora (Hu et al., 2022). Parallel datasets consist of pairs of texts where each text in the pair expresses the same meaning but in a different style. Non-parallel datasets, on the other hand, have no paired examples and simply exist as two independent mono-style corpora. For parallel datasets, *TST* can be formulated in such a way that instead of translating between languages, one can translate between styles following machine translation (Hu et al., 2022). However, obtaining suitable, sufficient parallel data for each desired style attribute is the biggest challenge (for further discussion, see Section 2.5).

2.3.2 Style and content are hard to separate

Style transfer text generation implies the need to distinguish content from style (Jin et al., 2022). In some scenarios, the line between content and style can be blurry. Thus, it becomes very difficult to separate style from meaning. A probable reason for this situation is that the subject on which an author is writing can also influence their choice of words and style (Jin et al., 2022). This interweaving of the style and semantics makes *TST* challenging. However, total disentanglement is impossible without inductive biases or some other forms of supervision (Locatello et al., 2019).

	Source Style	Target Style
Impolite → Polite:	Shut up! the video is starting!	Please be quiet, the video will begin shortly.
Negative → Positive:	The food is tasteless.	The food is delicious.
Informal → Formal:	The kid is freaking out.	That child is distressed.

Table 1: *TST* examples regarding sentiment, polarity, and formality.

2.3.3 No Standard Evaluation Measures

Human evaluation is regarded as the best indicator of quality, but unfortunately, it is expensive, slow, and lacks reproducibility (Belz et al., 2020), making it an infeasible approach to use on a daily basis to validate model performance. For this reason, we often rely on automated evaluation metrics to serve as a cheap and quick proxy for human judgment. The trade-off between style transfer accuracy and content preservation poses a very big challenge for evaluating *TST* tasks. Automated metrics alone do not adequately identify this challenge (Mir et al., 2019). There is currently neither a standard set of evaluation practices nor a clear definition of which exact aspects to evaluate (Mir et al., 2019). Further discussion on evaluation measures is in Section 2.6.

2.4 Datasets

To evaluate the *TST* models many datasets have been proposed over the years. We discuss a few popular datasets by individual subtasks as follows:

- Politeness Transfer – A compiled dataset on politeness with automatically labeled instances from the raw Enron e-mail corpus (Shetty and Adibi, 2004) was presented by (Madaan et al., 2020).
- Sentiment Transfer – Transfer text's polarity from positive to negative or vice-versa. Some popular datasets proposed for this task: Yelp (Shen et al., 2017), Amazon (He and McAuley, 2016), and IMDb (Dai et al., 2019).
- Formality Transfer – Grammarly's Yahoo Answers Formality Corpus (GYAFC) is the largest human-labeled parallel dataset that was proposed for formality transfer tasks by (Rao and Tetreault, 2018).
- Author's Style Re-writing – Xu et al. (2012) collected a parallel dataset that captured line-by-line modern interpretations of 16 Shakespeare's plays.
- Genre Transfer – Li et al. (2018) collected a caption dataset where each sentence was labeled as factual, romantic, or humorous.

2.5 Overview of Approaches

In general, *TST* approaches can be classified based on the data used for training.

2.5.1 Sequence to Sequence

For situations where parallel data is available, like most supervised methods, a standard sequence-to-sequence model(Sutskever et al., 2014) with the encoder-decoder structure can be used (Hu et al., 2022). This process is similar to machine translation and text summarization. The encoder-decoder architecture can then be implemented by either LSTM (Hochreiter and Schmidhuber, 1997) or the Transformer (Vaswani et al., 2017) architecture. For example, Jhamtani et al. (2017) trained a sequence-to-sequence model on a parallel corpus and then applied the model to translate modern English phrases to Shakespearean English. We have the seq2seq approach for parallel data, but since it's rare (see Section 2.3.1), there are unsupervised methods that use non-parallel data for the *TST* task.

The methods on non-parallel data can broadly be divided into three unsupervised approaches:

2.5.2 Prototype Editing

This process works by deleting only the parts of the sentences which represent the source style and replacing them with words with the target style while making sure that the resulting text is still fluent. The advantage of this approach is its simplicity and explainability. For example, Li et al. (2018) found that parts of a text that are associated with the original style can be replaced with new phrases associated with the target style. Madaan et al. (2020) first calculate the ratio of mean TF-IDF between the two attribute corpora for each n-gram, then normalize this ratio across all possible n-grams, and finally mark those n-grams with a normalized ratio higher than a pre-set threshold as attribute markers. The text was then fed into a sequence-to-sequence model to generate a fluent

text sequence in the target style. However, these approaches are not suitable for *TST* applications where simple phrase replacement is not enough or a correct way to transfer style. The style marker retrieval might not work if the datasets have a confounded style and contents. This is because they may lead to the wrong extraction of style markers, such as some content words.

2.5.3 Latent-space Disentanglement

This is the process of disentangling text into its content and attribute in an embedding latent space and then applying generative modeling. *TST* models first learn the latent representations of the content and the style of the given text. The latent representation of the original content is then combined with the latent representation of the desired target style to generate text in the target style. Techniques like back-translation and adversarial learning (Shen et al., 2017; Zhao et al., 2018; Fu et al., 2018; Prabhumoye et al., 2018; Hu et al., 2017) have been proposed to disentangle latent representations in content and style. As noted in Section 2.3.2, disentanglement is hard and these approaches typically have problems preserving content.

2.5.4 Pseudo-Parallel Corpus Creation

This process is used to train the model in a supervised way by generating pseudo-parallel data. One way of constructing pseudo-parallel data is through retrieval, i.e., extracting aligned sentence pairs from two mono-style corpora. Jin et al. (2019) constructed pseudo-parallel corpora by matching sentence pairs in the two style-specific corpora according to the *cosine similarity* of pre-trained sentence embeddings. Many other data augmentation methods have been explored by researchers (Shang et al., 2019; Jin et al., 2019; Nikolov and Hahnloser, 2018; Liao et al., 2018). Nevertheless, these approaches are limited by the lack of systematic evaluation of the generated pseudo-parallel datasets. The constructed pseudo-parallel corpora must reach a certain level of quality to be useful for *TST*.

2.6 Evaluation Measures

A successful style transfer output is one that portrays the correct target style along with preserving the original semantics of the text and maintaining natural language fluency.

2.6.1 Automatic Evaluation

It is the process that provides an economic, reproducible, and scalable way to assess the quality of generation results. There are several automated evaluation metrics proposed to measure the effectiveness of *TST* models (Pang, 2019b,a; Pang and Gimpel, 2019; Mir et al., 2019).

Style Transfer Accuracy The ability to transfer the style is measured using Style Transfer Accuracy (Hu et al., 2017; Shen et al., 2017; Fu et al., 2018; Luo et al., 2019b; John et al., 2019). Mostly, a binary style classifier (Moschitti et al., 2014) is pretrained separately to predict the style label of the input sentence and is then used to estimate the style transfer accuracy of the transferred style sentence. This is done by considering the target style as the ground truth.

Content Preservation In order to measure the amount of original content preserved after the style transfer procedure, some automated evaluation metrics from other NLG tasks have been adopted for *TST*. In *TST*, *BLEU* (Papineni et al., 2002) is computed the same as with machine translation, when parallel *TST* datasets or human references are available. Since most of the *TST* tasks assume a non-parallel setting and matching references of style transferred sentences are not always a feasible option, the evaluation metric *source-BLEU (sBLEU)* is adopted. In this method, a transferred sentence is compared to its source. The n-gram overlap with the source is considered a proxy for content preservation (Madaan et al., 2020). Cosine Similarity (Rahutomo et al., 2012) can also be calculated between the original sentence embeddings and transferred sentence embeddings (Fu et al., 2018). This methodology follows the idea that the embeddings of the two sentences should be close if most semantics is preserved.

Fluency One of the common goals for all NLG tasks is producing fluent outputs. A common approach to measuring the fluency of a sentence is using a language model (Kneser and Ney, 1995). In *TST* tasks, a pre-trained language model is used to compute the perplexity scores of the style-transferred sentences to evaluate the sentences' fluency (Mukherjee et al., 2022).

2.6.2 Human Evaluation

Human evaluation stands out from automatic evaluation due to its flexibility and comprehensiveness.

However, this evaluation approach is very challenging since the interpretation of text style can be subjective and vary from individual to individual (Pang, 2019b,a; Mir et al., 2019). In terms of evaluation types, there is point-wise scoring, wherein humans are asked to provide absolute scores of the model outputs, and pairwise comparison, wherein they are asked to judge which of the two outputs is better, or by providing a ranking for multiple outputs (Briakou et al., 2021). Human evaluations offer valuable insights into how well the *TST* algorithms can transfer style and generate sentences that are acceptable according to human standards. However, in *TST*, human evaluations are often underspecified and not standardized, which hampers the reproducibility of research in this field (Briakou et al., 2021).

2.7 Applications of TST

TST has a wide range of downstream applications in various NLP fields. A few popular examples are discussed below.

Chatbots (Kim et al., 2019) carried out a study that showcased the impact of chatbot's conversational style on users. (Li et al., 2016c) encoded personas of individuals in contextualized embeddings that helped in capturing the background information and style to maintain consistency in the generated responses. Firdaus et al. (2022) focused on generating polite personalized dialog responses in agreement with the user's profile and consistent with their conversational history.

Writing Aids Another important application of *TST* is enhancing the human writing experience (Can and Patton, 2004; Johnstone, 2009; Ashok et al., 2013). This application aids in people restyling their content to appeal to a variety of audiences, i.e., making a text polite, humorous, professional, or even Shakespearean.

Text Simplification Another inspiring application of *TST* is to automatically simplify content for better communication between experts and non-expert individuals in certain knowledge domains, thus lowering language barriers. For example, complicated legal, medical, or technical jargon is transferred into simple terms that a layman can comprehend (Cao et al., 2020).

Safety *TST* can also offer a means to neutralize subjective attitudes for certain texts where objectivity is strongly needed. For example, in the domains of news, encyclopedia, and textbooks. Such applications help in reshaping gender roles that are portrayed in writing (Clark et al., 2018). *TST* can also help in transforming hateful sentences into non-hateful ones. (Santos et al., 2018) propose an extension of a basic encoder-decoder architecture by including a collaborative classifier to deal with abusive language redaction.

3 Our works so far

In this section, we describe our experiments: (1) *Sentiment Transfer* (Mukherjee et al., 2022) and (2) *Polite Chatbot: A* TST *Application* (Mukherjee et al., 2023a). We focus here on research questions (see Section 1) (i) and (ii) in Section 3.1 and (v) in Section 3.2.

3.1 Sentiment Transfer

Text sentiment transfer is a sub-task of *TST*, which aims to flip the sentiment polarity of a sentence (positive to negative or vice versa) while preserving its sentiment-independent content (Section 2.4). We present a sentiment transfer model based on polarity-aware denoising, which accurately controls the sentiment attributes in generated text, preserving the content to a great extent and helping to balance the style-content trade-off. Our proposed model is structured around two key stages in the sentiment transfer process: better representation learning using a shared encoder and sentiment-controlled generation using separate sentiment-specific decoders.

3.1.1 Model Overview

Figure 1 shows the overview of our proposed architecture. Following Prabhumoye et al. (2018), we first translate the input text x in the base language to a chosen intermediate language $\bar{x}$ using a translation model.[1] Next, we prepare a noisy text x_{noise} from $\bar{x}$ using *polarity-aware noising* (described below) with word deletion or masking probabilities θ_N :

$$x_{noise} = Noise(\bar{x}; \theta_N). \quad (1)$$

We provide x_{noise} to the encoder of the $\bar{x} \rightarrow \hat{x}$ back-translation model (where $\hat{x}$ is a text in the base language with changed sentiment polarity). The encoder first converts the text to the latent representation z as follows:

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

[1]We work with English as the base language and German as an intermediate language.

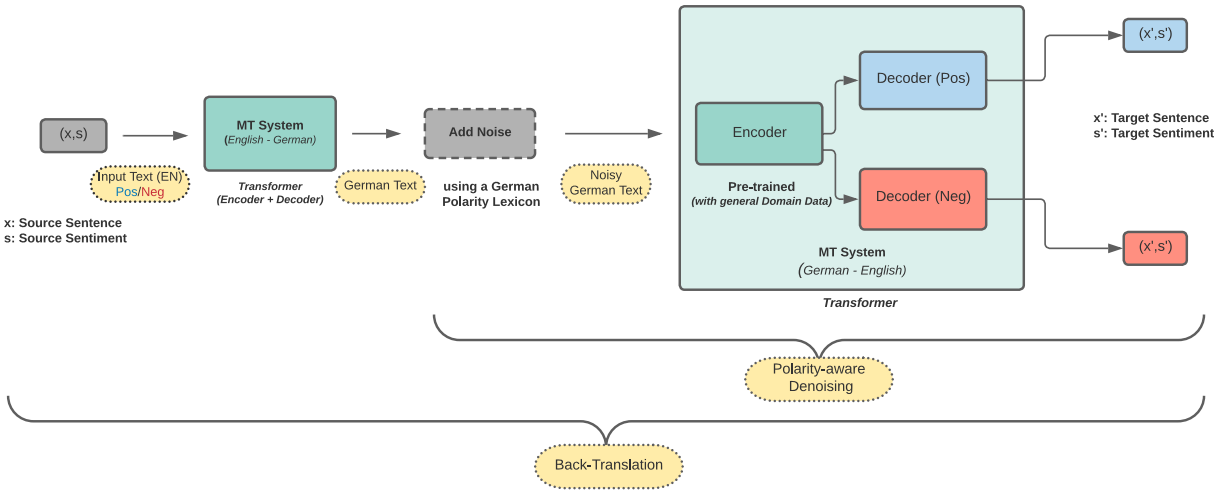

Figure 1: Our sentiment transfer pipeline. In the pipeline, we (1) *translate* the source sentence from English to German using a transformer-based machine translation (MT) system; (2) *apply noise* on the German sentence using a German polarity lexicon; (3) *encode* the German sentence to latent representation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

where θ_E represents the parameters of the encoder.

Two separate sentiment-specific decoders are trained to decode the original positive and negative inputs by passing in their latent representations z:

$$x_{pos} = Decoder_{pos}(z; \theta_{D_{pos}}) \quad (3)$$

$$x_{neg} = Decoder_{neg}(z; \theta_{D_{neg}}). \quad (4)$$

At inference time, sentiment transfer is achieved by decoding the shared latent representation using the decoder trained for the opposite sentiment, as follows:

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

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

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.

A pure back-translation approach (without any specific provisions for sentiment) is referred to as *Back-Translation* in our experiments. In addition to a pure back-translation model, we present several straightforward baselines. We have the back-translation with a style-identified token (*Style Tok*), then two separate models for both sentiment (*Two Sep. transformers:*), then shared encoders and separate decoders (*Shrd Enc + Two Sep Decoders*), then a pretrained encoder (*Pre Training Enc*) setup.

Polarity-Aware Denoising We devise a task-specific pre-training (Gururangan et al., 2020) scheme for improving the sentiment transfer abilities of the model. The idea of our pre-training scheme—*polarity-aware denoising*—is to first introduce noise, i.e. delete or mask a certain proportion of words in the intermediate German input to the back-translation step, then train the model to remove this noise, i.e. produce the original English sentence with no words deleted or masked. To decide which words get deleted or masked, we use automatically obtained sentiment polarity labels. This effectively adds more supervision to the task on the word level. We apply three different 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) both *general* and *polarity* words (each with a different probability).

We use polarity-aware denoising during encoder pretraining, following the shared encoder and separate decoder setup. The encoder is further fine-tuned during the sentiment transfer training.

3.1.2 Datasets

For experimental evaluation, we built a new English dataset for sentiment transfer, 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 sentiment transfer datasets. We also evaluated our models using IMDb Dataset (Dai et al., 2019) (for details, please refer (Mukherjee et al., 2022)).

3.1.3 Evaluations

Metrics To evaluate the performance of the models, we compare the generated samples along three different dimensions using automatic metrics: (1) style control, (2) content preservation, and (3) fluency.

Following prior work (see Section 2.6), we measure sentiment accuracy automatically by evaluating the sentiment of transferred sentences. We use a pre-trained transformer-based sentiment analysis pipeline[2] from Huggingface (Wolf et al., 2020).

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

Following previous work, we compute *BLEU* score (Papineni et al., 2002) between the transferred sentence and its source. Higher *BLEU* indicates higher n-gram overlap between the sentences, which is generally viewed as a proxy for content preservation. We also compute Sentence BERT (Reimers and Gurevych, 2019a) 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 the 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.

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 pivot words (based on NLTK Vader sentiment dictionary (Hutto and Gilbert, 2014)) have been masked. This allows measuring content preservation while ignoring the parts of the sentences that need to be changed.

Results Results on our Amazon Review data are shown in Table 2. Overall, there is clearly a tradeoff between preserving sentiment-independent content and achieving the desired target sentiment. Models which perform very well in sentiment transfer usually achieve worse results on content preservation.

The translation-only and style token baselines do not perform well in changing the sentiment. Using two separate decoders leads to major sentiment transfer improvements, but content preservation is poor. Using the pre-trained encoder has helped to improve content preservation, but sentiment transfer accuracy degrades significantly.

The main motivation for our work was to find a denoising strategy that 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. Additionally, our models show the ability to produce fluent sentences, as shown by the language model score: our models' scores are similar to the back-translation baseline; other models only reach higher language model scores when producing very short outputs.

Overall, our denoising approaches are able to balance well between sentiment transfer and content preservation. On content preservation, they perform much better than state-of-the-art models, and they stay competitive on style accuracy.

Human Evaluation As automated metrics for language generation do not correlate well with human judgments (Novikova et al., 2017), we conduct an in-house human evaluation with five expert annotators. We randomly select 100 sentences (50 for each sentiment) from our Amazon Review test set. The annotators rate model outputs using a 1-5 Likert scale for style control, content preservation, and fluency. We compare our best SCT_1 and SCT_2 models (selected above) with four state-of-the-art models: two of the most recent models (Wang et al., 2019; He et al., 2020), and the models with the 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.

We further examined a sample of the outputs in more detail to understand the behavior of different models. We found that state-of-the-art models tend to lose the content of the source sentence, as shown in the example outputs in Table 4. On the

[2] https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english

Table 2: Sentiment Transfer Task (Section 3.1): Automatic evaluation. *Accuracy*: Sentiment transfer accuracy. *Sim* and *B*: *Cosine similarity* and *BLEU* score between input and sentiment-transferred sentence. *M/Sim* and *M/B*: MaskSim and MaskBLEU (*similarity* and *BLEU* with polarity words masked). *LM*: Average log probability assigned by vanilla GPT-2 language model. *Len*: Average length of transferred sentences. *Avg*: Average of sentiment transfer accuracy, 100*MaskSim and MaskBLEU. Scores are based on a single run, with identical random seeds. The first two sections show our own baselines, and the third section shows our models with denoising (here in this thesis we only present the best settings denoted SCT_1 and SCT_2, other settings and the corresponding results are reported in (Mukherjee et al., 2022)). The bottom section shows a comparison with state-of-the-art models. Names of models with denoising reflect settings as follows: *W* denotes WMT pretraining data, *A* denotes Amazon finetuning data; the following tokens denote noise probability values associated with the respective data. *G/P* represents general/polarity token noising, *D/M* represents noising mode deletion/masking. E.g. *WG01-AG03-D*: noise probabilities on WMT and Amazon data are identical, noising by a deletion on general token noising is applied (with probabilities 0.1 and 0.3, respectively). *WG03P08-AG03P08-M*: noise probabilities on WMT and Amazon data are identical, noising by masking on both general and polarity token noising is applied (with probabilities 0.3 and 0.8, respectively).

Models	**Acc**	**Sim**	**M/Sim**	**B**	**M/B**	**LM**	**Len**	**Avg**
Back-Translation Only								
Back-translation only	0.4	0.828	0.768	28.0	45.3	-78.6	11.9	40.9
Our Baseline Models								
Style Tok	13.2	0.536	0.560	4.8	8.6	-52.1	7.6	25.9
Two Sep. transformers	89.3	0.394	0.611	6.8	19.6	-79.0	13.7	56.7
Shrd Enc + Two Sep Decoders	88.1	0.397	0.600	7.3	20.1	-78.0	12.5	56.0
Pre Training Enc	55.3	0.592	0.732	22.6	33.9	-93.3	13.4	54.1
Our Models (with Denoising)								
WG01-AG03-D (=SCT_2)	85.2	0.441	0.646	11.8	25.4	-79.8	13.1	58.4
WG03P08-AG03P08-M (=SCT_1)	82.0	0.460	0.665	13.7	27.4	-79.6	12.8	**58.6**
State-of-the-Art Models								
Shen et al. (2017)	88.6	0.346	0.513	3.2	18.3	-74.0	10.9	52.7
Li et al. (2018)	69.9	0.457	0.632	14.7	25.3	-85.1	12.2	52.8
Luo et al. (2019a)	92.4	0.279	0.468	0.0	9.1	-42.0	7.8	49.4
Prabhumoye et al. (2018)	93.5	0.308	0.504	0.9	15.2	-61.0	10.3	53.0
Wang et al. (2019)	79.3	0.385	0.545	10.6	20.3	-116.8	15.1	51.4
He et al. (2020)	91.5	0.352	0.542	9.5	21.8	-65.9	8.2	55.8

other hand, our models mostly preserve sentiment-independent content well while successfully transferring the sentiment. We conclude that with our models, there is a good balance between preserving the original sentiment-independent content and dropping the source sentiment, and existing state-of-the-art models tend to sacrifice one or the other.

3.2 Polite Chatbot: A *TST* Application

Building a chatbot agent that produces stylized and coherent responses can yield more engaging conversations (Niederhoffer and Pennebaker, 2002). Generating stylized dialogue responses has been investigated in various studies, with a broad understanding of style covering emotion (Zhou et al., 2018), personality (Li et al., 2016b) or politeness (Niu and Bansal, 2018). proposes a polite chatbot that can produce responses that are polite and coherent to the given context. In this study, a politeness transfer model is first used to generate polite synthetic dialogue pairs of contexts and polite utterances. Then, these synthetic pairs are employed to train a dialogue model. Automatic and human evaluations demonstrate that our method outperforms baselines in producing polite dialogue responses while staying competitive in terms of coherent to the given context.

3.2.1 Method

Our method consists of three steps (Figure 2). First, we train a politeness transfer model. Our goal here is to train a model that takes as input a neutral sentence x and outputs a sentence $\hat{x}$ that retains the content while increasing politeness. Second, we apply this politeness transfer model to generate synthetic polite chat data. Finally, we use the corpus $\hat{\mathcal{D}}$ to train a dialogue model.

Politeness Transfer Model Although we do not have parallel corpora available for politeness transfer, our transfer model is trained in a supervised

Table 3: Sentiment Transfer Task (Section 3.1): Human evaluation of sentiment transfer quality, content preservation, and fluency. Average of 1-5 Likert scale ratings on 100 examples from our Amazon Review data.

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
SCT_1 (WG03P08-AG03P08-M)	3.94	**2.61**	3.73
SCT_2 (WG01-AG03-D)	**3.99**	2.56	**3.79**

Table 4: Sentiment Transfer Task (Section 3.1): Example outputs comparison on samples from our Amazon Reviews dataset. Sentiment marker words (pivots) are colored. Note that our models preserve content better than most others.

	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 .
SCT_1 (WG03P08-AG03P08-M)	movie : a great deal of money : this movie is absolutely perfect .	my daughter hates it : my daughter .
SCT_2 (WG01-AG03-D)	this movie is a great deal of money.	my daughter hated it .

fashion on synthetic input-output pairs. These are obtained following Madaan et al. (2020): polite phrases (politeness markers) are identified using TF-IDF over polite and non-polite texts.[3] The markers are removed from polite texts on the input, and a sequence-to-sequence model is trained to increase sentence politeness by reconstructing the politeness markers on the output. Unlike Madaan et al. (2020), we do not use separate tagging and generation steps here and join the task into a single step. Specifically, we finetune a pre-trained language model for this task using standard cross-entropy loss.

Creating Synthetic Polite Data We apply our politeness transfer model to a dataset consisting of N dialogues $\mathcal{D} = \{C_1^{k_1}, ..., C_N^{k_N}\}$, where dialogue $C_i^{k_i}$ consists of k_i utterances $\{u_i^1, ..., u_i^{k_i}\}$. We create a corpus of context-utterance pairs $\hat{\mathcal{D}} = \{\langle C_1^1, \hat{u}_1^2\rangle, \langle C_1^2, \hat{u}_1^3\rangle, ..., \langle C_N^{K_N-1}, \hat{u}_N^{K_N}\rangle\}$. In other words, for every partial context, we add a polite version of the next utterance.

Dialogue Model We use a standard dialogue response generation model that produces a dialogue utterance u_i based on context $\mathbf{C} = \{u_1, ..., u_{i-1}\}$, trained using cross-entropy loss. We experiment with multiple pre-trained language models here. To achieve politeness in responses, we use the synthetic polite dialogue corpus $\hat{\mathcal{D}}$ obtained using our politeness transfer model.

[3]In principle, a much higher mean TF-IDF value over polite than non-polite texts means that a phrase is likely to be a politeness marker.

3.2.2 Datasets

Politeness Transfer We use the dataset of Madaan et al. (2020), i.e. preprocessed and filtered sentences from the Enron e-mail dataset (Shetty and Adibi, 2004) into ten buckets (P_0-P_9) based on the score of a politeness classifier by Niu and Bansal (2018). We use Madaan et al. (2020)'s TF-IDF-based approach to remove politeness markers from the sentences in the most polite P_9 bucket to prepare synthetic parallel data for training our politeness transfer models.

Dialogue To train our response generation models, we use DailyDialog (Li et al., 2017), an open-domain dataset of 13,118 human-human dialogues with 7.9 turns per dialogue on average.

3.2.3 Evaluations

We evaluate all dialogue models against three baselines: (1) vanilla version of the model, (2) model fine-tuned on unchanged DailyDialog data, (3)

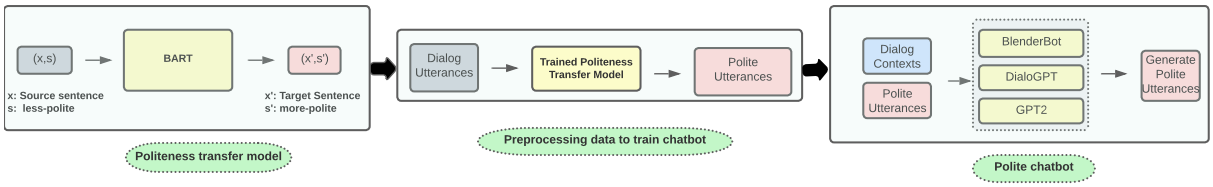

Figure 2: Our polite chatbot method: We (1) train the politeness transfer model; (2) generate synthetic training data by applying the transfer model to neutral utterances; (3) train the dialogue models using the synthetic data.

model finetuned on synthetic polite DailyDialog data generated in the same fashion as in our full model, but using Madaan et al. (2020)'s politeness transfer instead of ours.

Metrics Following prior work (Madaan et al., 2020; Niu and Bansal, 2018), we use automatic metrics for the evaluation of the models along two major dimensions: (1) style transfer and (2) content preservation and relevance. To measure politeness transfer quality, we compute *Polite Score*, which is defined as the average score given to the generated sequences by our politeness classifier, which we created by finetuning BERT (Devlin et al., 2019) on Madaan et al. (2020)'s Enron data (see Section 3.2.2).[4] Following prior work (Jin et al., 2022; Hu et al., 2022), we evaluate the relevance and content preservation using *embedding similarity* (Rahutomo et al., 2012) and *BLEU* score (Papineni et al., 2002). For *embedding similarity*, we use a pre-trained Sentence-BERT model (Reimers and Gurevych, 2019b) and *cosine similarity*. We use *BLEU-1* and *BLEU-2* to account for the expected different phrasing in polite outputs and the high output variance common to open-domain dialogue response generation.

Results Results of automatic metrics for dialogue modeling are shown in Table 5. The performance differences between the pre-trained models used are expected given the models' properties and intended use cases. While GPT-2 scores low on politeness, the dialogue-specific models obtain better results. As expected, all models perform much better in terms of content preservation after finetuning. Both ours and Madaan et al.'s politeness transfer result in an increase in politeness, and we can observe that our method consistently outperforms Madaan et al.'s. Moreover, our method is the only one that improves the Polite Score over the vanilla BlenderBot model. Finally, although the application of politeness transfer causes a decrease in content similarity with reference responses from DailyDialog, the drop is marginal, not consistent with all metrics, and could be caused by different phrasing.

Human Evaluation We have evaluated 50 model outputs for each variant of the BlenderBot model. The results are presented in Table 6. The human evaluation results mostly agree with our automatic evaluation results: our data preparation method performs better than Madaan et al. (2020)'s transfer in terms of politeness and is able to improve the base BlenderBot model. Both politeness-increasing methods cause a slight degradation in context coherency of the generated utterances; ours performs slightly worse in this aspect. However, our full approach yields more fluent outputs than the model trained on Madaan et al. (2020)'s politeness transfer.

We further examined a sample of the outputs in more detail to understand the behavior of different models in Table 7.

4 Future Work

In this section, We plan to address the research questions (see Section 1) (iii) in Section 4.1 and (iv) in Section 4.2:

4.1 Fixing multi-biases in multi-lingual settings

The use of biased language especially offensive and hateful language is a common problem of abusive behavior on online social media networks. The aim of our task is to fix biased text that contains words or phrases that are offensive, prejudiced, excluding, or hurtful. The challenge that arises, in this case, is the unavailability of parallel data (see Section 2.3.1). We plan to present a multi-lingual corpus of manually annotated *biased to neutral* 5k sentence pairs.

[4] Although the scale of politeness classes is not necessarily linear, we believe that this is still a good indicator of the overall politeness of the data.

	BlenderBot				DialoGPT				GPT-2			
Finetuned on	PS	BLEU-1,2		CS	PS	BLEU-1,2		CS	PS	BLEU-1,2		CS
Vanilla (no FT)	7.06	9.80	2.58	20.31	6.31	9.38	1.98	19.33	4.91	0.15	0.09	8.31
DailyDialog (DD)	7.11	17.21	7.25	45.80	6.14	11.72	2.60	38.44	5.08	7.82	2.13	29.72
DD + Madaan et al. (2020)	6.75	17.16	6.73	45.17	6.17	11.47	2.19	35.08	5.99	7.32	1.49	27.42
DD + Ours	**7.65**	17.03	6.85	41.80	**7.75**	11.44	2.57	35.03	**7.20**	5.65	1.03	26.80

Table 5: Polite Chatbot (Section 3.2): Evaluation results of polite dialog models. We indicate what version of the DailyDialog dataset (DD) was used for Finetuning (FT) if any. We measure the Polite Score (PS), BLEU score, and Content Similarity (CS). BLEU Score (of n-gram = 1,2) and CS are computed between predicted polite utterances and the original utterances.

BlenderBot finetuned on	Pol	CC	Flu
Vanilla (no FT)	3.46	1.16	4.64
DailyDialog (DD)	3.90	3.74	4.54
DD + Madaan et al. (2020)	3.50	3.06	3.98
DD + Ours	**4.26**	2.94	4.30

Table 6: Polite Chatbot (Section 3.2): Human Evaluation on BlenderBot outputs. We measured politeness (Pol), coherent to context (CC), and fluency (Flu).

We will start with utilizing a biased multilingual dataset presented by Kumar et al. (2021). The dataset consisted of four languages, namely, Meitei, Bangla, Hindi, and Indian English, and was collected from various social media platforms. In this work, a parallel dataset will be prepared using crowdsourcing by changing the tone of the biased texts to a more neutral mood, without altering the content of the text. To ensure quality control, experts, who will be part of this work, will review the rewrites of the workers. Further, we will experiment using two kinds of approaches: (i) models similarly used in low-resource MT (Haddow et al., 2022), and (ii) finetuning generative language models using our corpus.

4.2 Evaluating the content preservation for *TST*

The existing evaluation metrics for *TST* are relatively limited (Pang, 2019b; Pang and Gimpel, 2018; Mir et al., 2019). The main reason behind this is the entanglement between the semantic and stylistic properties of natural language (see Section 2.3.3). If the *TST* model transfers text from one style to another but omits or changes an important piece of information, it fails to preserve the meaning of the original text. On the flip side, if the model reproduces the source text exactly as is, it would have perfect content preservation, but fail completely in style transfer.

To evaluate content preservation more precisely, attempts have been made to first distinguish between semantic and stylistic components of text, and then meaningfully quantify the similarity of just the semantic component alone. While there is an open debate about whether it's possible to actually decouple style from content in free text, intuition leads us to believe that style attribute is expressed, at least in part, through select words (Sudhakar et al., 2019; Madaan et al., 2020).

In this task, we first aim to calculate sentence-level feature importance by applying a popular model interpretability technique called Integrated Gradients (Sundararajan et al., 2017) to our fine-tuned BERT style classifier which explains a model's prediction in terms of its features (Sudhakar et al., 2019). This method produces word attributions, which are essentially importance scores for each token in a sentence that indicate how much of the prediction outcome is attributed to that token. This approach allows us to selectively mask style-related tokens depending on their function within a particular sentence.

As some *TST* methods may also produce more complex rephrasings which also affect the style-independent part of the sentence (and make it hard to use word-based metrics for content preservation), we further plan to introduce another evaluation technique focused on understanding the underlying meaning (and whether it is preserved). To better understand the semantics of a sentence, requires breaking the linguistic input into a meaningful structure. Thus, we will calculate the semantic parsing and prepare semantic representations from the masked style-free source and transferred sentences. We then plan to provide a details analysis by comparing these structures(of source and transferred sentences) utilizing metrics such as SEMMATCH (Cai and Knight, 2013), and SEMBLEU (Song and Gildea, 2019).

Context : Nani Book Store. How can I help you? Do you have the 'The man and the Sea' by Ernest Hemingway?	
Ground Truth	**Yes, we have one copy left.**
Vanilla (no FT)	I would love to try it, but I don't think I'll be able to afford it.
DailyDialog (DD)	Yes, it ' s on that shelf. I ' ll get it for you right away.
DD + Madaan et al. (2020)	We do. Would you like a look at it?
DD + Ours	Yes, we do. Would you like to have a look at it? please let me know.

Table 7: Polite Chatbot (Section 3.2): A sample output using BlenderBot. Vanilla BlenderBot produces polite but irrelevant responses, and models finetuned on all DailyDialog data versions produce relevant responses, but ours is arguably the most polite.

In addition, we will further do the human evaluations to judge how well the results of automatic evaluation relate to human evaluation.

5 Conclusion

The thesis proposal described challenges in *TST* task. We showed how a *TST* model can learn and balance well between style transfer and content preservation in the absence of parallel data and also how the performance can be measured well using automatic metrics (see Section 3.1). We designed a sentiment transfer model using an extended transformer architecture and polarity-aware denoising. In addition to this, We introduced polarity-masked BLEU (MaskBLEU) and similarity score (MaskSim) for automatic evaluation of content preservation in this task. In order to build a *TST-based* downstream real-world application, we proposed a polite chatbot that can produce responses that are polite and coherent to the given context (see Section 3.2). In the future, we will continue to find ways to deal with the challenges in *TST* tasks. We plan to follow up on our experiments by introducing innovative automatic evaluation measures for the *TST* task (see Section 4.2). Further, we will present a manually annotated *biased to neutral* parallel corpus and a set of benchmark models to fix the biased texts on online social media networks (see Section 4.1).

References

Vikas Ganjigunte Ashok, Song Feng, and Yejin Choi. 2013. Success with style: Using writing style to predict the success of novels. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1753–1764.

Anya Belz, Shubham Agarwal, Ehud Reiter, and Anastasia Shimorina. 2020. Reprogen: Proposal for a shared task on reproducibility of human evaluations in nlg.

Eleftheria Briakou, Sweta Agrawal, Ke Zhang, Joel Tetreault, and Marine Carpuat. 2021. A review of human evaluation for style transfer. *arXiv preprint arXiv:2106.04747*.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752.

Fazli Can and Jon M Patton. 2004. Change of writing style with time. *Computers and the Humanities*, 38(1):61–82.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Liqun Chen, Shuyang Dai, Chenyang Tao, Haichao Zhang, Zhe Gan, Dinghan Shen, Yizhe Zhang, Guoyin Wang, Ruiyi Zhang, and Lawrence Carin. 2018. Adversarial text generation via feature-mover's distance. In *Advances in Neural Information Processing Systems*, pages 4666–4677.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd International Conference on Intelligent User Interfaces*, pages 329–340.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Yifan Fan, Xudong Luo, and Pingping Lin. 2020. A survey of response generation of dialogue systems. *International Journal of Computer and Information Engineering*, 14(12):461–472.

Xiaocheng Feng, Ming Liu, Jiahao Liu, Bing Qin, Yibo Sun, and Ting Liu. 2018. Topic-to-essay generation with neural networks. In *IJCAI*, pages 4078–4084.

Mauajama Firdaus, Arunav Shandilya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Being polite: Modeling politeness variation in a personalized dialog agent. *IEEE Transactions on Computational Social Systems*.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Suchin Gururangan, Ana Marasovic, 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 *Proc. ACL*, pages 8342–8360.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C Aggarwal, and Aston Zhang. 2022. Text style transfer: A review and experimental evaluation. *ACM SIGKDD Explorations Newsletter*, 24(1):14–45.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1587–1596. JMLR. org.

Clayton J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In *Proc. ICWSM*.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence to sequence models. In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. Imat: Unsupervised text attribute transfer via iterative matching and translation. 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 3088–3100.

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

Barbara Johnstone. 2009. Stance, style, and the linguistic individual. *Stance: sociolinguistic perspectives*, pages 29–52.

Zdeněk Kasner and Ondřej Dušek. 2020. Data-to-text generation with iterative text editing. *arXiv preprint arXiv:2011.01694*.

Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184 vol.1.

Xiang Kong, Bohan Li, Graham Neubig, Eduard Hovy, and Yiming Yang. 2019. An adversarial approach to high-quality, sentiment-controlled neural dialogue generation. *arXiv preprint arXiv:1901.07129*.

Ritesh Kumar, Enakshi Nandi, Laishram Niranjana Devi, Shyam Ratan, Siddharth Singh, Akash Bhagat, and Yogesh Dawer. 2021. The comma dataset v0. 2: Annotating aggression and bias in multilingual social media discourse. *arXiv preprint arXiv:2111.10390*.

Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin L Zhang. 2021. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. *arXiv preprint arXiv:2108.00449*.

Yuanmin Len, François Portet, Cyril Labbé, and Raheel Qader. 2020. Controllable neural natural language generation: comparison of state-of-the-art control strategies. In *WebNLG+: 3rd Workshop on Natural Language Generation from the Semantic Web*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and William B. Dolan. 2016c. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and 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)*, pages 1865–1874.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Yi Liao, Lidong Bing, Piji Li, Shuming Shi, Wai Lam, and Tong Zhang. 2018. Quase: Sequence editing under quantifiable guidance. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3855–3864.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2019. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proc. ICML*, volume 97, pages 4114–4124.

Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-task learning for speaker-role adaptation in neural conversation models. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 605–614, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Fuli Luo, Peng Li, Pengcheng Yang, Jie Zhou, Yutong Tan, Baobao Chang, Zhifang Sui, and Xu Sun. 2019a. Towards Fine-grained Text Sentiment Transfer. In *Proc. ACL*, pages 2013–2022.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019b. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5116–5122. AAAI Press.

Shuming Ma, Xu Sun, Wei Li, Sujian Li, Wenjie Li, and Xuancheng Ren. 2018. Query and output: Generating words by querying distributed word representations for paraphrase generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 196–206, New Orleans, Louisiana. Association for Computational Linguistics.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. *arXiv preprint arXiv:2004.14257*.

David D McDonald and James Pustejovsky. 1985. A computational theory of prose style for natural language generation. In *Second Conference of the European Chapter of the Association for Computational Linguistics*.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 2019. Evaluating style transfer for text. 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)*, pages 495–504.

Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors. 2014. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL.

Sourabrata Mukherjee and Ondrej Dusek. 2023. A basic introduction to text style transfer. Manuscript under review at 4EU+ International Workshop on Recent Advancements in Artificial Intelligence 2023.

Sourabrata Mukherjee, Vojta Hudecek, and Ondrej Dusek. 2023a. Polite chatbot: A text style transfer application. Manuscript under review at EACL-SRW 2023.

Sourabrata Mukherjee, Zdeněk Kasner, and Ondřej Dušek. 2022. Balancing the style-content trade-off in sentiment transfer using polarity-aware denoising. In *International Conference on Text, Speech, and Dialogue*, pages 172–186. Springer.

Sourabrata Mukherjee, Zdenek Kasner, and Ondrej Dusek. 2023b. Text style transfer: Applications and ethical impact. Manuscript under review at EACL-SRW 2023.

Jianmo Ni, Jiacheng Li, and Julian J. McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proc. EMNLP-IJCNLP*, pages 188–197.

Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Nikola I. Nikolov and Richard H. R. Hahnloser. 2018. Large-scale hierarchical alignment for author style transfer. *CoRR*, abs/1810.08237.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

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

Richard Yuanzhe Pang. 2019a. The daunting task of real-world textual style transfer auto-evaluation. *CoRR*, abs/1910.03747.

Richard Yuanzhe Pang. 2019b. Towards actual (not operational) textual style transfer auto-evaluation. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 444–445.

Richard Yuanzhe Pang and Kevin Gimpel. 2018. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. *arXiv preprint arXiv:1810.11878*.

Richard Yuanzhe Pang and Kevin Gimpel. 2019. Unsupervised evaluation metrics and learning criteria for non-parallel textual transfer. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 138–147.

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*.

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

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, volume 4, page 1.

Sudha Rao and Joel R. Tetreault. 2018. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In *Proc. NAACL-HLT*, pages 129–140.

Nils Reimers and Iryna Gurevych. 2019a. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proc. EMNLP-IJCNLP*, pages 3980–3990.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alexander M Rush, SEAS Harvard, Sumit Chopra, and Jason Weston. 2017. A neural attention model for sentence summarization. In *ACLWeb. Proceedings of the 2015 conference on empirical methods in natural language processing*.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. *arXiv preprint arXiv:1805.07685*.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: Cross projection in latent space. 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 4939–4948.

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.

Jitesh Shetty and Jafar Adibi. 2004. The enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*, 4(1):120–128.

Linfeng Song and Daniel Gildea. 2019. Sembleu: A robust metric for amr parsing evaluation. *arXiv preprint arXiv:1905.10726*.

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

Xiao Sun, Jia Li, Xing Wei, Changliang Li, and Jianhua Tao. 2020. Emotional editing constraint conversation content generation based on reinforcement learning. *Inf. Fusion*, 56:70–80.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

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, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. In *Advances in Neural Information Processing Systems*, pages 11034–11044.

Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018a. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. *arXiv preprint arXiv:1805.03616*.

Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018b. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4453–4460. International Joint Conferences on Artificial Intelligence Organization.

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 *Proc. EMNLP*, pages 38–45.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914.

Min Yang, Qiang Qu, Kai Lei, Jia Zhu, Zhou Zhao, Xiaojun Chen, and Joshua Z Huang. 2018. Investigating deep reinforcement learning techniques in personalized dialogue generation. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 630–638. SIAM.

Min Yang, Zhou Zhao, Wei Zhao, Xiaojun Chen, Jia Zhu, Lianqiang Zhou, and Zigang Cao. 2017. Personalized response generation via domain adaptation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1021–1024. ACM.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *35th International Conference on Machine Learning, ICML 2018*, pages 9405–9420. International Machine Learning Society (IMLS).

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.