Unsupervised Aspect-Level Sentiment Controllable Style Transfer

Mukuntha N S[§], Zishan Ahmad[§], Asif Ekbal, Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Bihar, India
{mukuntha.cs16,1821cs18,asif,pb}@iitp.ac.in

Abstract

Unsupervised style transfer in text has previously been explored through the sentiment transfer task. The task entails inverting the overall sentiment polarity in a given input sentence, while preserving its content. From the Aspect-Based Sentiment Analysis (ABSA) task, we know that multiple sentiment polarities can often be present together in a sentence with multiple aspects. In this paper, the task of aspect-level sentiment controllable style transfer is introduced, where each of the aspect-level sentiments can individually be controlled at the output. To achieve this goal, a BERT-based encoder-decoder architecture with saliency weighted polarity injection is proposed, with unsupervised training strategies, such as ABSA masked-language-modelling. Through both automatic and manual evaluation, we show that the system is successful in controlling aspect-level sentiments.

1 Introduction

With a rapid increase in the quality of generated text, due to the rise of neural text generation models (Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Sutskever et al., 2014; Vaswani et al., 2017), controllable text generation is quickly becoming the next frontier in the field of text generation. Controllable text generation is the task of generating realistic sentences whose attributes can be controlled. The attributes to control can be: (i). *Stylistic:* Like politeness, sentiment, formality etc, (ii). *Content:* Like information, entities, keywords etc. or (iii). *Ordering:* Like ordering of information, events, plots etc.

Controlling sentence level polarity has been well explored as a style transfer task. Zhang et al. (2018) used unsupervised machine translation techniques for polarity transfer in sentences. Yang et al. (2018) used language models as discriminators to achieve style (polarity) transfer in sentences. Li et al. (2018a) proposed a simpler method where they deleted the attribute markers and devise a method to replace or generate the target attribute-key phrases in the sentence.

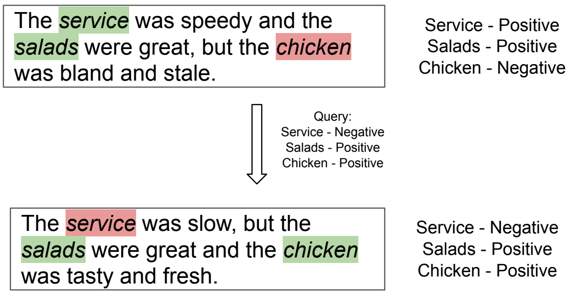
The service was speedy and the salads were great, but the chicken was bland and stale.
Service - Positive
Salads - Positive
Chicken - Negative
Query:
Service - Negative
Salads - Positive
Chicken - Positive
The service was slow, but the salads were great and the chicken was tasty and fresh.
Service - Negative
Salads - Positive
Chicken - Positive

Figure 1: An example of the proposed aspect-level sentiment style transfer task

In this paper we explore a more fine-grained style transfer task, where each aspect's polarities can be changed individually. Recent interest in Aspect-Based Sentiment Analysis (ABSA) (Pontiki et al., 2014) has shown that sentiment information can vary within a sentence, with differing sentiments expressed towards different aspect terms of target entities (e.g. 'food', 'service' in a restaurant domain). We introduce the task of aspect-level sentiment transfer - the task of rewriting sentences to transfer them from a given set of aspect-term polarities (such as 'positive sentiment' towards the service of a restaurant and a 'positive sentiment' towards the taste of the food) to a different set of aspect-term polarities (such as 'negative sentiment' towards the service of a restaurant and a 'positive' sentiment towards the taste of the food). This is a more challenging task than regular style transfer as the style attributes here are not the overall attributes for the whole sentence, but are localized to specific parts of the sentence, and multiple opposing at-

[§] equal contribution

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 303–312
December 4 - 7, 2020. ©2020 Association for Computational Linguistics

tributes could be present within the same sentence. The target of the transformation made needs to be localized and the other content expressed in the rest of the sentence need to be preserved at the output. An example of the task is shown in Figure 1.

For successful manipulation of the generated sentences, a few challenges need to be addressed: (i). The model should learn to associate the right polarities with the right aspects. (ii). The model needs to be able to correctly process the aspect-polarity query and accordingly delete, replace and generate text sequence to satisfy the query. (iii). The polarities of the aspects not in the query should not be affected. (iv). The non-attribute content and fluency of the text should be preserved.

We explore this task in an unsupervised setting (as is common with most style-transfer tasks due to the lack of an aligned parallel corpus) using only monolingual unaligned corpora. In this work, a novel encoder-decoder architecture is proposed to perform unsupervised aspect-level sentiment transfer. A BERT (Devlin et al., 2019) based encoder is used that is trained to understand aspect-specific polarity information. We also propose using a 'polarity injection' method, where saliency-weighted aspect-specific polarity information is added to the hidden representations from the encoder to complete the query for the decoder.

1.1 Motivation

The Aspect-Based Sentiment Analysis (ABSA) task shows that differing sentiments can be present within the same sentence, localized to different entities or parts of the text. The notion of styles in natural language can be used to refer to the attributes, such as sentiment, formality in content, emotion, sarcasm, etc. Similar to the sentiment, these other attributes can also be present localized to different entities taking differing values at each location. If we consider the style 'emotion' with the example "Although Alice infuriates me with her prattle and Bob scares me, I am quite happy about how things are turning out." - A single piece of text (such as a single sentence) can express an emotion, such as 'happiness' about an event while expressing 'fear' towards some entity and 'anger' towards a second entity. This shows that style transfer in language needs a more nuanced understanding. Especially when generating larger pieces of text, multiple such styles could intermingle, and differing styles can often be present together when discussing different topics and entities. Our work intends to take the first step towards a more controllable form of fine-grained style transfer with the task of aspect-level sentiment style transfer.

2 Related Work

In this section we present an overview of the related literature.

2.1 Sentiment Transfer

To the best of our knowledge, our current work is the first to tackle aspect-level sentiment transfer. Most of the previous works involving sentiment transfer (Li et al., 2018b; Yang et al., 2018; Shen et al., 2017; Xu et al., 2018; Prabhumoye et al., 2018; Wu et al., 2019) consider the style that is present throughout the sentence and seek to transfer only the overall sentiment polarities expressed. Tian et al. (2018) proposed a new training objective for content preservation during style transfer. They used Part-of-Speech (PoS) tagging to collect nouns at inputs, and expect them to be present at the output for content preservation. To achieve this, they proposed a PoS preservation constraint and 'Content Conditional Language Modelling'. They tested their system on sentiment style transfer task.

Wang et al. (2019) proposed a method that can also control the degree of polarity transfer in a sentence with multiple aspect categories present in it. Unlike their task which deals with *predefined aspect categories*, our task deals with *opinion target expressions*. Aspect categories are coarse entities that are few in number and predefined for a certain domain, while aspect-terms or opinion target expressions are fine-grained entities that are present in the text. They also did not investigate selectively transferring the polarity over a subset of aspects with multiple differing polarities at the output and only invert the overall polarity expressed by the sentence. Our method works across thousands of unique opinion target expressions (Table 1 shows the number unique target aspects present in each of our datasets).Our method also does not need these to be predefined, and so could be used to control the polarities of previously unseen target expressions as well.

2.2 Unsupervised Machine Translation

Previous works in unsupervised neural machine translation (Artetxe et al., 2017) and unsupervised style transfer (Zhang et al., 2018) have shown that, with only monolingual data, using a denoising auto-encoder loss and an on-the-fly back-translation loss can be very successful in achieving transfer. Both of these training steps are used as part of our method to train the network in an unsupervised fashion.

2.3 Natural Language Generation Architecture

Lai et al. (2019) proposed an adversarial training mechanism for Gated Recurrent Unit (GRU) based encoder-decoder model for sentiment polarity transfer and multiple-attribute transfer tasks. They split the training mechanism of their model into two phases, *viz.* (i). Style transfer phase and (ii). Reconstruction phase. Pryzant et al. (2020) proposed a method to remove subjective bias in the sentences. They proposed adding a 'join-embedding' weighted by a word subjective-bias probability to automatically edit the hidden states from the encoder. We adapt this 'join-embedding' method to inject weighted polarities into our encoder outputs as described in Section 3.5.

2.4 Aspect Based Sentiment Analysis

Aspect based sentiment analysis (ABSA) has been explored in a series of SemEval shared tasks. The task consists of both aspect term extraction and aspect sentiment prediction. Tay et al. (2018) proposed 'Aspect Fusion LSTM' to attend on the associative relationships between sentence words and aspect words to classify aspect polarities. Xu et al. (2019) proposed BERT based models for aspect term extraction and aspect-polarity classification tasks. We build similar BERT based aspect term-extraction and aspect-polarity classification models and use them to label Yelp reviews dataset. This dataset is then used for aspect-level sentiment controllable style transfer task in this paper.

3 Methodology

3.1 Problem Statement

Let us assume we have access to a corpora of labelled sentences $D = (x_1, l_1) \dots (x_n, l_n)$, where x_i is a sentence, and $l_i = \{(t_{i1}, p_{i1}) \dots (t_{im}, p_{im})\}$. Here, t_{ij} is an aspect-target or 'Opinion Target Expression' (Pontiki et al., 2014), and p_{ij} is the corresponding sentiment-polarity expressed towards t_{ij}, where $p_{ij} \in \{\text{``}positive\text{''}, \text{``}negative\text{''}\}$. A model is to be learned that takes as input (x, l_{tgt}) where x is the source sentence expressing some aspect-polarity set l_{src}, and outputs y that retain all the non-polarity content in x while expressing the aspect-polarity set l_{tgt}.

This is to be performed in an unsupervised manner, where we do not assume access to an aligned set of parallel sentences with the same content but different aspect-polarities.

The overall architecture consists of a Transformer (Vaswani et al., 2017) encoder-decoder neural network, where the encoder is BERT (Devlin et al., 2019). In this section, we describe the architecture and the training methodology used. The inputs provided to the model are the sentence, a list of aspects, their corresponding desired target polarities l_{tgt} and their corresponding per-token weights (explained in Section 3.5).

3.2 ABSA Input Representation

The BERT model (Devlin et al., 2019) was originally trained with two objectives: (i). A cloze objective where the classifier predicts missing words in a sequence, and (ii). A sentence-pair classification objective. For the sentence-pair objective, BERT was trained to take inputs as segment-pairs, where each segment has a different embedding added to it and are separated by a $[SEP]$ token. For our input representation, we construct such segment-pairs. The first segment consists of an aspect-polarity sequence SEG_A = "$T_1 P_1 [SEP_{ASP}] \dots T_k P_k [SEP_{ASP}]$", where T_i is the tokenized target aspect term and $P_i \in \{[POS], [NEG]\}$ is a polarity corresponding to it. $[SEP_{ASP}]$" is a separator token. $[POS]$, $[NEG]$ are the special tokens corresponding to the 'positive' and 'negative' sentiments, for which the unused tokens from the BERT vocabulary were used. The second segment SEG_B consists of a sentence expressing some sentiment towards these targets.

3.3 Preconditioning the BERT-encoder for ABSA Input Representations

We precondition the BERT encoder to better understand the ABSA task and to learn the token-embeddings for $[POS]$ and $[NEG]$ with MLM pre-training (a cloze objective) (c.f. Figure 2). For each data instance, with an equal probability, we randomly mask out either (i). all the polarity tokens from aspect-polarity sequence (SEG_A), or (ii). random tokens from the sentence (SEG_B) and train the encoder to correctly predict the masked-

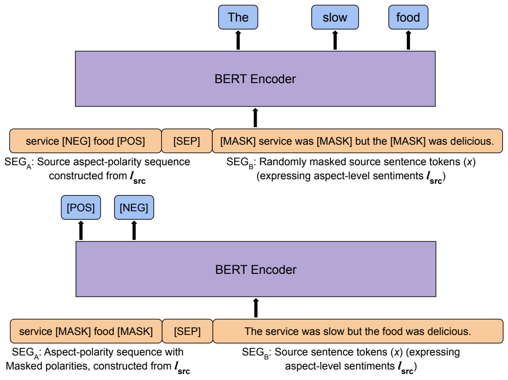
The
slow
food
BERT Encoder
service [NEG] food [POS]
[SEP]
[MASK] service was [MASK] but the [MASK] was delicious.
SEG_A: Source aspect-polarity sequence constructed from l_src
SEG_B: Randomly masked source sentence tokens (x) (expressing aspect-level sentiments l_src)
[POS]
[NEG]
BERT Encoder
service [MASK] food [MASK]
[SEP]
The service was slow but the food was delicious.
SEG_A: Aspect-polarity sequence with Masked polarities, constructed from l_src
SEG_B: Source sentence tokens (x) (expressing aspect-level sentiments l_src)

Figure 2: BERT Encoder pre-training

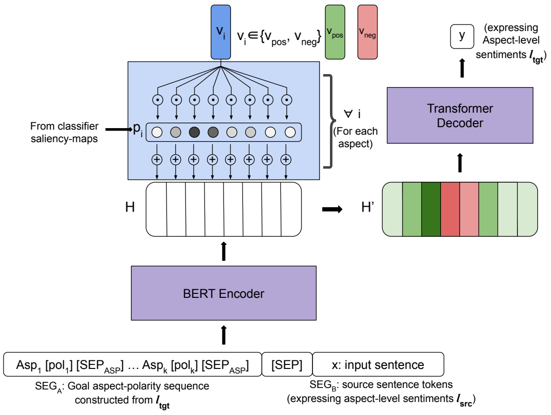
v_i
v_i∈{v_pos, v_neg}
v_pos
v_neg
y
(expressing Aspect-level sentiments l_tgt)
From classifier saliency-maps
p_i
∀ i (For each aspect)
Transformer Decoder
H
H'
BERT Encoder
Asp_1 [pol_1] [SEP_ASP] … Asp_k [pol_k] [SEP_ASP]
[SEP]
x: input sentence
SEG_A: Goal aspect-polarity sequence constructed from l_tgt
SEG_B: source sentence tokens (expressing aspect-level sentiments l_src)

Figure 3: The encoder-decoder network used, with the polarity injection.

out tokens. When the polarities get masked, the encoder learns to correctly understand the aspect-level sentiment polarities from a sentence. When the words from the sentence get masked, the encoder also learns to correctly predict attribute markers corresponding to a given aspect and a sentiment. For example, it would learn associations between the markers, such as 'personable' (or 'rude') when given an aspect-term, such as 'staff' with a polarity '[POS]' (or '[NEG]') as opposed to an aspect-marker, such as 'delicious', which cannot be used with 'staff'.

3.4 Encoder-Decoder Architecture

To convert a sentence from one set of aspect-level polarities to another, the input to the encoder consists of the target aspect-level polarities l_{tgt} as SEG_A with the source sentence x passed as SEG_B. The full architecture is shown in Figure 3. The source sentence x expresses some source aspect-level polarities l_{src} which is not provided to the model. The polarity-injection (explained in Section 3.5) adds the weighted target polarities l_{tgt} into the hidden-representation H from the BERT-encoder to obtain H' which is passed to the decoder. The decoder is trained to output the target sentence y which consists of the same content as present in x but expressing the target aspect-level polarities l_{tgt}. This architecture is trained in an unsupervised fashion as explained in Section 3.6.

3.5 Polarity Injection

In Pryzant et al. (2020), authors showed that the hidden states of the encoder can be edited by adding weighted vectors to indicate subjective-bias, before being input to the decoder. They proposed this as a method to join the results from two sub-modules in their system. Here, we extend this to cover multiple attributes - the 'positive' and 'negative' sentiments, and substitute the supervised model they train with saliency-based weights. We inject (add) weighted amounts of two vectors corresponding to these two attributes to edit the hidden states output by the encoder. For each aspect, the vector added corresponds to the desired target polarity of this aspect, and the amount added to a given token depends on the saliency-based weight for this token calculated from the gradient for this aspect's polarity from a classification model (described in Section 4.1.1).

More formally, the polarity injection is calculated from equation 1. $H = [h_1, h_2, \dots h_k]$ that denotes the hidden-state output from the encoder, and $H' = [h'_1, h'_2, \dots h'_k]$ are the new hidden-states calculated after polarity-injection. The number p_{ij} denotes the saliency-based weightage for token j with respect to aspect i. Figure 3 shows the polarity-injection architecture. v_{pos} and v_{neg} are the special vector-embeddings, which have the same size as the hidden dimension, and trained to denote the positive and negative sentiment, respectively.

$$h'_j = h_j + \sum_{i=1}^{k} p_{ij} \cdot v_i \tag{1}$$

$$v_i = \begin{cases} v_{pos} & \text{if } pol_i \text{ desired is positive} \\ v_{neg} & \text{if } pol_i \text{ desired is negative} \end{cases} \tag{2}$$

where pol_i is the target (desired) polarity from l_{tgt} for the i^th aspect-term. For calculating p_{ij}, saliency-maps obtained for each aspect from the polarity classifier described in 4.1 are used. Saliency-maps (Simonyan et al., 2014) are calculated with the gradient of the loss at the input, as given in equation

Dataset	No. of Sentences	No. of Target Aspects	No. of Unique Target Aspects
SemEval (Train and Validation)	2,242	4,016	1,437
SemEval (Test)	401	513	269
Yelp (Train and Validation)	361,968	471,820	47,750

Table 1: Data distribution for the restaurant domain. The Yelp dataset does not contain target aspects and their polarities extracted, and these were extracted with a classifier trained on the SemEval training data

3. The s_{tok} values for all the tokens tok in the sentence x are normalized between 0 and 1 for each (target t, sentence x) pair to obtain the p_{ij} values. Since the saliency-maps produce high values for the tokens that are important in calculating the sentiment's polarity, adding the 'positive' or 'negative' embedding weighted by these probabilities would provide hints to the decoder about the important words to be rewritten with the required sentiment. The p_{ij} values over Segment-A is set to 1 over all the tokens corresponding to the i$^{\text{th}}$ aspect term ($T_i P_i$) (see Section 3.2) and 0 over the other tokens.

$$s_{tok} = \left| \frac{\partial L(y_t; x, t, \theta)}{\partial emb_{tok}} \right|; \forall tok \in x \qquad (3)$$

One-Zero Alternative to Saliency: To test for performance in the absence of any saliency information, we also propose using a one-zero setup. Here, p_{ij} is set to 1 over the tokens corresponding to the i$^{\text{th}}$ aspect-term and 0 elsewhere. So in this setup, v_{pos} gets added to the tokens corresponding to the positive aspects and v_{neg} gets added to the tokens corresponding to the negative aspects. For example, in Figure 1, v_{pos} gets added to the sub-word tokens corresponding to the word 'salads' and 'chicken', while v_{neg} gets added to the sub-word tokens corresponding to the word 'service'.

3.6 Unsupervised Training

For training the model in an unsupervised setting, we alternate training steps between a denoising auto-encoding objective and a back-translation objective. During the denoising step, we add random noise to the sentence part of the input SEG_B. We also randomly mask the polarities in the aspect-polarity sequence in SEG_A with a small probability to ensure the model learns to generate outputs using the polarity injection clues. During the back-translation step, a random query l_{tgt} aspect-polarity sequence is used to produce an intermediate translation (using the model), and the same model is trained to regenerate the original input when provided the aspect-polarity sequence from the original input sentence.

4 Experiments

In this section we report the datasets used for the experiments and the implementation details.

4.1 Datasets

Text generation tasks require huge amounts of data, however there are no aspect-sentiment annotated datasets that are large enough for our task. Fortunately, aspect extraction and aspect-sentiment classification tasks have been well explored and have several publicly available datasets. We used datasets (only restaurant domain) from SemEval 2014, 2015 and 2016 (ABSA task) to train BERT based aspect extraction and aspect-sentiment classification systems. We only consider positive and negative polarities for our experiments.

For the task of aspect-level sentiment style transfer, we use Yelp dataset. Since this dataset does not contain aspect-level polarity information or the target-aspects extracted, we use our BERT-based target-extraction model and BERT-based polarity classification model which were trained on the SemEval ABSA training data, to generate aspect-level sentiment data from the Yelp reviews dataset. Table 1 shows some statistics from the datasets.

4.1.1 Aspect based Sentiment Analysis with BERT

A pipeline of BERT-based models was trained for target-extraction and aspect-level polarity classification over the SemEval dataset. These are the models used to extract target-aspects and their polarities from the Yelp dataset. The target extraction task was posed as a sequential token classification problem with BERT using the IOB2 format (SANG, 1999). This BERT model was fed the whole sentence as the input segment and it obtained an F1-score of 0.8012 (evaluation carried out similar to Sang and Buchholz (2000)). The sentiment-polarity prediction task is posed as a sentence-pair classification problem using BERT, with the sentence provided as the first segment and the aspect-term as the second segment. This model obtained an F1-score of 0.9080 for the positive po-

Model	Classifier Score (Overall)	Classifier Score (1-Aspect)	Classifier Score (2-Aspects)	Classifier Score (3-or-more Aspects)	BLEU Score
BERT-Baseline (BB)	0.5158	0.4983	0.5448	0.5036	36.0683
BB + MLM pretraining (BB-MLM)	0.5298	0.5433	0.5310	0.5145	35.4601
BB-MLM + one-zero polarity injection	0.5415	0.5675	0.5276	0.5290	35.8244
BB-MLM + saliency-based polarity injection	**0.5918**	**0.6125**	**0.5828**	**0.5797**	**39.3838**

Table 2: Results of automatic evaluation. The overall classifier score is calculated over all queries. The other columns show the score calculated only on queries with one aspect, two aspects or three or more aspects. The classifier scores are calculated on the full test set, while the BLEU scores are measured with reference-outputs for a subset of 100 queries.

Model	Att	Con	Gra
BERT-Baseline (BB)	2.48	3.99	3.96
BB + MLM pretraining (BB-MLM)	2.64	3.95	4.04
BB-MLM + one-zero polarity injection	2.80	4.00	**4.05**
BB-MLM + saliency-based polarity injection	**2.98**	**4.08**	**4.05**

Table 3: Results of manual evaluation. Here, 'Att' stands for attribute match, 'Con' stands for content preservation and 'Gra' stands for grammaticality or fluency. Manual evaluation is performed on a subset of 100 queries from all the test set queries, and averaged scores are shown.

larity and 0.8239 for the negative polarity on the ABSA restaurant dataset. Using this classifier, for each (Sentence, Target) pair the gradient of the loss was taken at the input token embeddings and normalized to obtain the saliency-based weights used for polarity-injection.

4.2 Implementation Details

All the models were implemented using PyTorch (Paszke et al., 2017). The BERT model was implemented using the transformers library (Wolf et al., 2019). Models are trained with an initial learning rate of 1e-4 with a linear schedule and a warmup (Vaswani et al., 2017), using the Adam Optimizer (Kingma and Ba, 2019). Mini-batches of size 32 were used during training. A linear schedule was used for the weight of the loss from the denoising auto-encoding step, which was set to decrease from 1 to 0.1 for the first 30,000 optimization steps and then decrease linearly to 0 over the next 70,000 steps. The models were each trained for 8 epochs on the Yelp dataset. The random masking probability used during pre-training was 0.25. During the denoising step, a probability of 0.25 was used for dropping words, and words were shuffled with a window-size of 3.

5 Results and Analysis

5.1 Evaluation

The evaluation metrics we use are an extension of the metrics used for evaluating the sentiment transfer task by previous work (such as Li et al. (2018b); Wang et al. (2019)). The evaluation was done with the SemEval test dataset. Queries were generated from this data by randomly inverting a subset (non-null, improper subset) of the polarities expressed at the input. For queries with 2 or more aspects, as many queries were generated as there were aspects in the sentence with different random inversions, resulting in a total of 513 evaluation queries. A sample consisting of 100 queries from the test set was used for manual evaluation.

5.1.1 Automatic Evaluation

For automatic evaluation, we use a classifier score and a BLEU score. The results for automatic evaluation are shown in Table 2.

Classifier Score: We use an aspect-level sentiment polarity classifier to measure how many of the outputs express the necessary target polarities (Li et al., 2018b). We use the classifier described in 4.1.1 for the polarity prediction. We define the *classifier score* to be the fraction of aspect-level sentiment polarities (predicted by the classifier from the output) that match with the desired aspect-level polarity (from the query). While averaging, each

Input	Query	Model Output
overall, decent food at a good price, with friendly people.	food - negative people - positive	overall , mediocre food at a good price , with friendly people .
	food - positive people - negative	overall , decent food at a good price , with rude people .
the waiter was attentive , the food was delicious and the views of the city were great	waiter - negative food - positive views of the city - positive	the waiter was inattentive , the food was delicious and the views of the city were great .
	waiter - positive food - negative views of the city - positive	the waiter was attentive , the food was disappointing but the views of the city were great .
	waiter - positive food - negative views of the city - negative	the waiter was attentive , the food was disappointing but the views of the city were terrible .

Table 4: Example outputs from the full model with saliency-based polarity injection with different aspect-level polarity queries.

aspect-level sentiment in a query was treated as a separate instance.

BLEU Scores: Like in Li et al. (2018b); Gan et al. (2017), human reference outputs were written for 100 of the queries. Three Human experts were asked to rewrite the reviews with as much content preserved as possible, without compromising fluency. These experts had good language abilities and having satisfactory knowledge in the relevant area. We report BLEU scores for the models against these references. A BLEU score could be treated as a measure of content preservation from the input or the output fluency.

5.1.2 Manual Evaluation

Following the previous methods (Li et al., 2018b; Wang et al., 2019) for manual evaluation of style transfer, workers were asked to rate the output sentences on the Likert-scale (1 to 5) for three criteria - Attribute match to the query set of aspect-level polarities (Att), Fluency (Gra) measuring the naturalness of the output and Content preservation (Con). They were shown the source sentence, the query aspect-level polarities and the model output. The results of manual evaluation are shown in Table 3 [1]

5.2 Error Analysis

The importance of each component in our model is shown through an ablation study in Table 2 and Table 3. From the classifier-based score, we see that the full model with saliency-based polarity injection is the most successful in transferring sentiment-level polarities. Polarity injection, even without saliency information is seen to be useful. The models with polarity injection are especially better at transferring sentiments when three or more aspects are present, showing that the polarity signals are useful in localizing the style attributes with multiple targets present. The model using saliency-based weighting for the polarity injection has a significantly higher classifier score. This could be because of the saliency information acting like an adversarial white-box attack on the classifier, making it easier to obtain higher classifier scores.

The Content preservation (Con) scores and BLEU scores for the baseline models are significantly high, but these models also show poor Attribute match (Att) scores. This means that many of the sentiments at the output were left untransferred resulting in the poor Att score, while large parts of the input text were copied over to the output resulting in the larger BLEU and Content preservation scores. The improved Content preservation (Con) scores and the Fluency (Gra) from the model without saliency information to the model with saliency-based weighting shows that the attribute transfer with saliency-based info is more successful in inverting the correct polarities, while maintaining the content and fluency, due to the added information about the words to be edited.

The Table 5 shows how the model outputs change with different components of the model are ablated. With queries involving mostly positive or negative attributes, the saliency-based polarity injection supports the localized inversion of sentiment in the output. Outputs also show how polarity injection helps produce the required change with more content and fluency preserved, by selectively editing the correct words.

[1] Inter-annotator agreement measured through the Krippendorff's alpha was found to be 0.92, 0.82, 0.87 for 'Att', 'Con', and 'Gra' respectively.

Input	the veal and the mushrooms were cooked perfectly .
Query	veal - positive, mushrooms - negative
BERT-Baseline (BB)	the veal and the mushrooms were not cooked perfectly .
BB + MLM pretraining (BB-MLM)	the veal and the mushrooms were over cooked perfectly .
BB-MLM + one-zero polarity injection	the veal was gross and the mushrooms were over cooked .
BB-MLM + saliency-based polarity injection	loved the veal and the mushrooms were over cooked .
Input	the waiter was attentive , the food was delicious and the views of the city were great .
Query	waiter - negative, food - positive, views of the city - positive
BERT-Baseline (BB)	the waiter was attentive , the food was delicious and the city were great .
BB + MLM pretraining (BB-MLM)	the waiter was attentive , the food was delicious and the views of the city were great .
BB-MLM + one-zero Polarity injection	the waiter was attentive , the food was delicious and the views of the city were great .
BB-MLM + saliency-based polarity injection	the waiter was inattentive , the food was delicious and the views of the city were great .
Input	for 7 years they have put out the most tasty, most delicious food and kept it that way...
Query	food - negative
BERT-Baseline (BB)	for 7 years they have put out the most tasty food and kept it that way.
BB + MLM pretraining (BB-MLM)	for 7 years they have put out the most greasy food and bland food that way...
BB-MLM + one-zero polarity injection	for 6 years they have put out the most bland food and kept it that way...
BB-MLM + saliency-based polarity injection	for 7 years they have put out the most bland food and kept it that way...

Table 5: Example outputs from the SemEval data showing aspect-level sentiment transfer from the ablated models. Aspects colored red (negative) or green (positive) indicate their sentiment.

Input	i must say i am surprised by the bad reviews of the restaurant earlier in the year , though .
Query	restaurant - negative
Output	i must say i am surprised by the bad reviews of the restaurant earlier in the year , though .
Comment	No change. Sentiment here is implied and latent.
Input	the space is limited so be prepared to wait up to 45 minutes - 1 hour , but be richly rewarded when you savor the delicious indo-chinese food
Query	space - positive, indo-chinese food - positive
Output	the space is extensive so be prepared to 10 - 15 - 20 + minutes , but delicious chinese food .
Comment	Disfluency and dropped content due to the length of input and the negative sentiment implied through the word 'waiting'.
Input	i'd be horrified if my staff were turning away customers so early and so rudely!
Query	staff - positive
Output	i'd be delighted if my staff were turning away customers so early and nicely!
Comment	Lower naturalness of the output from real-world knowledge that turning away customers is bad.
Input	i had fish and my husband had the filet - both of which exceeded our expectations .
Query	fish - negative, filet - positive
Output	i had fish and my husband had the filet - both of which exceeded our expectations .
Comment	The attribute markers for fish and filet are shared, making transfer difficult. Significant rewriting of the input in needed to produce acceptable fluent output.

Table 6: Example sentences that show difficulty in transferring sentiment.

To understand the errors in outputs better, the outputs marked with low Att, Con and Gra scores were examined. Some of these outputs are shown and discussed in Table 6. Many of the failures in Attribute match were found to be due to the complexity involved in the language, such as when the sentiment expressed towards a target is implicit from the content of the review. The absence of attribute markers also makes it harder to convert sentiment. Most outputs with low Con and Gra scores were found to contain very long sentences. The models were trained on the Yelp dataset which mostly contained smaller sentences. Failed examples with multiple different polarities at the output were often also due to the attribute markers towards different aspects being shared in the input sentence.

Such examples require significant rewriting and reordering to produce sentences of acceptable fluency, and our method seems most successful when making localized changes such as with word replacements.

6 Conclusion and Future Work

In this paper, the task of aspect-level sentiment style transfer has been introduced, where stylistic attributes can be localized to different parts of a sentence. We have proposed a BERT-based encoder-decoder architecture with saliency-based polarity injection and show that it can be successful at the task when trained in an unsupervised setting. The experiments have been conducted on an aspect level polarity tagged benchmark dataset related to the restaurant domain. This work is hopefully an important initial step in developing a fine-grained controllable style transfer system. In the future, we would like to explore the ability to transfer such systems to data-sparse domains, and explore injecting attributes such as emotions to targets attributes in larger pieces of text.

Acknowledgments

Authors duly acknowledge the support from the Project titled Sevak-An Intelligent Indian Language Chatbot, Sponsored by SERB, Govt. of India (IMP/2018/002072).

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

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.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709.

Diederik P Kingma and J Adam Ba. 2019. A method for stochastic optimization. arxiv 2014. *arXiv preprint arXiv:1412.6980*, 434.

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 3570–3575.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018a. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018b. Delete, retrieve, generate: A simple approach to sentiment and style transfer. *arXiv preprint arXiv:1804.06437*.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

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

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 480–489.

EFTK SANG. 1999. Representing text chunks. In *Proceedings of EACL'99*, pages 173–179.

Erik Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

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.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*. Iclr.

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

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis. In *Thirty-second AAAI conference on artificial intelligence*.

Youzhi Tian, Zhiting Hu, and Zhou Yu. 2018. Structured content preservation for unsupervised text style transfer. *arXiv preprint arXiv:1810.06526*.

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

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. " mask and infill": Applying masked language model to sentiment transfer. *arXiv preprint arXiv:1908.08039*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232*.

Jingjing Xu, Xu Sun, Qi Zeng, Xuancheng Ren, Xiaodong Zhang, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. *arXiv preprint arXiv:1805.05181*.

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.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*.