# Politeness Transfer: A Tag and Generate Approach

**Aman Madaan** *, **Amrith Setlur** *, **Tanmay Parekh** *, **Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, Shrimai Prabhumoye**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
{amadaan, asetlur, tparekh}@cs.cmu.edu

## Abstract

This paper introduces a new task of politeness transfer which involves converting non-polite sentences to polite sentences while preserving the meaning. We also provide a dataset of more than 1.39 million instances automatically labeled for politeness to encourage benchmark evaluations on this new task. We design a *tag* and *generate* pipeline that identifies stylistic attributes and subsequently generates a sentence in the target style while preserving most of the source content. For politeness as well as five other transfer tasks, our model outperforms the state-of-the-art methods on automatic metrics for content preservation, with a comparable or better performance on style transfer accuracy. Additionally, our model surpasses existing methods on human evaluations for grammaticality, meaning preservation and transfer accuracy across all the six style transfer tasks. The data and code is located at https://github.com/tag-and-generate/

## 1 Introduction

Politeness plays a crucial role in social interaction, and is closely tied with power dynamics, social distance between the participants of a conversation, and gender (Brown et al., 1987; Danescu-Niculescu-Mizil et al., 2013). It is also imperative to use the appropriate level of politeness for smooth communication in conversations (Coppock, 2005), organizational settings like emails (Peterson et al., 2011), memos, official documents, and many other settings. Notably, politeness has also been identified as an interpersonal style which can be decoupled from content (Kang and Hovy, 2019). Motivated by its central importance, in this paper we study the task of converting non-polite sentences to polite sentences while preserving the meaning.

Prior work on text style transfer (Shen et al., 2017; Li et al., 2018; Prabhumoye et al., 2018; Rao and Tetreault, 2018; Xu et al., 2012; Jhamtani et al., 2017) has not focused on politeness as a style transfer task, and we argue that defining it is cumbersome. While native speakers of a language and cohabitants of a region have a good working understanding of the phenomenon of politeness for everyday conversation, pinning it down as a definition is non-trivial (Meier, 1995). There are primarily two reasons for this complexity. First, as noted by (Brown et al., 1987), the phenomenon of politeness is rich and multifaceted. Second, politeness of a sentence depends on the culture, language, and social structure of both the speaker and the addressed person. For instance, while using "please" in requests made to the closest friends is common amongst the native speakers of North American English, such an act would be considered awkward, if not rude, in the Arab culture (Kádár and Mills, 2011).

We circumscribe the scope of politeness for the purpose of this study as follows: First, we adopt the data driven definition of politeness proposed by (Danescu-Niculescu-Mizil et al., 2013). Second, we base our experiments on a dataset derived from the Enron corpus (Klimt and Yang, 2004) which consists of email exchanges in an American corporation. Thus, we restrict our attention to the notion of politeness as widely accepted by the speakers of North American English in a formal setting.

Even after framing politeness transfer as a task, there are additional challenges involved that differentiate politeness from other styles. Consider a common directive in formal communication, "send me the data". While the sentence is not impolite, a rephrasing "could you please send me the data" would largely be accepted as a more polite way of phrasing the same statement (Danescu-Niculescu-Mizil et al., 2013). This example brings out a distinct characteristic of politeness. It is easy to pinpoint the signals for *politeness*. However,

* authors contributed equally to this work.



*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881
July 5 - 10, 2020. 

cues that signal the *absence* of politeness, like direct questions, statements and factuality (Danescu-Niculescu-Mizil et al., 2013), do not explicitly appear in a sentence, and are thus hard to objectify. Further, the other extreme of politeness, impolite sentences, are typically riddled with curse words and insulting phrases. While interesting, such cases can typically be neutralized using lexicons. For our study, we focus on the task of transferring the non-polite sentences to polite sentences, where we simply define non-politeness to be the absence of both politeness and impoliteness. Note that this is in stark contrast with the standard style transfer tasks, which involve transferring a sentence from a well-defined style polarity to the other (like positive to negative sentiment).

We propose a *tag* and *generate* pipeline to overcome these challenges. The *tagger* identifies the words or phrases which belong to the original style and replaces them with a tag token. If the sentence has no style attributes, as in the case for politeness transfer, the tagger adds the tag token in positions where phrases in the target style can be inserted. The *generator* takes as input the output of the tagger and generates a sentence in the target style. Additionally, unlike previous systems, the outputs of the intermediate steps in our system are fully realized, making the whole pipeline interpretable. Finally, if the input sentence is already in the target style, our model won't add any stylistic markers and thus would allow the input to flow as is.

We evaluate our model on politeness transfer as well as 5 additional tasks described in prior work (Shen et al., 2017; Prabhumoye et al., 2018; Li et al., 2018) on content preservation, fluency and style transfer accuracy. Both automatic and human evaluations show that our model beats the state-of-the-art methods in content preservation, while either matching or improving the transfer accuracy across six different style transfer tasks(§5). The results show that our technique is effective across a broad spectrum of style transfer tasks. Our methodology is inspired by Li et al. (2018) and improves upon several of its limitations as described in (§2).

Our main contribution is the design of politeness transfer task. To this end, we provide a large dataset of nearly 1.39 million sentences labeled for politeness (https://github.com/tag-and-generate/politeness-dataset). Additionally, we hand curate a test set of 800 samples (from Enron emails) which are annotated as requests. To the best of our knowledge, we are the first to undertake politeness as a style transfer task. In the process, we highlight an important class of problems wherein the transfer involves going from a neutral style to the target style. Finally, we design a "tag and generate" pipeline that is particularly well suited for tasks like politeness, while being general enough to match or beat the performance of the existing systems on popular style transfer tasks.

## 2 Related Work

Politeness and its close relation with power dynamics and social interactions has been well documented (Brown et al., 1987). Recent work (Danescu-Niculescu-Mizil et al., 2013) in computational linguistics has provided a corpus of *requests* annotated for politeness curated from Wikipedia and StackExchange. Niu and Bansal (2018) uses this corpus to generate polite dialogues. Their work focuses on contextual dialogue response generation as opposed to content preserving style transfer, while the latter is the central theme of our work. Prior work on Enron corpus (Yeh and Harnly, 2006) has been mostly from a socio-linguistic perspective to observe social power dynamics (Bramsen et al., 2011; McCallum et al., 2007), formality (Peterson et al., 2011) and politeness (Prabhakaran et al., 2014). We build upon this body of work by using this corpus as a source for the style transfer task.

Prior work on style transfer has largely focused on tasks of sentiment modification (Hu et al., 2017; Shen et al., 2017; Li et al., 2018), caption transfer (Li et al., 2018), persona transfer (Chandu et al., 2019; Zhang et al., 2018), gender and political slant transfer (Reddy and Knight, 2016; Prabhumoye et al., 2018), and formality transfer (Rao and Tetreault, 2018; Xu et al., 2019). Note that formality and politeness are loosely connected but independent styles (Kang and Hovy, 2019). We focus our efforts on carving out a task for politeness transfer and creating a dataset for such a task.

Current style transfer techniques (Shen et al., 2017; Hu et al., 2017; Fu et al., 2018; Yang et al., 2018; John et al., 2019) try to disentangle source style from content and then combine the content with the target style to generate the sentence in the target style. Compared to prior work, "Delete, Retrieve and Generate" (Li et al., 2018) (referred to as DRG henceforth) and its extension (Sudhakar et al., 2019) are effective methods to generate out-

puts in the target style while having a relatively high rate of source content preservation. However, DRG has several limitations: (1) the delete module often marks content words as stylistic markers and deletes them, (2) the retrieve step relies on the presence of similar content in both the source and target styles, (3) the retrieve step is time consuming for large datasets, (4) the pipeline makes the assumption that style can be transferred by deleting stylistic markers and replacing them with target style phrases, (5) the method relies on a fixed corpus of style attribute markers, and is thus limited in its ability to generalize to unseen data during test time. Our methodology differs from these works as it does not require the retrieve stage and makes no assumptions on the existence of similar content phrases in both the styles. This also makes our pipeline faster in addition to being robust to noise.

Wu et al. (2019) treats style transfer as a conditional language modelling task. It focuses only on sentiment modification, treating it as a cloze form task of filling in the appropriate words in the target sentiment. In contrast, we are capable of generating the entire sentence in the target style. Further, our work is more generalizable and we show results on five other style transfer tasks.

## 3 Tasks and Datasets

### 3.1 Politeness Transfer Task

For the politeness transfer task, we focus on sentences in which the speaker communicates a requirement that the listener needs to fulfill. Common examples include imperatives "*Let's stay in touch*" and questions that express a proposal "*Can you call me when you get back?*". Following Jurafsky et al. (1997), we use the umbrella term "action-directives" for such sentences. The goal of this task is to convert action-directives to polite requests. While there can be more than one way of making a sentence polite, for the above examples, adding gratitude ("*Thanks* and let's stay in touch") or counterfactuals ("*Could* you please call me when you get back?") would make them polite (Danescu-Niculescu-Mizil et al., 2013).

**Data Preparation** The Enron corpus (Klimt and Yang, 2004) consists of a large set of email conversations exchanged by the employees of the Enron corporation. Emails serve as a medium for exchange of requests, serving as an ideal application for politeness transfer. We begin by pre-processing the raw Enron corpus following Shetty and Adibi (2004). The first set of pre-processing[1] steps and de-duplication yielded a corpus of roughly 2.5 million sentences. Further pruning[2] led to a cleaned corpus of over 1.39 million sentences. Finally, we use a politeness classifier (Niu and Bansal, 2018) to assign politeness scores to these sentences and filter them into ten buckets based on the score ($P_0$-$P_9$; Fig. 1). All the buckets are further divided into train, test, and dev splits (in a 80:10:10 ratio).

For our experiments, we assumed all the sentences with a politeness score of over 90% by the classifier to be polite, also referred as the $P_9$ bucket (marked in green in Fig. 1). We use the train-split of the $P_9$ bucket of over 270K polite sentences as the training data for the politeness transfer task. Since the goal of the task is making action directives more polite, we manually curate a test set comprising of such sentences from test splits across the buckets. We first train a classifier on the switchboard corpus (Jurafsky et al., 1997) to get dialog state tags and filter sentences that have been labeled as either action-directive or quotation.[3] Further, we use human annotators to manually select the test sentences. The annotators had a Fleiss's Kappa score ($\kappa$) of 0.77[4] and curated a final test set of 800 sentences.

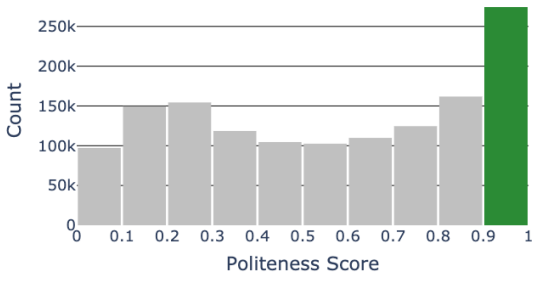


Figure 1: Distribution of Politeness Scores for the Enron Corpus

In Fig. 2, we examine the two extreme buckets with politeness scores of $< 10\%$ ($P_0$ bucket) and $> 90\%$ ($P_9$ bucket) from our corpus by plotting

[1]Pre-processing also involved steps for tokenization (done using spacy (Honnibal and Montani, 2017)) and conversion to lower case.

[2]We prune the corpus by removing the sentences that 1) were less than 3 words long, 2) had more than 80% numerical tokens, 3) contained email addresses, or 4) had repeated occurrences of spurious characters.

[3]We used AWD-LSTM based classifier for classification of action-directive.

[4]The score was calculated for 3 annotators on a sample set of 50 sentences.

10 of the top 30 words occurring in each bucket. We clearly notice that words in the $P_9$ bucket are closely linked to polite style, while words in the $P_0$ bucket are mostly content words. This substantiates our claim that the task of politeness transfer is fundamentally different from other attribute transfer tasks like sentiment where both the polarities are clearly defined.

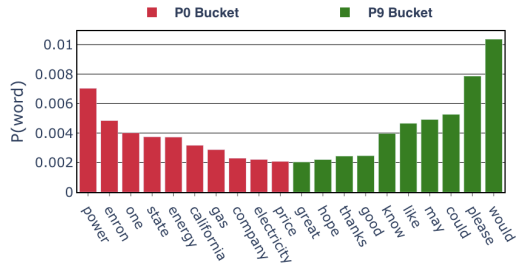


Figure 2: Probability of occurrence for 10 of the most common 30 words in the $P_0$ and $P_9$ data buckets

### 3.2 Other Tasks

The **Captions** dataset (Gan et al., 2017) has image captions labeled as being factual, romantic or humorous. We use this dataset to perform transfer between these styles. This task parallels the task of politeness transfer because much like in the case of politeness transfer, the captions task also involves going from a style neutral (factual) to a style rich (humorous or romantic) parlance.

For sentiment transfer, we use the **Yelp** restaurant review dataset (Shen et al., 2017) to train, and evaluate on a test set of 1000 sentences released by Li et al. (2018). We also use the **Amazon** dataset of product reviews (He and McAuley, 2016). We use the Yelp review dataset labelled for the **Gender** of the author, released by Prabhumoye et al. (2018) compiled from Reddy and Knight (2016). For the **Political** slant task (Prabhumoye et al., 2018), we use dataset released by Voigt et al. (2018).

## 4 Methodology

We are given non-parallel samples of sentences $\mathbf{X_1} = \{\mathbf{x}_1^{(1)} \dots \mathbf{x}_n^{(1)}\}$ and $\mathbf{X_2} = \{\mathbf{x}_1^{(2)} \dots \mathbf{x}_m^{(2)}\}$ from styles $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively. The objective of the task is to efficiently generate samples $\hat{\mathbf{X}}_\mathbf{1} = \{\hat{\mathbf{x}}_1^{(2)} \dots \hat{\mathbf{x}}_n^{(2)}\}$ in the target style $\mathcal{S}_2$, conditioned on samples in $\mathbf{X_1}$. For a style $\mathcal{S}_v$ where $v \in \{1, 2\}$, we begin by learning a set of phrases ($\Gamma_v$) which characterize the style $\mathcal{S}_v$. The presence of phrases from $\Gamma_v$ in a sentence $\mathbf{x}_i$ would associate the sentence with the style $\mathcal{S}_v$. For example, phrases like "pretty good" and "worth every penny" are characteristic of the "positive" style in the case of sentiment transfer task.

We propose a two staged approach where we first infer a sentence $z(\mathbf{x}_i)$ from $\mathbf{x}_i^{(1)}$ using a model, the tagger. The goal of the tagger is to ensure that the sentence $z(\mathbf{x}_i)$ is agnostic to the original style ($\mathcal{S}_1$) of the input sentence. Conditioned on $z(\mathbf{x}_i)$, we then generate the transferred sentence $\hat{\mathbf{x}}_i^{(2)}$ in the target style $\mathcal{S}_2$ using another model, the generator. The intermediate variable $z(\mathbf{x}_i)$ is also seen in other style-transfer methods. Shen et al. (2017); Prabhumoye et al. (2018); Yang et al. (2018); Hu et al. (2017) transform the input $\mathbf{x}_i^{(v)}$ to a latent representation $z(\mathbf{x}_i)$ which (ideally) encodes the content present in $\mathbf{x}_i^{(v)}$ while being agnostic to style $\mathcal{S}_v$. In these cases $z(\mathbf{x}_i)$ encodes the input sentence in a continuous latent space whereas for us $z(\mathbf{x}_i)$ manifests in the surface form. The ability of our pipeline to generate observable intermediate outputs $z(\mathbf{x}_i)$ makes it somewhat more interpretable than those other methods.

We train two independent systems for the tagger & generator which have complimentary objectives. The former identifies the style attribute markers $a(x_i^{(1)})$ from source style $\mathcal{S}_1$ and either replaces them with a positional token called [TAG] or merely adds these positional tokens without removing any phrase from the input $x_i^{(1)}$. This particular capability of the model enables us to generate these tags in an input that is devoid of any attribute marker (i.e. $a(x_i^{(1)}) = \{\}$). This is one of the major differences from prior works which mainly focus on removing source style attributes and then replacing them with the target style attributes. It is especially critical for tasks like politeness transfer where the transfer takes place from a non-polite sentence. This is because in such cases we may need to add new phrases to the sentence rather than simply replace existing ones. The generator is trained to generate sentences $\hat{\mathbf{x}}_i^{(2)}$ in the target style by replacing these [TAG] tokens with stylistically relevant words inferred from target style $\mathcal{S}_2$. Even though we have non-parallel corpora, both systems are trained in a supervised fashion as sequence-to-sequence models with their own distinct pairs of inputs & outputs. To create parallel training data, we first estimate the style markers $\Gamma_v$ for a given style $\mathcal{S}_v$ & then use these to curate style free sentences with [TAG]


Figure 3: Our proposed approach: *tag* and *generate*. The tagger infers the interpretable style free sentence $z(\mathbf{x}_i)$ for an input $\mathbf{x}_i^{(1)}$ in source style $\mathcal{S}_1$. The generator transforms $\mathbf{x}_i^{(1)}$ into $\hat{\mathbf{x}}_i^{(2)}$ which is in target style $\mathcal{S}_2$.

tokens. Training data creation details are given in sections §4.2, §4.3.

Fig. 3 shows the overall pipeline of the proposed approach. In the first example $\mathbf{x}_1^{(1)}$, where there is no clear style attribute present, our model adds the [TAG] token in $z(\mathbf{x}_1)$, indicating that a target style marker should be generated in this position. On the contrary, in the second example, the terms "ok" and "bland" are markers of negative sentiment and hence the tagger has replaced them with [TAG] tokens in $z(\mathbf{x}_2)$. We can also see that the inferred sentence in both the cases is free of the original and target styles. The structural bias induced by this two staged approach is helpful in realizing an interpretable style free tagged sentence that explicitly encodes the content. In the following sections we discuss in detail the methodologies involved in (1) estimating the relevant attribute markers for a given style, (2) tagger, and (3) generator modules of our approach.

### 4.1 Estimating Style Phrases

Drawing from Li et al. (2018), we propose a simple approach based on n-gram tf-idfs to estimate the set $\Gamma_v$, which represents the style markers for style $v$. For a given corpus pair $\mathbf{X}_1, \mathbf{X}_2$ in styles $\mathcal{S}_1, \mathcal{S}_2$ respectively we first compute a probability distribution $p_1^2(w)$ over the n-grams $w$ present in both the corpora (Eq. 2). Intuitively, $p_1^2(w)$ is proportional to the probability of sampling an n-gram present in both $\mathbf{X}_1, \mathbf{X}_2$ but having a much higher tf-idf value in $\mathbf{X}_2$ relative to $\mathbf{X}_1$. This is how we define the impactful style markers for style $\mathcal{S}_2$.

$$\eta_1^2(w) = \frac{\frac{1}{m}\sum_{i=1}^{m} \text{tf-idf}(w, \mathbf{x}_i^{(2)})}{\frac{1}{n}\sum_{j=1}^{n} \text{tf-idf}(w, \mathbf{x}_j^{(1)})} \quad (1)$$

$$p_1^2(w) = \frac{\eta_1^2(w)^\gamma}{\sum_{w'} \eta_1^2(w')^\gamma} \quad (2)$$

where, $\eta_1^2(w)$ is the ratio of the mean tf-idfs for a given n-gram $w$ present in both $\mathbf{X}_1, \mathbf{X}_2$ with $|\mathbf{X}_1| = n$ and $|\mathbf{X}_2| = m$. Words with higher values for $\eta_1^2(w)$ have a higher mean tf-idf in $\mathbf{X}_2$ vs $\mathbf{X}_1$, and thus are more characteristic of $\mathcal{S}_2$. We further smooth and normalize $\eta_1^2(w)$ to get $p_1^2(w)$. Finally, we estimate $\Gamma_2$ by

$$\Gamma_2 = \{w : p_1^2(w) \geq k\}$$

In other words, $\Gamma_2$ consists of the set of phrases in $\mathbf{X}_2$ above a given style impact $k$. $\Gamma_1$ is computed similarly where we use $p_2^1(w), \eta_2^1(w)$.

### 4.2 Style Invariant Tagged Sentence

The tagger model (with parameters $\theta_t$) takes as input the sentences in $\mathbf{X}_1$ and outputs $\{z(\mathbf{x}_i) : \mathbf{x}_i^{(1)} \in \mathbf{X}_1\}$. Depending on the style transfer task, the tagger is trained to either (1) identify and replace style attributes $a(\mathbf{x}_i^{(1)})$ with the token tag [TAG] (replace-tagger) or (2) add the [TAG] token at specific locations in $\mathbf{x}_i^{(1)}$ (add-tagger). In both the cases, the [TAG] tokens indicate positions where the generator can insert phrases from the target style $\mathcal{S}_2$. Finally, we use the distribution $p_1^2(w)/p_2^1(w)$ over $\Gamma_2/\Gamma_1$ (§4.1) to draw samples of attribute-markers that would be replaced with the [TAG] token during the creation of training data.

The first variant, replace-tagger, is suited for a task like sentiment transfer where almost every sentence has some attribute markers $a(\mathbf{x}_i^{(1)})$ present in it. In this case the training data comprises of pairs where the input is $\mathbf{X}_1$ and the output is $\{z(\mathbf{x}_i) : \mathbf{x}_i^{(1)} \in \mathbf{X}_1\}$. The loss objective for replace-tagger is given by $\mathcal{L}_r(\theta_t)$ in Eq. 3.

$$\mathcal{L}_r(\theta_t) = -\sum_{i=1}^{|\mathbf{X}_1|} \log P_{\theta_t}(z(\mathbf{x}_i)|\mathbf{x}_i^{(1)}; \theta_t) \quad (3)$$

The second variant, add-tagger, is designed for cases where the transfer needs to happen from style neutral sentences to the target style. That is, $\mathbf{X}_1$ consists of style neutral sentences whereas $\mathbf{X}_2$ consists of sentences in the target style. Examples of

such a task include the tasks of politeness transfer (introduced in this paper) and caption style transfer (used by Li et al. (2018)). In such cases, since the source sentences have no attribute markers to remove, the tagger learns to add [TAG] tokens at specific locations suitable for emanating style words in the target style.

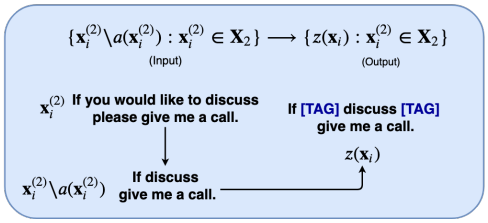


Figure 4: Creation of training data for add-tagger.

The training data (Fig. 4) for the add-tagger is given by pairs where the input is $\{\mathbf{x}_i^{(2)} \backslash a(\mathbf{x}_i^{(2)}) : \mathbf{x}_i^{(2)} \in \mathbf{X}_2\}$ and the output is $\{z(\mathbf{x}_i) : \mathbf{x}_i^{(2)} \in \mathbf{X}_2\}$. Essentially, for the input we take samples $\mathbf{x}_i^{(2)}$ in the target style $\mathcal{S}_2$ and explicitly remove style phrases $a(\mathbf{x}_i^{(2)})$ from it. For the output we replace the same phrases $a(\mathbf{x}_i^{(2)})$ with [TAG] tokens. As indicated in Fig. 4, we remove the style phrases "you would like to" and "please" and replace them with [TAG] in the output. Note that we only use samples from $\mathbf{X}_2$ for training the add-tagger; samples from the style neutral $\mathbf{X}_1$ are not involved in the training process at all. For example, in the case of politeness transfer, we only use the sentences labeled as "polite" for training. In effect, by training in this fashion, the tagger learns to add [TAG] tokens at appropriate locations in a style neutral sentence. The loss objective ($\mathcal{L}_a$) given by Eq. 4 is crucial for tasks like politeness transfer where one of the styles is poorly defined.

$$\mathcal{L}_a(\theta_t) = -\sum_{i=1}^{|\mathbf{X}_1|} \log P_{\theta_t}(z(\mathbf{x}_i)|\mathbf{x}_i^{(2)} \backslash a(\mathbf{x}_i^{(2)}); \theta_t) \quad (4)$$

### 4.3 Style Targeted Generation

The training for the generator model is complimentary to that of the tagger, in the sense that the generator takes as input the tagged output $z(\mathbf{x}_i)$ inferred from the source style and modifies the [TAG] tokens to generate the desired sentence $\hat{\mathbf{x}}_i^{(v)}$ in the target style $\mathcal{S}_v$.

$$\mathcal{L}(\theta_g) = -\sum_{i=1}^{|\mathbf{X}_v|} \log P_{\theta_g}(\mathbf{x}_i^{(v)}|z(\mathbf{x}_i); \theta_g) \quad (5)$$

The training data for transfer into style $\mathcal{S}_v$ comprises of pairs where the input is given by $\{z(\mathbf{x}_i) : x_i^{(v)} \in \mathbf{X}_v\,,\, v \in \{1, 2\}\}$ and the output is $\mathbf{X}_v$, i.e. it is trained to transform a style agnostic representation into a style targeted sentence. Since the generator has no notion of the original style and it is only concerned with the style agnostic representation $z(\mathbf{x}_i)$, it is convenient to disentangle the training for tagger & generator.

Finally, we note that the location at which the tags are generated has a significant impact on the distribution over style attributes (in $\Gamma_2$) that are used to fill the [TAG] token at a particular position. Hence, instead of using a single [TAG] token, we use a set of positional tokens $[\text{TAG}]_t$ where $t \in \{0, 1, \dots\ T\}$ for a sentence of length $T$. By training both tagger and generator with these positional $[\text{TAG}]_t$ tokens we enable them to easily realize different distributions of style attributes for different positions in a sentence. For example, in the case of politeness transfer, the tags added at the beginning ($t = 0$) will almost always be used to generate a token like "Would it be possible ..." whereas for a higher $t$, $[\text{TAG}]_t$ may be replaced with a token like "thanks" or "sorry."

## 5 Experiments and Results

**Baselines** We compare our systems against three previous methods. DRG (Li et al., 2018), Style Transfer Through Back-translation (BST) (Prabhumoye et al., 2018), and Style transfer from non-parallel text by cross alignment (Shen et al., 2017) (CAE). For DRG, we only compare against the best reported method, delete-retrieve-generate. For all the models, we follow the experimental setups described in their respective papers.

**Implementation Details** We use 4-layered transformers (Vaswani et al., 2017) to train both tagger and generator modules. Each transformer has 4 attention heads with a 512 dimensional embedding layer and hidden state size. Dropout (Srivastava et al., 2014) with p-value 0.3 is added for each layer in the transformer. For the politeness dataset the generator module is trained with data augmentation techniques like random word shuffle, word drops/replacements as proposed by (Im

| | Politeness | | | | Gender | | | | Political | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | BL-s | MET | ROU | Acc | BL-s | MET | ROU | ACC | BL-s | MET | ROU |
| CAE | **99.62** | 6.94 | 10.73 | 25.71 | 65.21 | 9.25 | 14.72 | 42.42 | 77.71 | 3.17 | 7.79 | 27.17 |
| BST | 60.75 | 2.55 | 9.19 | 18.99 | 54.4 | 20.73 | 22.57 | 55.55 | **88.49** | 10.71 | 16.26 | 41.02 |
| DRG | 90.25 | 11.83 | 18.07 | 41.09 | 36.29 | 22.9 | 22.84 | 53.30 | 69.79 | 25.69 | 21.6 | 51.8 |
| OURS | 89.50 | **70.44** | **36.26** | **70.99** | **82.21** | **52.76** | **37.42** | **74.59** | 87.74 | **68.44** | **45.44** | **77.51** |

Table 1: Results on the Politeness, Gender and Political datasets.

et al., 2017). We empirically observed that these techniques provide an improvement in the fluency and diversity of the generations. Both modules were also trained with the BPE tokenization (Sennrich et al., 2015) using a vocabulary of size 16000 for all the datasets except for Captions, which was trained using 4000 BPE tokens. The value of the smoothing parameter $\gamma$ in Eq. 2 is set to 0.75. For all datasets except Yelp we use phrases with $p_1^2(w) \geq k = 0.9$ to construct $\Gamma_2$, $\Gamma_1$ (§4.1). For Yelp $k$ is set to 0.97. During inference we use beam search (beam size=5) to decode tagged sentences and targeted generations for tagger & generator respectively. For the tagger, we re-rank the final beam search outputs based on the number of [TAG] tokens in the output sequence (favoring more [TAG] tokens).

**Automated Evaluation** Following prior work (Li et al., 2018; Shen et al., 2017), we use automatic metrics for evaluation of the models along two major dimensions: (1) style transfer accuracy and (2) content preservation. To capture accuracy, we use a classifier trained on the nonparallel style corpora for the respective datasets (barring politeness). The architecture of the classifier is based on AWD-LSTM (Merity et al., 2017) and a softmax layer trained via cross-entropy loss. We use the implementation provided by fastai.[5] For politeness, we use the classifier trained by (Niu and Bansal, 2018).[6] The metric of transfer accuracy **(Acc)** is defined as the percentage of generated sentences classified to be in the target domain by the classifier. The standard metric for measuring content preservation is BLEU-self **(BL-s)** (Papineni et al., 2002) which is computed with respect to the original sentences. Additionally, we report the BLEU-reference **(BL-r)** scores using the human reference sentences on the Yelp, Amazon and Captions datasets (Li et al., 2018). We also report ROUGE **(ROU)** (Lin, 2004) and METEOR **(MET)** (Denkowski and Lavie, 2011) scores. In particular, METEOR also uses synonyms and stemmed forms of the words in candidate and reference sentences, and thus may be better at quantifying semantic similarities.

Table 1 shows that our model achieves significantly higher scores on BLEU, ROUGE and METEOR as compared to the baselines DRG, CAE and BST on the Politeness, Gender and Political datasets. The BLEU score on the Politeness task is greater by 58.61 points with respect to DRG. In general, CAE and BST achieve high classifier accuracies but they fail to retain the original content. The classifier accuracy on the generations of our model are comparable (within 1%) with that of DRG for the Politeness dataset.

In Table 2, we compare our model against CAE and DRG on the Yelp, Amazon, and Captions datasets. For each of the datasets our test set comprises 500 samples (with human references) curated by Li et al. (2018). We observe an increase in the BLEU-reference scores by 5.25, 4.95 and 3.64 on the Yelp, Amazon, and Captions test sets respectively. Additionally, we improve the transfer accuracy for Amazon by 14.2% while achieving accuracies similar to DRG on Yelp and Captions. As noted by Li et al. (2018), one of the unique aspects of the Amazon dataset is the absence of similar content in both the sentiment polarities. Hence, the performance of their model is worse in this case. Since we don't make any such assumptions, we perform significantly better on this dataset.

While popular, the metrics of transfer accuracy and BLEU have significant shortcomings making them susceptible to simple adversaries. BLEU relies heavily on n-gram overlap and classifiers can be fooled by certain polarizing keywords. We test this hypothesis on the sentiment transfer task by a *Naive Baseline*. This baseline adds *"but overall it sucked"* at the end of the sentence to transfer it to negative sentiment. Similarly, it appends *"but overall it was perfect"* for transfer into a positive sentiment. This baseline achieves an average accuracy score of 91.3% and a BLEU score of 61.44 on the Yelp

[5] https://docs.fast.ai/

[6] This is trained on the dataset given by (Danescu-Niculescu-Mizil et al., 2013).

| | Yelp | | | | | Amazon | | | | | Captions | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc | BL-s | BL-r | MET | ROU | Acc | BL-s | BL-r | MET | ROU | Acc | BL-s | BL-r | MET | ROU |
| CAE | 72.1 | 19.95 | 7.75 | 21.70 | 55.9 | 78 | 2.64 | 1.68 | 9.52 | 29.16 | 89.66 | 2.09 | 1.57 | 9.61 | 30.02 |
| DRG | **88.8** | 36.69 | 14.51 | 32.09 | 61.06 | 52.2 | 57.07 | 29.85 | **50.16** | 79.31 | **95.65** | 31.79 | 11.78 | 32.45 | 64.32 |
| OURS | 86.6 | **47.14** | **19.76** | **36.26** | **70.99** | **66.4** | **68.74** | **34.80** | 45.3 | **83.45** | 93.17 | **51.01** | **15.63** | **43.67** | **79.51** |

Table 2: Results on the Yelp, Amazon and Captions datasets.

| | Con | | Att | | Gra | |
|---|---|---|---|---|---|---|
| | DRG | Ours | DRG | Ours | DRG | Ours |
| Politeness | 2.9 | **3.6** | 3.2 | **3.6** | 2.0 | **3.7** |
| Gender | 3.0 | **3.5** | - | - | 2.2 | **2.5** |
| Political | 2.9 | **3.2** | - | - | 2.5 | **2.7** |
| Yelp | 3.0 | **3.7** | 3 | **3.9** | 2.7 | **3.3** |

Table 3: Human evaluation on Politeness, Gender, Political and Yelp datasets.

dataset. Despite high evaluation scores, it does not reflect a high rate of success on the task. In summary, evaluation via automatic metrics might not truly correlate with task success.

**Changing Content Words** Given that our model is explicitly trained to generate new content only in place of the TAG token, it is expected that a well-trained system will retain most of the non-tagged (content) words. Clearly, replacing content words is not desired since it may drastically change the meaning. In order to quantify this, we calculate the fraction of non-tagged words being changed across the datasets. We found that the non-tagged words were changed for only 6.9% of the sentences. In some of these cases, we noticed that changing non-tagged words helped in producing outputs that were more natural and fluent.

**Human Evaluation** Following Li et al. (2018), we select 10 unbiased human judges to rate the output of our model and DRG on three aspects: (1) content preservation **(Con)** (2) grammaticality of the generated content **(Gra)** (3) target attribute match of the generations **(Att)**. For each of these metrics, the reviewers give a score between 1-5 to each of the outputs, where 1 reflects a poor performance on the task and 5 means a perfect output. Since the judgement of signals that indicate gender and political inclination are prone to personal biases, we don't annotate these tasks for target attribute match metric. Instead we rely on the classifier scores for the transfer. We've used the same instructions from Li et al. (2018) for our human study. Overall, we evaluate both systems on a total of 200 samples for Politeness and 100 samples each for Yelp, Gender and Political.

Table 3 shows the results of human evaluations. We observe a significant improvement in content preservation scores across various datasets (specifically in Politeness domain) highlighting the ability of our model to retain content better than DRG. Alongside, we also observe consistent improvements of our model on target attribute matching and grammatical correctness.

**Qualitative Analysis** We compare the results of our model with the DRG model qualitatively as shown in Table 4. Our analysis is based on the linguistic strategies for politeness as described in (Danescu-Niculescu-Mizil et al., 2013). The first sentence presents a simple example of the *counterfactual modal* strategy inducing *"Could you please"* to make the sentence polite. The second sentence highlights another subtle concept of politeness of *1st Person Plural* where adding *"we"* helps being indirect and creates the sense that the burden of the request is shared between speaker and addressee. The third sentence highlights the ability of the model to add *Apologizing* words like *"Sorry"* which helps in deflecting the social threat of the request by attuning to the imposition. According to the *Please Start* strategy, it is more direct and insincere to start a sentence with *"Please"*. The fourth sentence projects the case where our model uses *"thanks"* at the end to express gratitude and in turn, makes the sentence more polite. Our model follows the strategies prescribed in (Danescu-Niculescu-Mizil et al., 2013) while generating polite sentences.[7]

**Ablations** We provide a comparison of the two variants of the tagger, namely the replace-tagger and add-tagger on two datasets. We also train and compare them with a *combined* variant.[8] We train these tagger variants on the Yelp and Captions datasets and present the results in Table 5. We observe that for Captions, where we transfer a factual (neutral) to romantic/humorous sentence, the add-

[7]We provide additional qualitative examples for other tasks in the supplementary material.

[8]Training of combined variant is done by training the tagger model on the concatenation of training data for add-tagger and replace-tagger.

| Input | DRG Output | Our Model Output | Strategy |
|---|---|---|---|
| what happened to my personal station? | what happened to my mother to my co??? | could you please let me know what happened to my personal station? | Counterfactual Modal |
| yes, go ahead and remove it. | yes, please go to the link below and delete it. | yes, we can go ahead and remove it. | 1st Person Plural |
| not yet-i'll try this wkend. | not yet to say-i think this will be a <unk> long. | sorry not yet-i'll try to make sure this wk | Apologizing |
| please check on metromedia energy, | thanks again on the energy industry, | please check on metromedia energy, thanks | Mitigating please start |

Table 4: Qualitative Examples comparing the outputs from DRG and Our model for the Politeness Transfer Task

tagger provides the best accuracy with a relatively negligible drop in BLEU scores. On the contrary, for Yelp, where both polarities are clearly defined, the replace-tagger gives the best performance. Interestingly, the accuracy of the add-tagger is $\approx 50\%$ in the case of Yelp, since adding negative words to a positive sentence or vice-versa neutralizes the classifier scores. Thus, we can use the add-tagger variant for transfer from a polarized class to a neutral class as well.

To check if the combined tagger is learning to perform the operation that is more suitable for a dataset, we calculate the fraction of times the combined tagger performs add/replace operations on the Yelp and Captions datasets. We find that for Yelp (a polar dataset) the combined tagger performs 20% more replace operations (as compared to add operations). In contrast, on the CAPTIONS dataset, it performs 50% more add operations. While the combined tagger learns to use the optimal tagging operation to some extent, a deeper understanding of this phenomenon is an interesting future topic for research. We conclude that the choice of the tagger variant is dependent on the characterstics of the underlying transfer task.

| | Yelp | | Captions | |
|---|---|---|---|---|
| | Acc | BL-r | Acc | BL-r |
| Add-Tagger | 53.2 | 20.66 | **93.17** | 15.63 |
| Replace-Tagger | **86.6** | 19.76 | 84.5 | 15.04 |
| Combined | 72.5 | **22.46** | 82.17 | **18.5**1 |

Table 5: Comparison of different *tagger* variants for Yelp and Captions datasets

## 6 Conclusion

We introduce the task of politeness transfer for which we provide a dataset comprised of sentences curated from email exchanges present in the Enron corpus. We extend prior works (Li et al., 2018; Sudhakar et al., 2019) on attribute transfer by introducing a simple pipeline – $tag$ & $generate$ which is an interpretable two-staged approach for content preserving style transfer. We believe our approach is the first to be robust in cases when the source is style neutral, like the "non-polite" class in the case of politeness transfer. Automatic and human evaluation shows that our approach outperforms other state-of-the-art models on content preservation metrics while retaining (or in some cases improving) the transfer accuracies.

## Acknowledgments

This material is based on research sponsored in part by the Air Force Research Laboratory under agreement number FA8750-19-2-0200. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory or the U.S. Government. This work was also supported in part by ONR Grant N000141812861, NSF IIS1763562, and Apple. We would also like to acknowledge NVIDIA's GPU support. We would like to thank Antonis Anastasopoulos, Ritam Dutt, Sopan Khosla, and, Xinyi Wang for the helpful discussions.

## References


Philip Bramsen, Martha Escobar-Molano, Ami Patel, and Rafael Alonso. 2011. Extracting social power relationships from natural language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 773–782, Stroudsburg, PA, USA. Association for Computational Linguistics.

Penelope Brown, Stephen C Levinson, and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Khyathi Chandu, Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2019. my way of telling a story: Persona based grounded story generation. In *Proceedings of the Second Workshop on Storytelling*, pages 11–21.

Liz Coppock. 2005. Politeness strategies in conversation closings. *unpublished paper available online at http://www. stanford. edu/˜ coppock/face. pdf (last accessed 23 December 2007)*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91. Association for Computational Linguistics.

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*.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

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. International World Wide Web Conferences Steering Committee.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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 of *Proceedings of Machine Learning Research*, pages 1587–1596, International Convention Centre, Sydney, Australia. PMLR.

Daniel Im Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. 2017. Denoising criterion for variational auto-encoding framework. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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.

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, Florence, Italy. Association for Computational Linguistics.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.

Dániel Z Kádár and Sara Mills. 2011. *Politeness in East Asia*. Cambridge University Press.

Dongyeop Kang and Eduard Hovy. 2019. xslue: A benchmark and analysis platform for cross-style language understanding and evaluation.

Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. In *CEAS*.

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. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Andrew McCallum, Xuerui Wang, and Andrés Corrada-Emmanuel. 2007. Topic and role discovery in social networks with experiments on enron and academic email. *J. Artif. Int. Res.*, 30(1):249–272.

Ardith J Meier. 1995. Defining politeness: Universality in appropriateness. *Language Sciences*, 17(4):345–356.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models. *arXiv preprint arXiv:1708.02182*.

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

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

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: A case study on the Enron corpus. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 86–95, Portland, Oregon. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Emily E. Reid, and Owen Rambow. 2014. Gender and power: How gender and gender environment affect manifestations of power. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1965–1976, Doha, Qatar. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. "transforming" delete, retrieve, generate approach for controlled text style transfer. 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 3267–3277, Hong Kong, China. Association for Computational Linguistics.

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

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Mask and infill: Applying masked language model for sentiment transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5271–5277. International Joint Conferences on Artificial Intelligence Organization.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations. *arXiv preprint arXiv: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.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems*, pages 7287–7298.

Jen-Yuan Yeh and Aaron Harnly. 2006. Email thread reassembly using similarity matching. In *Conference on Email and Anti-Spam*. Conference on Email and Anti-Spam.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational*

*Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

| Non-polite Input | DRG | Our Model |
|---|---|---|
| jon - - please use this resignation letter in lieu of the one sent on friday . | - i think this would be a good idea if you could not be a statement that harry 's signed in one of the schedule . | jon - sorry - please use this resignation letter in lieu of the one event sent on |
| if you have a few minutes today, give me a call | i'll call today to discuss this. | if you have a few minutes today, please give me a call at |
| anyway you can let me know. | anyway, i'm sure i'm sure. | anyway please let me know as soon as possible |
| yes, go ahead and remove it. | yes, please go to the link below and delete it. | yes, we can go ahead and remove it. |
| can you explain a bit more about how those two coexist ? also ..... | i can explain how the two more than <unk> i can help with mike ? | can you explain a bit more about how those two coexist ? also thanks |
| go ahead and sign it - i did . | go away so we can get it approved . | we could go ahead and sign it - i did look at |

Table 6: Additional Qualitative Examples of outputs from our Model and DRG for the Politeness Transfer Task

| Task | Non-polite Input | DRG | Our Model |
|---|---|---|---|
| Fem → Male | my husband ordered the brisket . | my wife had the best steak . | my wife ordered the brisket . |
| Fem → Male | i ' m a fair person . | i ' m a good job of the <unk> . | i ' m a big guy . |
| Male → Fem | my girlfriend and i recently stayed at this sheraton . | i recently went with the club . | my husband and i recently stayed at this office . |
| Male → Fem | however , once inside the place was empty . | however , when the restaurant was happy hour for dinner . | however , once inside the place was super cute . |
| Pos → Neg | good drinks , and good company . | horrible company . | terrible drinks , terrible company. |
| Pos → Neg | i will be going back and enjoying this great place ! | i will be going back and enjoying this great ! | i will not be going back and enjoying this garbage ! |
| Neg → Pos | this is the reason i will never go back . | this is the reason i will never go back . | so happy i will definitely be back . |
| Neg → Pos | salsa is not hot or good . | salsa is not hot or good . | salsa is always hot and fresh . |
| Dem → Rep | i am confident of trumps slaughter . | i am mia love | i am confident of trumps administration . |
| Dem → Rep | we will resist trump | we will impeach obama | we will be praying for trump |
| Rep → Dem | video : black patriots demand impeachment of obama | video : black police show choose | video : black patriots demand to endorse obama |
| Rep → Dem | mr. trump is good ... but mr. marco rubio is great ! ! | thank you mr. good ... but mr. kaine is great senator ! ! | mr. schumer is good ... but mr. pallone is great ! ! |
| Fact → Rom | a woman is sitting near a flower bed overlooking a tunnel . | a woman is sitting near a flower overlooking a tunnel, determined to | a woman is sitting near a brick rope , excited to meet her boyfriend . |
| Fact → Rom | two dogs play with a tennis ball in the snow . | two dogs play with a tennis ball in the snow . | two dogs play with a tennis ball in the snow celebrating their friendship . |
| Fact → Hum | three kids play on a wall with a green ball . | three kids on a bar on a field of a date . | three kids play on a wall with a green ball fighting for supremacy . |
| Fact → Hum | a black dog plays around in water . | a black dog plays in the water . | a black dog plays around in water looking for fish . |

Table 7: Additional Qualitative Examples of our Model and DRG for other Transfer Tasks