

Semi-supervised Formality Style Transfer using Language Model Discriminator and Mutual Information Maximization

Kunal Chawla

Georgia Institute of Technology
Atlanta, GA

kunalchawla@gatech.edu

Diyi Yang

Georgia Institute of Technology
Atlanta, GA

dyang888@gatech.edu

Abstract

Formality style transfer is the task of converting informal sentences to grammatically-correct formal sentences, which can be used to improve performance of many downstream NLP tasks. In this work, we propose a semi-supervised formality style transfer model that utilizes a language model-based discriminator to maximize the likelihood of the output sentence being formal, which allows us to use maximization of token-level conditional probabilities for training. We further propose to maximize mutual information between source and target styles as our training objective instead of maximizing the regular likelihood that often leads to repetitive and trivial generated responses. Experiments showed that our model outperformed previous state-of-the-art baselines significantly in terms of both automated metrics and human judgement. We further generalized our model to unsupervised text style transfer task, and achieved significant improvements on two benchmark sentiment style transfer datasets.

1 Introduction

Text style transfer is the task of changing the style of a sentence while preserving the content. It has many useful applications, such as changing emotion of a sentence, removing biases in natural language, and increasing politeness in text (Sennrich et al., 2016; Pryzant et al.; Rabinovich et al., 2017; Yang et al., 2019; Chen et al., 2018).

There is a wide availability of “informal” data from online sources, yet current Natural Language Processing (NLP) tasks and models could not leverage or achieve good performance for such data due to informal expressions, and grammatical, spelling and semantic errors. Hence, formality style transfer, a specific style transfer task that aims to preserve the content of an informal sentence while making it semantically and grammatically correct, has

Informal	<i>I flippin' LOVE that movie, sweeeet!</i>
Formal	<i>I truly enjoy that movie.</i>
Informal	<i>we was hanging out a little.</i>
Formal	<i>We were spending a small amount of time together.</i>

Table 1: Examples of (formal, informal) sentence pairs.

recently received a growing amount of attention. Some examples are given in Table 1.

The most widely-used models for formality style transfer are based on a variational auto-encoder architecture, trained on parallel text data of (informal, formal) style sentence pairs with same content (Jing et al., 2019). However, there is still a lot of inconsistencies between human-generated sentences and outputs of current models, largely due to the limited availability of parallel data. In contrast, large amount of data consisting of sentences with just either informal or formal labels is relatively easier to collect. To tackle the training data bottleneck, we propose a semi-supervised approach for formality style transfer, using both human-annotated parallel data and large amount of unlabeled data.

Following the success of Generative Adversarial Nets (GAN) (Goodfellow et al., 2014), binary classifiers are often used on the generator outputs in unsupervised text style transfer to ensure that transferred sentences are similar to sentences in the target domain (Shen et al., 2017; Hu et al., 2017). However, Yang et al. (2018) showed that using a Language Model instead of a binary classifier can provide stronger, more stable training loss to the model, as it leverages probability of belonging to the target domain for each token in the sentence. We extend this line of work to semi-supervised formality style transfer, and propose to use two language models (one for source style and another for target) to help the model utilize information from both styles for training.

Moreover, style transfer models are usually

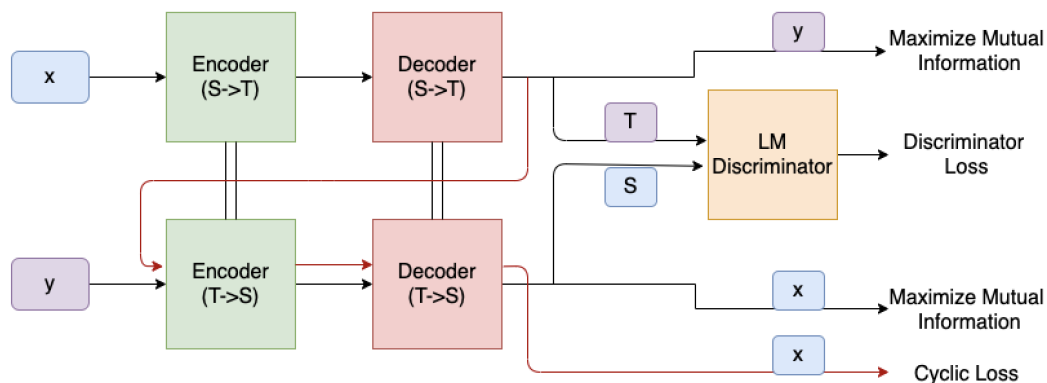


Figure 1: Model architecture. Here $(x, y) \in D$ is a (source, target) style sentence pair with same content, and S and T are source and target styles respectively. The parameters for encoder and decoder are shared across forward and backward style transfer directions. The red arrow corresponds to the cyclic reconstruction loss. Cyclic and discriminator losses are trained on $x \in U$, unsupervised class-labeled data.

trained by maximizing $P(y|x)$, where (x, y) is a (informal, formal) sentence pair. Such models tend to generate trivial outputs, often involving high-frequency phrases in the target domain (Li et al., 2016b). Building on prior work, to introduce more diversity and connections between the input and output, we propose to maximize mutual information (MMI) between source and target styles, which take into account not only the dependency of output on input, but also the likelihood that the input corresponds to the output. While this has only been done at test-time so far, we extend this approach to train our model with MMI objective.

We evaluate our proposed models that incorporate both the language model discriminators and mutual information maximization on Grammarly Yahoo Answers Corpus (GYAFC) Dataset (Rao and Tetreault, 2018). Experiments showed that our simple semi-supervised formality style transfer model outperformed state-of-the-art methods significantly, in terms of both automatic metrics (BLEU) and human evaluation. We further show that our approach can be used for unsupervised style transfer, as demonstrated by significant improvements over baselines on two sentiment style benchmarks: Yelp and Amazon Sentiment Transfer Corpus, where parallel data is not available. We have publicly released our code at <https://github.com/GT-SALT/FormalityStyleTransfer>.

2 Related Works

Sequence-to-Sequence Models Text style transfer is often modeled as a sequence-to-sequence

(seq2seq) task (Yang et al., 2018; Xu et al., 2019; Li et al., 2018). A classical architecture for seq2seq models is variational autoencoders (VAE) which uses an “encoder” to encode the input sentence into a hidden representation, and then uses a “decoder” to generate the new sentences (Shen et al., 2017; Hu et al., 2017; Jing et al., 2019). Long Short Term Memory (LSTMs) (Hochreiter and Schmidhuber, 1997), and more recently, self-attention based CNN architectures (Vaswani et al., 2017) are often used as base architectures for such models.

Pre-training of the encoders on multiple tasks and datasets has been shown to be effective (Devlin et al., 2018; Liu et al., 2019) in improving performances of individual tasks. These models are often trained with the cross-entropy loss (Vaswani et al., 2017) on the output tokens, or in other words, maximizing $P(y|x)$ where (x, y) is a pair of source and target style sentence respectively. Li et al. (2016b) showed that maximizing mutual information (MMI) $M(x, y)$ during test-time between the source and target instead can lead to more diverse and appropriate outputs in seq2seq models. Some other works (Zhang et al., 2018) maximize a variational lower bound on pairwise mutual information. We use a denoising auto-encoder BART (Lewis et al., 2019) trained with MMI objective.

Semi-Supervised and Unsupervised Style Transfer Some approaches like Li et al. (2018) and Lai et al. (2019) focus on deleting style-related keywords to make content style-independent. However, other works hypothesize that content and style cannot be separated, and use techniques such as back-translation (Lample et al., 2019),

cross-projection between styles in latent space (Shang et al., 2019a), reinforcement learning-based one step model (Luo et al., 2019), and iterative matching and translation (Jin et al., 2019). Following Goodfellow et al. (2014), using a generator along with a style classifier is often used for unsupervised tasks (Shen et al., 2017; Hu et al., 2017; Fu et al., 2018). However, recent work suggests (Yang et al., 2018) that using Language Models instead of CNN discriminators can result in more fluent, meaningful outputs. Maximizing likelihood of reconstruction of the input from the generated output has been used in both image generation (Zhu et al., 2017) and text style transfer (Shang et al., 2019b; Luo et al., 2019; Logeswaran et al., 2018) to improve performance. Motivated by these work, we use language models for our discriminator, and maximize cyclic reconstruction likelihood as part of our training objective.

Formality Style Transfer Grammarly (Rao and Tetreault, 2018) released a large-scale dataset for Formality Style Transfer, and tested several rule-based and deep neural networks-based baselines. CNN-based discriminators and cyclic reconstruction objective have been used (Xu et al., 2019) in a semi-supervised setting. Wang et al. (2019) used a combination of original and rule-based processed sentences to train the model. There is also evidence that using multi-task learning (Niu et al., 2018) and models pretrained on a large scale corpus (Wang et al., 2019) improve performance. This work uses a BART model (Lewis et al., 2019) pretrained on CNN-DM dataset (Nallapati et al., 2016) for our base architecture.

3 Method

This section presents our semi-supervised formality style transfer model. We detail the task and our base architecture in Section 3.1. We add a language model-based discriminator to the model, described in Section 3.2, and explain the maximization of mutual information in Section 3.3. The final architecture for our model is summarized in Section 3.4 and shown in Figure 1

3.1 Formality Style Transfer

Define T (“formal” in our case) as the target style and S (“informal”) as the source style for the formality style transfer task. Let D be the parallel dataset containing (source, target) style sentence pairs and U be the additional unlabeled

data, denoted by U_S for sentences with source style and U_T for sentences with target style.

Our base model is a variational auto-encoder mechanism G that generates sentences of target style. The goal is to maximize $P(y|x; \theta_G)$ where θ_G are the parameters of the model. This is done by cross-entropy loss over the target sentence tokens and generated output probabilities. To leverage Maximum Mutual Information objective, as described in Section 3.3, we make the model bi-directional. It can be used to transfer source style to target style as well as target style to source style. Hence, an additional input $c \in \{S, T\}$ is passed to G specifying the style to which the sentence is to be converted. Hence, our objective for base model is to maximize $P(y|x, T; \theta_G)$.

3.2 Language Model Discriminator

We add a Language model(LM) based discriminator to the model. It functions as a binary classifier which scores the formality of the output generated by the decoder. It includes two language models trained independently on informal and formal data. The “score” of a sentence by a language model is calculated by the product of locally normalized probabilities of each token given the previous tokens. Let x be a sentence from P with label c , then

$$LM(x) = \prod_{i=0}^{len(x)} P(x_i|x_{0:i-1}; \theta_{LM}) \quad (1)$$

where x_i are the tokens in x and θ_{LM} are the parameters of the language model. The softmax-normalized score of the sentence by the language models is interpreted as the classifier score:

$$P(T|x) = \frac{e^{LM_T(x)}}{e^{LM_T(x)} + e^{LM_S(x)}} \quad (2)$$

The language model discriminator is pre-trained on source and target data from P with the cross entropy loss:

$$\theta_C^* = \min_{\theta_C} \sum_{x \in U} (-\log P(c|x; \theta_C)) \quad (3)$$

where c is the label of x , θ_C are the parameters of the LM discriminator and θ_C^* are the trained parameters. The weights are then frozen for the training. A common training objective ((Wang et al., 2019; Fu et al., 2018)) is to minimize the sum of translation loss L_{trans} and discriminator loss L_{disc} ,

defined as:

$$\begin{aligned}
L_{trans}((x, y) \in D) &= -\log(P(y|x, T; \theta_G)) \\
L_{disc}(x \in U_S) &= -\log(P(T|x, T; \theta_G, \theta_C)) \\
L_{disc}(x \in U_T) &= -\log(P(S|x, S; \theta_G, \theta_C)) \\
\theta_G^* &= \min_{\theta_G} \left(\sum_{(x,y) \in D} L_{trans}(x, y) + \sum_{x \in U} L_{disc}(x) \right)
\end{aligned} \tag{4}$$

3.3 Maximum Mutual Information Objective

As discussed, instead of using usual translation loss which maximizes $P(y|x; \theta_G)$ and often produces trivial and repetitive content, we chose to maximize pairwise mutual information between the source and the target:

$$\begin{aligned}
\hat{y} &= \operatorname{argmax}_y \log \frac{P(x, y)}{P(y)P(x)} \\
&= \operatorname{argmax}_y (\log P(y|x) - \log P(y))
\end{aligned} \tag{5}$$

Following (Li et al., 2016b), we introduce a parameter λ ‘‘forward-translation weight’’ to generalize the MMI objective and adjust the relative weights of forwards and backwards translation:

$$\begin{aligned}
\hat{y} &= \operatorname{argmax}_y (\log P(y|x) - (1 - \lambda) \log P(y)) \\
&= \operatorname{argmax}_y (\lambda \log P(y|x) + (1 - \lambda) \log P(x|y))
\end{aligned}$$

The translation loss thus becomes:

$$\begin{aligned}
L_{trans}((x, y) \in D) &= \lambda \log P(y|x, T; \theta_G) \\
&\quad + (1 - \lambda) \log P(x|y, S; \theta_G)
\end{aligned} \tag{6}$$

3.4 Overall Model Architecture

Making the model bi-directional also allows us to leverage unsupervised data using cyclical reconstruction loss L_{cycle} , which encourages a sentence translated to the opposite style and back to be similar to itself (Shang et al., 2019b). Let $G(x, c)$ be the output of the model for a sentence x with target style c . Then

$$\begin{aligned}
L_{cycle}(x \in U_S) &= -\log P(x|G(x, T), S; \theta_G) \\
L_{cycle}(x \in U_T) &= -\log P(x|G(x, S), T; \theta_G)
\end{aligned}$$

Let w_{disc} and w_{cycle} denote the weights for discriminator and cyclic loss respectively. The overall loss function L for the training step is:

$$\begin{aligned}
L &= \sum_{(x,y) \in D} L_{trans}(x, y) \\
&\quad + \sum_{x \in U} (w_{disc} L_{disc}(x) + w_{cycle} L_{cycle}(x))
\end{aligned} \tag{7}$$

Dataset	Train	Valid	Test
E&M	52595	2877	1416
F&R	51967	2788	1432
BookCorpus	214K	-	-
Twitter	211K	-	-
Yelp (Positive)	270K	2000	500
Yelp (Negative)	180K	2000	500
Amazon (Positive)	277K	985	500
Amazon (Negative)	278K	1015	500

Table 2: The statistics of train, validation and test sets of all used datasets.

4 Experiments

4.1 Dataset

We used Grammarly’s Yahoo Corpus Dataset (GYAFC) (Rao and Tetreault, 2018) as our parallel data for supervised training. The dataset is divided into two sub-domains- ‘‘Entertainment and Music’’ (E&M) and ‘‘Family and Relationships’’ (F&R). For the unsupervised data, we crawled Twitter data for informal data, and we used BookCorpus data (Zhu et al., 2015) for the formal data. In the pre-training step, we train the language model discriminator on the unannotated informal and formal data. The detailed process of the data collection is given in the Appendix. The statistics of datasets are in Table 2.

4.2 Pre-processing and Experiment Setup

The text was pre-processed with Byte Pair Encoding(BPE) (Shibata et al., 1999) with a vocabulary size of 50,000. For pre-training, we trained the LM Discriminator with the unsupervised data with cross entropy loss. For training, we merged both datasets of GYAFC and used the training objective as described in Section 3.4 to train the model.

We used Fairseq (Ott et al., 2019) library built on top of PyTorch (Paszke et al., 2019) to run our experiments. We used BART-large (Lewis et al., 2019) model pretrained on CNN-DM summarization data (Nallapati et al., 2016) for our base encoder and decoder. BART was chosen because of its bidirectional encoder which uses words from both left and right for training, as well as superior performance on text generation tasks. Its training objective of reconstruction from noisy text data fits our task well. We chose the model pre-trained on CNN-DM dataset because of the relevance of the decoder pre-trained on formal words to our task.

Both decoder and the encoder have 12 layers each with 16 attention heads and a hidden embed-

ding size of 1024. We shared the weights for encoder and decoder across the forward and backward translation, using a special input token to the encoder. For the language models, we used a Transformer (Vaswani et al., 2017) decoder with 4 layers and 8 attention heads per layer.

One NVIDIA RTX 2080 Ti with 11GB memory was used to run the experiments with the max token size of 64. We also used update frequency 4, increasing the effective batch size. Adam Optimizer (Kingma and Ba, 2015) was used to train the model, and the parameters learning rate, λ , w_{disc} and w_{cycle} were fine-tuned. The model was selected based on perplexity of informal to formal translation on validation data. Beam search (size = 10) was used to generate sentences. A length penalty (= 2.0) was used to reduce redundancy in the output sentence. Further details on model parameters are mentioned in Appendix.

4.3 Evaluation Metrics

The result was evaluated with BLEU (Papineni et al., 2002). We used word tokenizer and corpus BLEU calculator from Natural Language Toolkit (NLTK) (Loper and Bird, 2002) to calculate the BLEU score. Due to the subjective nature of the task, BLEU does not capture the output of the model well. Hence, we also used human annotations for some of the models.

Amazon Mechanical Turk was used to evaluate 100 randomly sampled sentences from each dataset of GYAFC. To increase annotation quality, we required workers located in US to have a 98% approval rate and at least 5000 approved HITs for their previous work on MTurk. Each sentence was annotated by 3 workers, who rated each generated sentence using the following metrics, following (Rao and Tetreault, 2018):

- **Content:** Annotators judge if the source and translated sentence convey the same information on a scale of 1-6: 6: Completely equivalent, 5: Mostly equivalent, 4: Roughly equivalent, 3: Not equivalent but share some details, 2: Not equivalent but on same topic, 1: Completely dissimilar.
- **Fluency:** Workers score the clarity and ease of understanding of the translated sentence on a scale from 1-5: 5: Perfect, 4: Comprehensible, 3: Somewhat Comprehensible, 2: Incomprehensible, 1: Incomplete.

- **Formality:** Workers rate the formality of the translated sentence on a scale of -3 to 3. -3: Very Informal, -2: Informal, -1: Somewhat Informal, 0: Neutral, 1: Somewhat Formal, 2: Formal and 3: Very Formal.

We also provided detailed definitions and examples to workers, which are described together with annotation interface in Appendix. The intra-class correlation was estimated using ICC-2k (Random sample of k raters rate each target) and calculated using Pingouin (Vallat, 2018) Python package. It varied from 0.521-0.563 for various models, indicating moderate agreement (Koo and Li, 2016). We then averaged the three human-provided labels to obtain the rating for each sentence.

4.4 Baselines and Model Variants

We compared our approach with several baseline methods as follows:

- **SimpleCopy:** Simply copying the source sentence as the generated output.
- **Target:** Human-generated outputs.
- **Rule-based** (Rao and Tetreault, 2018): Using hand-made rules.
- **NMT** (Jhamtani et al., 2017): A LSTM encoder-decoder model with attention.
- **Transformer** (Vaswani et al., 2017): A Transformer architecture with the same configuration as our encoder and decoder.

We also compared our model with previous state-of-the-art works:

- **Hybrid Annotations** (Xu et al., 2019): Uses CNN-based discriminator and cyclic reconstruction loss in a semi-supervised setting.
- **NMT Multi-Task** (Niu et al., 2018): Solves two tasks: monolingual formality transfer and formality-sensitive machine translation jointly using multi-task learning.
- **Pretrained w/ Rules** (Wang et al., 2019): Uses a pre-trained OpenAI GPT-2 model and a combination of original and rule-based processed sentences to train the model.

The performances for these works were taken from the respective papers. We also introduced several variants of our model for comparison:

Model	E&M				F&R			
	BLEU	Content	Fluency	Formality	BLEU	Content	Fluency	Formality
SimpleCopy	50.28	-	-	-	51.66	-	-	-
Target	99.99	5.54	4.79	2.31	100.00	5.54	4.79	2.30
Rule-based	60.37	-	-	-	66.40	-	-	-
NMT (Jhamtani et al., 2017)	68.41	-	-	-	74.22	-	-	-
Transformer (Vaswani et al., 2017)	67.97	-	-	-	74.20	-	-	-
Hybrid Annotations (Xu et al., 2019)*	69.63	5.22	4.62	1.97	74.43	5.29	4.53	2.04
NMT Multi-task (Niu et al., 2018)	72.13	-	-	-	75.37	-	-	-
Pretrained w/ Rules (Wang et al., 2019)	72.70	5.38	4.51	1.67	76.87	5.64	4.63	1.78
Dual Reinforcement** (Luo et al., 2019)	-	-	-	-	41.9	-	-	-
Ours Base	74.66	4.93	4.33	1.82	78.89	5.06	4.36	1.84
w/ CNN discriminator*	75.04	-	-	-	79.05	-	-	-
w/ LM discriminator*	75.65	5.33	4.69	2.30	79.50	5.35	4.66	2.31
w/ LM* + MMI	76.19	-	-	-	79.92	-	-	-
Ours*	76.52	5.35	4.81	2.38	80.29	5.42	4.74	2.31

Table 3: Results on GYAFC Dataset. An average of 3 runs was used for each model to calculate BLEU. Models with * leverage extra data via semi-supervised methods. ** represents unsupervised models. The description for the models is given in Section 4.3. The best scores (besides the target) for each metric are in **bold**.

- **Ours Base.** Pretrained uni-directional auto-encoder architecture from BART (Lewis et al., 2019) fine-tuned on our data.
- **Ours w/ CNN Discriminator:** A CNN architecture with 3 layers used on the output of the decoder. The discriminant was trained with unsupervised class-labeled data.
- **Ours w/ LM Discriminator:** Two transformer-based language models with 4 layers, used on the output of the decoder.
- **Ours w/ LM + MMI:** Model trained with MMI objective and LM discriminator.
- **Ours:** Ours Base model trained with LM discriminator, MMI objective, and cyclic reconstruction loss.

4.5 Results

The results are summarized in Table 2. Compared to various baselines such as Pretrained w/ Rules (Wang et al., 2019), our proposed models achieved significant improvements with 3.82 absolute increase of BLEU on E&M and an increase of 3.42 on F&R. By utilizing the language model discriminator and mutual information maximization, Ours achieved state-of-the-art results on both subsets of the GYAFC dataset in terms of BLEU, boosting the BLEU to 76.52 and 80.29 on E&M and F&R respectively. Our contributions increase the score by 2-3 points compared to the fine-tuned BART baseline as well. This validates the effectiveness of our semi-supervised formality style transfer models.

Details on runtime and memory requirements can be found in Appendix. Our contributions increase the performance without increasing the test-time or memory requirements significantly.

Consistent with this quantitative result, human annotation results showed that Ours produced more fluent and more formal outputs compared to our selected baselines. Pretrained w/ Rules was rated to have better content preservation, but lower fluency and formality. This is possibly due to different approaches taken to deal with slang and idiomatic expressions in language, as described in Section 5.3 (Type 7). Wang et al. (2019) tends to keep the content at the cost of formality of the output, while Xu et al. (2019) and our model often ignore the content. For example, our model’s output of “*the two boys rednecked as hell play guitar*” is “*The two boys play guitar.*”, omitting details like “red-neck” which are rarely mentioned in formal language.

Moreover, we observed that there are comparable human annotation results between Target and Ours. Our model achieved slightly higher scores on the formality of the sentences compared to human-generated outputs. This may suggest that our model has a tendency to increase the formality of a sentence, even if it loses a bit of meaning preservation. We also found that additional unsupervised data helps: compared to Ours Base, language model discriminator improves performance significantly (with BLEU scores from 74.66 to 75.65, and from 78.89 to 79.50). Note that our method is generic, and can be further combined with baseline methods, such as Wang et al. (2019);

Model	Sentence
Informal Hybrid Annotations (Xu et al., 2019) Pretrained w/ rules (Wang et al., 2019) Ours Human-Annotation	fidy cent he is fine and musclar Fidy Cent is fine and Muslim. Fidy Cent is a fine and musclar artist. 50 Cent is fine and muscular. 50 Cent is fine and muscular.
Informal Hybrid Annotations Pretrained w/ rules Ours Human-Annotation	Plus she is a cray ****. She is a clay. She is a cray ****. She is not very nice. Also, she is a mentally unstable woman.
Informal Hybrid Annotations Pretrained w/ rules Ours Human-Annotation	So far i haven't heard that shes come back here (Arkansas)? I have not heard that she is in Arkansas. So far, I have not heard that she is coming back here(Arkansas). So far I have not heard that she has returned to Arkansas. So far I have not heard that she returned to Arkansas.

Table 4: Some sample outputs from various models.

Niu et al. (2018).

We notice that BLEU does not necessarily correlate well with improved fluency, which is consistent with previous studies (Rao and Tetreault, 2018; Lin and Och, 2004). Many fluent sentences did not capture the meaning of the sentence well, which reduces BLEU. Conversely, it is possible to have high intersection with the gold label sentence but still not be fluent.

Some qualitative results from our best-performing model (by BLEU score in Table 3), Xu et al. (2019), Wang et al. (2019) and target sentences, are provided in Table 4. We observed that our model consistently generates better translations compared to the previous methods, especially in terms of dealing with proper nouns, informal phrases and grammatical mistakes.

4.6 Testing on Unsupervised data

We further extended our method to unsupervised tasks, using only cyclic reconstruction and Language Discriminator losses as our training objective. Sentiment Transfer corpus (Li et al., 2018) from Yelp and Amazon was used for evaluation. The statistics are given in Table 2. The corpora include separate negative and positive sentiment data without parallel data. We followed the evaluation protocol and baselines from Li et al. (2018). In addition to BLEU, we used two additional metrics for evaluation: (1) **Accuracy**: The percentage of sentences successfully translated into positive, as measured by a separate pre-trained classifier. (2)

G-Score: The geometric Mean of accuracy and BLEU scores. We rank our models by G-Score, following Xu et al. (2012), since there is a trade-off between accuracy and BLEU, as changing more words can get better accuracy but lower content preservation.

We used the script and sentiment classifier from Li et al. (2018) to evaluate our outputs. Results were averaged for the two directions: positive-to-negative sentiment transfer and negative-to-positive sentiment transfer, with 500 sentences in the test set for each direction.

We compared our results with previous state-of-the-art approaches. Style Embedding and Multi Decoding (Fu et al., 2018) learn an embedding of the source sentence such that a decoder can use it to reconstruct the sentence, but a discriminator, which tries to identify the source attribute using this encoding, fails. Cross-Aligned (Shen et al., 2017) also encodes the source sentence into a vector, but the discriminator looks at the hidden states of the RNN decoder.

Li et al. (2018) extract content words by deleting style-related phrases, retrieves relevant target-related phrases and combines them using a neural model. They provide three variants of their model. Word-level Conditional GAN (Lai et al., 2019) also tries to separate content and style with a word-level conditional architecture. Dual Reinforcement (Luo et al., 2019) uses reinforcement learning for bidirectional translation without separating style and content. Iterative Matching (Jin et al., 2019)

Model	Yelp			Amazon		
	ACC	BLEU	G-Score	ACC	BLEU	G-Score
SimpleCopy	2.4	18.0	6.57	18.9	39.2	27.2
Target	69.6	100.0	83.4	41.3	99.9	64.2
Cross Aligned (Shen et al., 2017)	73.7	3.1	15.1	74.1	0.4	5.4
Style Embedding (Fu et al., 2018)	8.7	11.8	10.1	43.3	10.0	20.8
Multi Decoding (Fu et al., 2018)	47.6	7.1	18.4	68.3	5.0	18.5
Template Based (Li et al., 2018)	81.7	11.8	31.0	68.7	27.1	43.1
Retrieve Only (Li et al., 2018)	95.4	0.4	6.2	70.3	0.9	8.0
Delete Only (Li et al., 2018)	85.7	7.5	25.4	45.6	24.6	33.5
Delete & Retrieve (Li et al., 2018)	88.7	8.4	27.3	48.0	22.8	33.1
Dual Reinforcement (Luo et al., 2019)	85.6	13.9	34.5	-	-	-
Word-level Conditional GAN (Lai et al., 2019)	87.8	9.6	29.1	77.4	6.7	22.8
Iterative Matching (Jin et al., 2019)	87.9	4.3	19.4	-	-	-
Ours	86.2	14.1	34.9	68.9	28.6	44.4

Table 5: Results on Sentiment Transfer datasets. Results were averaged across two directions: negative-to-positive and positive-to-negative sentiment transfer. An average of three runs was used for each directions. Here ACC and GM mean Accuracy and G-Score respectively. The best scores (besides the target) for each metric are in **bold**.

iteratively refines imperfections in the alignment of semantically similar sentences from the source and target dataset. We used the performance numbers for these approaches from either the original papers when the evaluation protocol is similar to ours or by evaluating publicly released outputs of the models.

We achieved state-of-the-art results on both Yelp and Amazon Sentiment Transfer corpus, as shown in Table 5. Our model attains slightly lower accuracy on sentiment classification of output sentences, but preserves more content compared to previous models, resulting in the highest G-Score on both datasets. This suggests that our approach can generalize well to unsupervised style transfer tasks.

5 Model Analysis and Discussion

Although our model performed well on formality style transfer, there is still a gap compared to human performance. To understand why the task is challenging and how future research could advance this direction, we take a closer look at formality dataset, model generation errors, and certain challenges that existing approaches struggle with.

5.1 Effect of Forward Translation Weight

As mentioned in Section 3.3, MMI objective is equivalent to a weighted sum of source-to-target and target-to-source translation. We show the effect of forward translation weight, λ in Figure 2, and find that using MMI objective helps performance

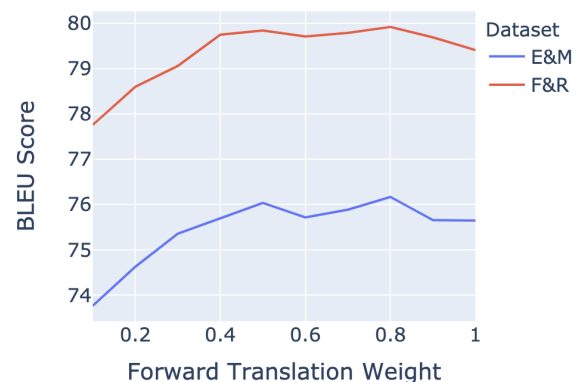


Figure 2: Performance with forward translation weight

as compared to baseline translation loss (which corresponds to $\lambda = 1.0$). However, equivalent weighing of the two directions (corresponding to $\lambda = 0.5$) does not result in the best performance: a bias towards the informal to formal direction ($\lambda = 0.8$) gives better BLEU scores. We posit that this could be because unlike formal sentences, informal sentences do not follow a particular style: they vary from structurally correct with some mistakes to just a collection of telegram-style keywords, and hence the objective of generating this should be assigned less importance than the forward task.

5.2 Cyclic and Discriminator Loss

In our model, we used unsupervised class labeled data to train our model using cyclic and discrim-

inator loss. We also conducted experiments to use these losses for parallel data as well. However, training on parallel data using these objectives in addition to MMI objective did not result in additional improvements, while increasing the training time and memory requirements. Partially, this could be because maximizing target sentence probability already captures the target style, hence discriminator loss does not help. Similarly, maximizing Mutual Information ensures that target-to-source translation is also a maximisation objective during training, hence reducing the effectiveness of cyclic reconstruction loss. Therefore, we concluded that maximizing mutual information during training is sufficient for parallel data.

5.3 Challenges in Formality Text Transfer

We conduct a thorough examination of the GYAFC dataset and categorize the challenges into the following categories:

1. **Informal Phrases and Abbreviations:** Presence of “informal” phrases (*what the hell*), emojis (.:)) and abbreviations (*omg, brb*).
2. **Missing Context:** A lack of context of the conversation (for example, “*It had to be the chickin*”) or a lack of punctuation or proper capitalization cues (“*can play truth or dare or snake and ladders*”).
3. **Named Entities:** Proper nouns and popular references like “Fifty Cent” or “eBay” should not be changed despite the wrong pluralization and capitalization, respectively. This is worsened by the lack of any capitalization or punctuation cues to find named entities.
4. **Sarcasm and Rhetorical Questions:** Rhetorical questions, sarcastic language and negations have been long-standing problems in NLP (Li et al., 2016a). For example, “*sure, because this is so easy*” is sarcastic and should not be translated literally.
5. **Repetition:** Informal text often has a lot of redundant information. For example, “*I used to work at the store and met him while i was working there.*” can be formally structured as “*I met him while i was working at the store.*”.
6. **Spellings and Grammar Errors:** This is prevalent in most (>60%) informal sentences.

Type	Input (%)	Output (%)	Resolved (%)
1	16	7	56
2	5	4	20
3	12	3	75
4	2	2	0
5	3	0	100
6	61	7	89
7	5	0	100

Table 6: The breakdown of challenge types for formality style transfer, and their percentages in the source *Input*, generated *Output*, and the percentage of challenges successfully resolved by our model.

7. **Slang and Idiomatic Expressions:** Some sentences have words especially nouns, adjectives and adverbs that can be considered as slang, idiom, and even discriminatory.

We randomly sampled 100 sentences from the dataset to estimate the prevalence of such challenges. We also examined the output from our model to analyze if a challenge has been solved or still presents an issue to the model. The result is summarized in Table 6. We found that our models resolved most spelling and grammatical mistakes (Type 6), and performs well with avoiding repetition (Type 5). However, missing context, informal expressions and named entities continue to be challenging. One major challenge is the inability to correct sarcastic/rhetorical sentences (Type 4).

6 Conclusion

This work introduces a semi-supervised formality style transfer model that utilizes both a language model based discriminator to maximize the likelihood of the output sentences being formal, and a mutual information maximization loss during training. Experiments conducted on a large-scale formality corpus showed that our simple method significantly outperformed previous approaches in terms of both automatic metrics and human judgement. We also demonstrated that our model can be generalized well to unsupervised style transfer tasks. We also discussed specific challenges that current approaches faced with this task.

Acknowledgements

We thank the reviewers and members of Georgia Tech SALT group for their feedback on this work. We acknowledge the support of NVIDIA Corporation with the donation of GPU used for this work.

References

- Wei-Fan Chen, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Learning to flip the bias of news headlines. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 79–88.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.
- Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *AAAI*.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14*, page 2672–2680, Cambridge, MA, USA. MIT Press.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.
- 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, ICML'17*, page 1587–1596. JMLR.org.
- 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, Copenhagen, Denmark. Association for Computational Linguistics.
- 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 3097–3109, Hong Kong, China. Association for Computational Linguistics.
- Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2019. Neural style transfer: A review. *IEEE transactions on visualization and computer graphics*.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.
- Terry K Koo and M. Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15 2:155–63.
- Chih-Te Lai, Yi-Te Hong, Hong-You Chen, Chi-Jen Lu, and Shou-De Lin. 2019. Multiple text style transfer by using word-level conditional generative adversarial network with two-phase training. 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 3579–3584, Hong Kong, China. Association for Computational Linguistics.
- Guillaume Lample, Sandeep Subramanian, Eric Michael Smith, Ludovic Denoyer, Marc’Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *ICLR*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016a. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational 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, New Orleans, Louisiana. Association for Computational Linguistics.
- Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.
- Lajanugen Logeswaran, Honglak Lee, and Samy Bengio. 2018. Content preserving text generation with

- attribute controls. In *Advances in Neural Information Processing Systems*, pages 5103–5113.
- Edward Loper and Steven Bird. 2002. [Nltk: The natural language toolkit](#). In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, page 63–70, USA. Association for Computational Linguistics.
- Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. [A dual reinforcement learning framework for unsupervised text style transfer](#).
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence RNNs and beyond](#). In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.
- Xing Niu, Sudha Rao, and Marine Carpuat. 2018. [Multi-task neural models for translating between styles within and across languages](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. [fairseq: A fast, extensible toolkit for sequence modeling](#). In *Proceedings of NAACL-HLT 2019: Demonstrations*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.
- Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. [Automatically neutralizing subjective bias in text](#).
- Ella Rabinovich, Shachar Mirkin, Raj Patel, Mount Carmel, Lucia Specia, and Shuly Wintner. 2017. [Personalized machine translation: Preserving original author traits](#).
- Sudha Rao and Joel R. Tetreault. 2018. [Dear sir or madam, may i introduce the yafc corpus: Corpus, benchmarks and metrics for formality style transfer](#). In *NAACL-HLT*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Controlling politeness in neural machine translation via side constraints](#). pages 35–40.
- Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019a. [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.
- Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019b. [Semi-supervised text style transfer: Cross projection in latent space](#).
- 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.
- Yusuke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, and Takeshi Shinohara. 1999. [Byte pair encoding: A text compression scheme that accelerates pattern matching](#).
- Raphael Vallat. 2018. [Pingouin: statistics in python](#). *The Journal of Open Source Software*, 3(31):1026.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. [Harnessing pre-trained neural networks with rules for formality style transfer](#). In *EMNLP/IJCNLP*.
- Ruo Chen Xu, Tao Ge, and Furu Wei. 2019. [Formality style transfer with hybrid textual annotations](#). *ArXiv*, abs/1903.06353.
- Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. [Paraphrasing for style](#). In *Proceedings of COLING 2012*, pages 2899–2914, Mumbai, India. The COLING 2012 Organizing Committee.
- Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. [Let's make your request more persuasive: Modeling persuasive strategies via](#)

semi-supervised neural nets on crowdfunding platforms. 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 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. [Unsupervised text style transfer using language models as discriminators](#). In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7287–7298. Curran Associates, Inc.

Y. Zhang, M. Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and W. Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *NeurIPS*.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei Efros. 2017. [Unpaired image-to-image translation using cycle-consistent adversarial networks](#). pages 2242–2251.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. [Aligning books and movies: Towards story-like visual explanations by watching movies and reading books](#). In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), ICCV '15*, page 19–27, USA. IEEE Computer Society.