Utrecht University

Assessing Sentiment Transfer Models on "Real" Text Style Transfer Tasks

Student: ZiYuan Wang,
Supervisor: Guanyi Chen, Marijn Schraagen,
Student Number: 9380744,
Program: Master Computing Science, Utrecht University,
Date: 29/8/2023

Contents

Abstract

With the development of Artificial Intelligence, style transformation has made progress in computer vision and natural language processing. Style transfer originated in the field of vision and has been extended to the field of text, defined as preserving the content while changing the style of the text, for example, attributes such as politeness, formality, and humor. However, for one of the popular sentiment style transfer tasks, we argue that it should not belong to style transfer. We selected five representative datasets and models, and evaluated them in real style transfer tasks, revealing that sentiment transfer may not be a real style transfer.

1 Introduction

Artificial intelligence is becoming more and more important nowadays [Vinuesa et al.(2020)]. The escalating importance of artificial intelligence is due to its transformative impact on various fields. In recent years, AI has grown exponentially, revolutionizing industries such as healthcare, finance, and communications. This thesis delves into the realm of text style transfer, which is a key aspect of artificial intelligence. In AI, the study of style transfer has made lots of progress in many areas, including computer vision and natural language processing. Style transfer is a technique designed to convert the style of a given piece of content (e.g., image or text) to another style while keeping its basic information intact. It is interesting that the concept of style transfer was first introduced in the field of computer vision, where it has been widely applied to tasks such as artistic style transfer and image-to-image translation. Style transfer in computer vision is a technique that allows the users to recompose the content of an image in the style of another. The success of these methods inspired researchers to explore the possibility of applying similar techniques to natural language processing. So what is text style transfer? Text style transfer aims to change some attributes in the given text, such as politeness, formality, humor, and others while not changing the original content. Text style transfer is a rapidly growing field of research that aims to automatically modify the style of a given text while preserving its content [Liu et al.(2020)]. This task has broad applications, ranging from adding politeness to sentences [Madaan et al.(2020)] to removing toxic sentences in social media posts [Laugier et al.(2021)].

However, in the field of style transfer, there are issues with the definition of “style" in many studies. While sentiment transfer has been widely studied in the literature [Madaan et al.(2020)], it is more about content than style. For example, simply substituting the word "happy" with "sad" can effectively shift the sentiment of a sentence, but it doesn’t necessarily alter its style. Hence, it’s crucial for style transformations to maintain the original meaning of the sentence while changing its style.

The style of a sentence should be a characteristic of another domain, such as formality [Wegmann and Nguyen(2021)], politeness, etc.

Because of the issue in sentiment transfer, there is a need to re-examine the models and algorithms in the sentiment transfer area. Therefore, in this thesis, we will choose five models, with good performance in sentiment transfer, to evaluate the performance of these models or algorithms on real style transfer tasks to see if they still work effectively. To demonstrate the real style transfer tasks, we choose five datasets, containing different areas of text style transfer. The specifics of the chosen model as well as the dataset will be developed in detail below. By conducting these experiments, we hope to demonstrate that a portion of the research on sentiment transfer is not true style transfer.

Overall, we believe our study can shed light on the current state-of-the-art in text style transfer and provide valuable insights for future research. By evaluating the models on real style transfer tasks, we hope to contribute to the development of more robust and accurate models for text style transfer. Additionally, our study can also serve as a benchmark for researchers and practitioners in this field, providing them with a better understanding of the capabilities and limitations of different approaches for style transfer.

This thesis is structured as follows: Chapter 2 discusses related work, e.g., section 2.1 introduces natural language processing and natural language generation, and section 2.2 discusses the concept of style. Chapter 3 discusses the research questions of this thesis. Chapter 4 discusses the dataset chosen for this thesis and the ideas behind its selection. Chapter 5 discusses the models selected for this thesis and the principles behind their representation. Chapter 6 lists the experimental protocol of this thesis. Chapter 7 lists the experimental results of this thesis and their analysis. The last chapter summarizes the findings of this thesis and the outlook for the future.

2 Literature Review

2.1 Background

2.1.1 What is NLP

NLP [O'Connor and McDermott(2001)] is currently a popular direction in the field of artificial intelligence research, which aims to enable computer p rograms to understand, accept and process human language. Human language is an abstract symbolic representation, that contains rich semantic information, and humans can easily understand the meaning of it. Computers, on the other hand, can only process numerical information on and cannot understand human language directly, so they need to convert human language numerically. Not only that, human communication is contextual, which is also a huge challenge for computers. Before the advent of artificial intelligence, it was almost impossible to write language models using traditional programming methods, but over time, we have many algorithms that can understand human language.

The main direction of NLP research is to enable computers to understand the meaning of human language, and to achieve this, models need to be trained to have the ability to analyze syntax, and to slice and dice utterances. This involves tasks such as part-of-speech tagging [Voutilainen(2003)], syntactic parsing [Van Gompel and Pickering(2007)], and semantic analysis [Goddard(2011)], which enable computers to identify the relationships between words and phrases in a sentence and to infer their meaning. NLP techniques can also be used to perform sentiment analysis [Medhat et al.(2014)], which involves identifying the emotional tone of a piece of text, and text classification, which involves assigning a document to one or more predefined categories based on its content.

In summary, NLP is a rapidly growing area of research and development that has led to significant advances in our ability to process, understand, and generate human language. With the growing corpus for training and the development of new machine-learning techniques, the potential applications of NLP are vast and diverse.

2.1.2 Tasks of NLP

Tasks in NLP can be simply classified into these categories: Sequence labeling, Classification tasks, Sentence relationship judgment, Generative tasks. Sequence labelling is a form of natural language processing task used to annotate each token in a text sequence, including word segmentation [Saffran et al.(1996)], POS tagging [Straka and Straková(2017)]. The objective of word segmentation is to divide the text into individual words or phrases, POS tagging is to assign grammatical categories to each word, named entity recognition is to identify entities with specific meanings in the text, such as person names, organization names, and location names, and semantic labelling is to identify sentiments, opinions, attitudes, and other aspects in the text.

Classification tasks refer to the classification or labelling of text into predefined categories or tags. Text classification involves categorizing text into multiple pre-defined categories, such as news, sports, entertainment, etc., whereas sentiment analysis is used to determine the polarity of the text's sentiment, such as positive, negative, or impartial.

Sentence relationship judgement is a form of natural language processing task designed to identify the relationship between two given sentences. Natural language inference tasks aim to determine the logical relationship between two sentences, such as entailment, contradiction, or irrelevance. QA tasks aim to answer factual questions relating to the input text.

Generative tasks, such as machine translation [Koehn(2009)] and text summarization, generate new text related to the input text. Text machine translation is the process of translating text from one language to another, whereas text summarization is the process of summarising and simplifying the given text for improved comprehension.

2.1.3 What is NLG

The goal of Natural Language Generation (NLG), a branch of natural language processing (NLP), is to automatically produce language that is similar to that of humans [McDonald(2010), Chen(2022)]. NLG aims to make it possible for computers to produce text output that is coherent, meaningful, and culturally appropriate. NLG can be applied in a variety of contexts, including chatbots, automated journalism, product descriptions on e-commerce sites, and business intelligence reporting. Virtual assistants called chatbots can have natural-language conversations with people. Chatbots can answer to user inquiries in real-time using NLG technology by providing pertinent and contextually appropriate responses [Chen et al.(2020), Zeng et al.(2021), Zheng et al.(2022)]. Automated journalism is the process of creating news items from structured data, such as stock market or sports scores, automatically. NLG can be used in e-commerce to automatically create product descriptions, helping businesses save time and resources. NLG is used in business intelligence reporting to automatically provide summaries and insights from massive amounts of data. NLG can assist in conveying findings in a clear and useful way.

NLG can be performed using various techniques, including rule-based approaches [Goodman et al.(19
Chen et al.(2019), Chen and van Deemter(2020), Chen and van Deemter(2023)], template-based approaches [McRoy et al.(2003)], and machine learning-based approaches [Khan et al.(2016), Chen et al.(2018), Chen et al.(2023), Same et al.(2022)]. In rule-based approaches, NLG systems generate text using rules that have been manually crafted. The principles specify how the output should be generated given particular input conditions. Using input data such as temperature, humidity, and precipitation, a rule-based approach might generate a weather report using a set of predefined principles.

Template-based approaches employ pre-defined templates with data-fillable place-

holders. The templates are intended to encapsulate the structure of the text that must be produced. The NLG system replaces the placeholders with the pertinent data to generate the final output. When the output's structure is fixed, but its content must be generated dynamically based on input data, template-based approaches are frequently employed. For example, a template-based strategy could be utilised to generate personalized emails based on user preferences.

Machine translation is an important part of NLP, which focuses on automatically translating text [Bar-Hillel(1960)] from one language to another. It can be mainly divided into three steps: "pre-processing, translation model, and post-processing". Pre-processing is the process of normalizing the sentences in the source language, dividing the excessively long sentences into several short sentences by punctuation, filtering some inflections and words that are irrelevant to the meaning, grouping some numbers and expressions that are not standardized into sentences that conform to the specification, and so on. Translation module is the process of translating the input character units and sequences into the target language sequences, which is the most critical and core area in machine translation. Throughout the history of machine translation development, the translation module can be divided into three categories: rule-based translation, statistical-based translation and neural network-based translation. Nowadays, neural network-based machine translation has become the mainstream method, and the effect is far more than the first two types of methods. The post-processing module is to convert the translation results into case, modeling units for splicing, and special symbols for processing, so that the translation results are more in line with people's reading habits. Machine translation techniques are widely used in style transfer, for example, back-translation techniques utilize machine translation.

Approaches based on machine learning generate text output using algorithms that learn from vast quantities of data. These approaches are frequently more flexible than rule-based and template-based approaches because they can recognise patterns and structures in the data that are difficult to specify explicitly. Machine learning-based NLG systems can be trained on large text datasets and can learn to generate text that sounds natural and resembles human-generated text. In applications such as chatbots, where the system must generate responses to user queries that sound natural, machine-learning-based approaches are frequently employed.

The approach chosen is dependent on the specific use case and available resources. Rule-based and template-based approaches may be easier to implement, but less flexible than machine learning-based approaches. Approaches based on machine learning necessitate vast quantities of training data and computational resources but generate text with a more natural tone. The discipline of NLG is advancing rapidly, and researchers are constantly investigating new techniques to enhance the quality and adaptability of NLG systems.

2.2 What is Style

Style can be defined as a particular way of doing something, which includes the way something is presented, expressed, or performed. Style [Dictionary(2002)] is an artistic concept, a representative look of an artwork in its entirety. Style is different from the general artistic characteristics, which are relatively stable and inherent through the artwork, reflecting the intrinsic characteristics of the time, nation, or artist's thought and aesthetics. The essence is the artist's unique and distinctive expression of aesthetics, with infinite richness. Artists are formed by different life experiences, artistic qualities, emotional tendencies, and aesthetic differences, and are influenced by the historical conditions of the times, society, and nationality. Subject matter and genre, art disciplines also have a restraining effect on the style of works

In writing, style refers to the way a writer uses language to express their ideas, thoughts, or feelings [Arsyad and Adila(2018)]. It includes elements such as word choice, sentence structure, tone, and voice. Different styles of writing are used for different purposes, such as persuasive writing, informative writing, and creative writing. Each style has its own characteristics and conventions, which writers can use to achieve their intended effects. For example, persuasive writing often uses emotional appeals and persuasive language [Walton(2005)] to convince readers of a particular point of view, while informative writing [Kroll(1986)] uses clear and concise language to provide information to readers.

In art, styles are characterized by diversity and homogeneity. The infinitely rich diversity of the real world itself, the different creative personalities of artists, and the diversity of aesthetic needs of art appreciators determine the diversity of art styles. An artist's style [Van Noord et al.(2015)] can be influenced by various factors such as cultural, social, and historical contexts, personal experiences, and artistic movements. Even the works of the same artist do not exclude the possibility of having various styles. It is the diversity of artistic styles that greatly contributes to the prosperity and development of art. On the other hand, the diverse styles of the same artist present a dominant stylistic characteristic on the whole due to the constraints of their creative personalities; the stylistic distinction between different artists cannot but be constrained by the aesthetic needs and artistic development of a certain era, nation and class in which they live together, thus showing the consistency of styles. The diversity and consistency of styles are interrelated and permeable, presenting an intricate phenomenon that should be distinguished when making art criticism. In the plastic arts, stylistic diversity and homogeneity often have very distinct expressions. For example, the outstanding creations of Italian Renaissance art are different from Michelangelo's majestic, Da Vinci's deep and Raphael's elegant, while Romanesque, Gothic, Renaissance and Baroque are typical styles of their respective eras. Of course, an era also has an era of art style, which is due to the formation of people in a period of time have a relatively

close aesthetic tendency. For example, the Han Dynasty mostly advocated a simple and muddy art style, and the 18th century France was popular with the highly decorative Rococo style, etc

In fashion, style refers to a person's distinctive manner of dressing, grooming, and accessorizing. Different fashion styles can be associated with different subcultures, such as punk, goth, hip-hop, and preppy. Moreover, fashion styles can change over time, reflecting cultural shifts and historical events.

Musical style [Crocker(1986)] refers to the representative and unique look of a musical work as a whole. It is the type of music. Song style is similar to other artistic styles in that it is relatively stable, internal and profound, and can reflect more essentially the external mark of the inner characteristics of the era, nation or musician's personal ideology, aesthetic ideals and spiritual temperament through songs. The formation of the song style is a sign that the era, nation or musician has transcended the infantile stage in the understanding and realization of music and has freed itself from various patterned constraints, thus tending to or reaching maturity.

Overall, style has a vital role in many aspects of human expression and communication. It can be used to convey messages, create a feeling of identity, or elicit specific emotional and behavioural responses from audiences. Understanding the intricacies and conventions of diverse writing styles is critical for constructing accurate and effective models and algorithms for style transfer tasks in the field of natural language processing and style transfer.

2.3 Style Transfer in Computer Vision

Style transfer entails acquiring a mapping function from a source domain to a target domain. Typically, the mapping function is learnt using techniques such as deep neural networks, which can learn complex mappings between two domains. Image processing, natural language processing, and music are just some of the disciplines in which these techniques have been implemented. In image processing, the source domain is typically an input image, whereas the target domain is the intended image style. For example, applying the style of a painting or drawing to a content image transforms the input image into a painting or drawing. The source domain in natural language processing is the original text, while the target domain is the desired style of the output text.

Style transfer in the image is a technique of machine learning that involves transferring the artistic style from one image to another [Gatys et al.(2016)]. It works by learning the statistical properties of the style image and applying them to the content image. By separating the style and content information of an image, it becomes possible to apply the style of one image to the content of another. In recent years, style transfer has become increasingly popular in the field of computer vision. It is a pow-

erful technique that can be used to create artistic effects on images, such as turning a photograph into a painting or sketch, by transferring the style of one image onto another [Liu et al.(2017)]. This is typically achieved using deep neural networks, which are trained on large datasets of images to learn how to extract and manipulate features in the images. The resulting models can then be used to perform style transfer on new images.

One of the most common applications of style transfer in computer vision is style generate images. This involves applying the style of a painting or other artwork to a photograph or other image, resulting in a new image that combines the content of the original image with the style of the artwork. The resulting images can be highly expressive and visually striking and have been used in a variety of contexts, such as advertising, graphic design, and digital art.

Another important application of style transfer in computer vision is image-to-image translation. This involves using a style image to translate a content image into a different domain, such as turning a daytime image into a nighttime image [Meng et al.(2019)]. This technique has been used in a variety of contexts, such as video editing, image editing, and virtual reality, and has the potential to enable new forms of creative expression and artistic production.

Style transfer entails acquiring a mapping function from a source domain (i.e., the initial input) to a target domain (i.e., the desired output). This mapping can be learned using a variety of techniques, including deep neural networks, which are capable of learning complex mappings between two domains. In the case of image style transfer, the source domain is the input image, whereas the target domain is the intended style of the output image. Similarly, in natural language processing, the source domain is the original text and the target domain is the desired text style.

2.4 Text Style

Text style transfer is a fast expanding field of study, and several works have been done that cover various facets of this endeavour. We conducted a thorough literature review of recent research in order to give a comprehensive overview of the state of the art in this area. Our review covers a broad range of topics related to text style transfer, including different types of styles, approaches for style transfer, several datasets, and evaluation methods. We also highlight some of the challenges that remain in this field and opportunities for future research.

2.4.1 Corpora

Parallel Corpora Two basic methods to specify text style transfer learning are parallel corpora and non-parallel corpora. In parallel corpora, pairs of texts are available

in both the source and target styles. This means that the same content is expressed in two different styles, such as formal and casual, and the model can learn to transfer the style of one text to the other. Parallel corpora are often used in supervised learning, as the desired output style is known in advance. One key area of focus is formality transfer, which aims to modify the formality of a text, such as making a casual text more formal or vice versa. In [Rao and Tetreault(2018)], [Wu et al.(2020)] and [Ma et al.(2021)], the authors both used the GYAFC dataset provided by YAHOO. Datasets such as GYAFC and Yelp datasets [Ma et al.(2021)] have been used to evaluate the quality of formality transfer, and metrics such as accuracy and F1 score have been used to assess the performance of different models. Although they are dealing with the same dataset, the ideas used and the way they are handled is different from one author to another. For example, some researchers use conventional ideas such as Pivot-Based, teacher-student, back translation [Wu et al.(2020)], reinforcement learning algorithms [Upadhyay et al.(2022)] or large-scale pre-training models such as GPT2, BERT. [Liu et al.(2020)]

In another area of research related to text style transfer, some researchers have explored the task of politeness transfer. Politeness is an important aspect of language that can impact how a message is received by its intended audience. Some studies have focused on developing models that can automatically generate more polite language [Madaan et al.(2020)], while others have aimed to remove the toxic meaning of words and phrases [Laugier et al.(2021)] [Atwell et al.(2022)]. These approaches have been applied in various contexts, such as social media platforms and online forums, where the tone of messages can often be confrontational and aggressive. By improving the politeness of language in these contexts, it may be possible to reduce hostility and improve online communication. While still in its early stages, politeness transfer research shows promise for creating more positive and respectful online environments. Several works have proposed methods for politeness transfer, including template-based approaches, rule-based models, and neural network-based models.

Non-Parallel Corpora In contrast, non-parallel corpora contain texts in only one style and are often used in unsupervised learning. The model learns to capture style information from the input text and generate text in the desired style without explicit style labels. Non-parallel corpora are easier to obtain and can be useful when labeled data is scarce. However, since the model is not explicitly trained in the desired style, it may not always produce text that is stylistically appropriate or grammatically correct. Since there are more non-parallel databases, there are more types of style transfers that can be involved. In [Toshevska and Gievska(2021)], The author conducted a study of Shakespeare's stylistic transformations. In this essay [Khalid and Srinivasan(2022)], the author explores sensory style in the literary genre. In addition, in the area of positive

and negative style transformation, the article also conducted research on it [Li et al.(2020), Hu and He(2021)].

2.4.2 Approaches

Supervised Learning Supervised learning is a machine learning technique that involves training a model on a labeled dataset to learn the mapping between the input and output variables.

There are many supervised learning methods, of which the sequence-to-sequence approach is a representative one. In Seq2Seq, two recurrent neural networks (RNNs) are trained together: an encoder and a decoder.

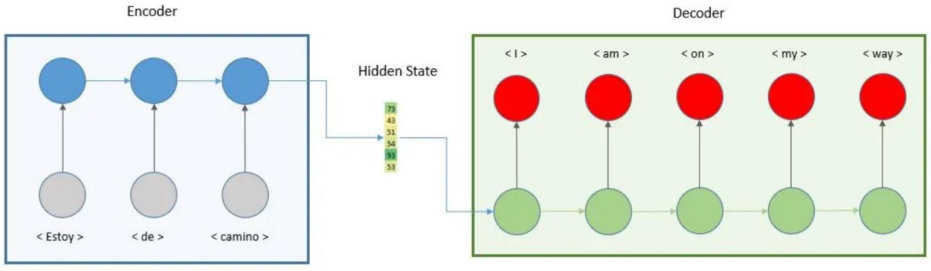

Figure 1: Encoder and decoder [BM([n. d.])]

The encoder receives a sequence of input data, such as sentences in a language, and processes them into a fixed-length vector representation. Then, the decoder receives the context vector and generates a string of output data. Jhamtani et al. [Jhamtani et al.(2017)] used this method to solve a style transfer problem. They present the problem of converting modern English into Shakespearean style. Due to the small parallel dataset they used, the previous phrase translation method [Xu et al.(2012)] did not perform well. The authors therefore propose a sequence-to-sequence neural model with a pointer network component that copies words directly from the input and demonstrate that this approach is much better than previous phrase translations. Formality transfer has also evolved considerably in supervised learning. Wang et al. [Wang et al.(2020)] noticed the problem of regular Sequence to Sequence when studying formality transfer. The authors argue that since the datasets of parallel data are generally relatively small, e.g., the GYAFC dataset [Rao and Tetreault(2018)] has only 50k data per domain, and the sentences after style transfer are very close to the original utterances. as shown in Fig below. Therefore, direct application of sequence-to-sequence (Seq2Seq) model not only risks overfitting the training set, but also may not take full advantage of the nature of informal and formal sentences. The authors

propose a sequence-to-sequence model that shares the latent space. After experiments, the authors demonstrate that the model effectively improves the quality of the Seq2Seq model's performance in formality transfer area.

Informal sentence:	**Formal sentence:**
I do not know are u ready for one ?	I do not know. Are you ready for one?
Sounds like a rhetorical question :)	It sounds like a rhetorical question.
what r ya talking abt	What are you talking about ?

Figure 2: Examples of formality style transfer [Wang et al.(2020)]

NLP techniques have also evolved significantly in recent years in pre-training large models. GPT is a family of large-scale language models developed by OpenAI that uses unsupervised learning techniques to pre-train large amounts of text data. Wang et al. [Wang et al.(2019)] also investigates the application of the GYAFC dataset[Rao and Tetreault(2018)] in the direction of formality transfer. Unlike they did in 2020 [Wang et al.(2020)], they try to introduce a large-scale pre-trained model into it. The authors argue that the introduction of GPT models is effective because the GYAFC dataset is small. The authors feed strings of raw informal sentences and pre-processed sentences into the encoder, and use two encoders to encode raw informal text and rule pre-processed text separately, and then develop a hierarchical attention-based model to aggregate the information.

Unsupervised Learning Unsupervised learning approaches in text style transfer are gaining popularity in recent years due to the lack of parallel data in many style transfer tasks. Parallel data is scarce and expensive for many styles of text such as sentiment. For unsupervised learning, it can leverage large-scale unlabeled data and pre-trained models to learn robust representations of text and can generate diverse and natural texts in the target style by using techniques such as word replacement, back translation, pseudo-parallel data, or GAN network.

There are several typical methods using unsupervised learning. Word replacement is one of the typical methods. Word replacement doesn't mean that we can only replace words in a sentence. On the other hand, deleting, replacing, or adding phrases to a sentence are all possible ways to change the style of a sentence. Li et al. [Li et al.(2018)] proposed a simple example to change the style. In studying sentiment transfer, the authors found that style transfer can often be accomplished by changing several attribute tokens - words or phrases in a sentence that denote a particular attribute - while leaving the rest of the sentence largely unchanged. The authors identify attribute tokens by looking for phrases in the corpus that occur more frequently in sentences with one attribute than another, remove the negative tokens from them, and then search for

similar sentences from the positive corpus to pair and train. The authors argue that such an approach is more efficient than simply using sentences in generative adversarial networks [Zhao et al.(2018)].

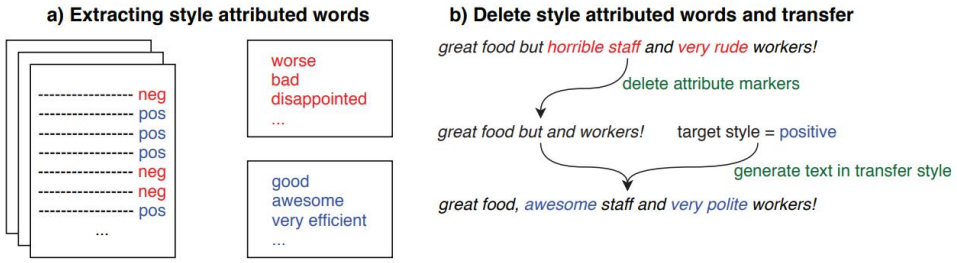

Figure 3: Word delete method [Li et al.(2018)]

Back-translation is another mainstream unsupervised style transfer technique. The specific process can be simply understood as, given a source text of a specific style, first use a model to transform the text into the target style. Then, the translated text is back-translated into the original style using the same model. The resulting text is then compared with the original source text, and if the two texts are similar enough, they can be added to the training data of the style conversion model. The quality of the data can be effectively improved by this step. Prabhumoye et al. [Prabhumoye et al.(2018)] argues that relying on substitution, deletion and insertion of [Li et al.(2018)] for style transfer, while proving that gas retains the meaning of the sentence well, may not be correct for style transfer, and the variety of styles it can transfer is limited. The authors therefore hypothesize - relying on Rabinovich et al. [Rabinovich et al.(2016)] who show that authorship features are confounded by manual and automatic machine translation - that grounding in back-translation is a plausible way to reword sentences while reducing their stylistic features. The authors propose the use of back-translation to rewrite sentences and reduce the effect of the original style, based on which the model is trained.

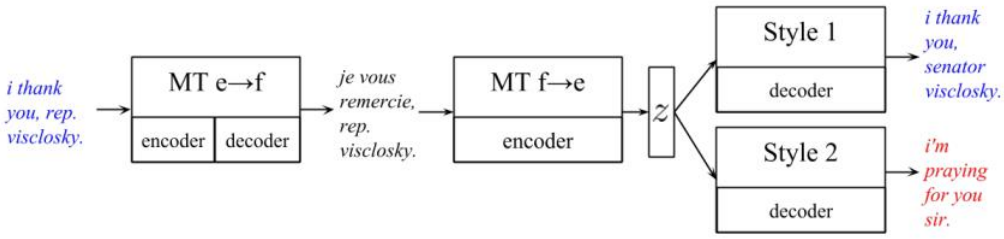

Figure 4: Back translation [Prabhumoye et al.(2018)]

GAN-based models, which stand for Generative Adversarial Networks, are a type

of unsupervised learning approach that have been applied to text style transfer. It consists of two networks: the generative network G, which is used to fit the data distribution, and the discriminative network D, which is used to determine whether the input is "real" or not. In the training process, the generative network G "tricks" D by accepting a random noise to mimic the s in the training set as much as possible, while D discriminates the real data and the output of the generative network as much as possible, thus forming a game process between the two networks. Ideally, the game results in a generative model that can be "faked".

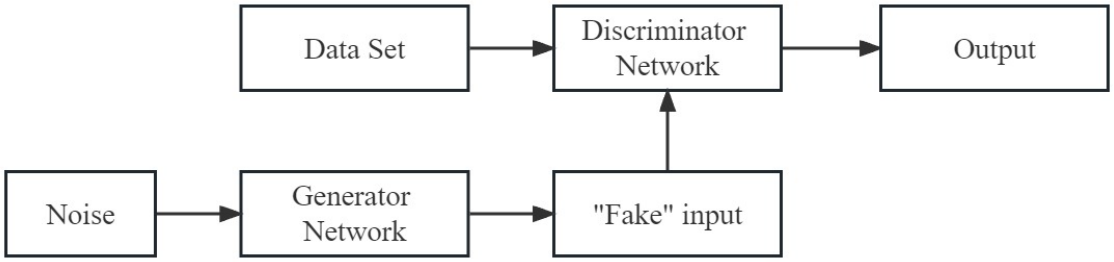

Figure 5: GAN network

In text style transfer, The generator network learns to generate text that is indistinguishable from the target style, while the discriminator network tries to distinguish between the generated text and text from the target style. Previous methods of non-parallel text style transfer using adversarial training often suffer from content leaking, which causes unintended changes in the content during the style transfer process. Lai et al. [Lai et al.(2019)] proposed a new adversarial training model with a word-level conditional architecture and a two-phase training procedure. This work successfully solved the leaking problem and gained good performance on style transfer tasks.

Pseudo-parallel data generation models are a type of unsupervised learning approach for text style transfer that aims to generate a large amount of pseudo-parallel data from monolingual corpora without explicit style labels. Previous methods suffered from the limitation of preserving the intended or affective meaning of a sentence while transferring its style. Jin et al. [Jin et al.(2019)] argue that a style transfer system must be able to (1) produce sentences that match the target attributes, (2) preserve the content of the source sentences, and (3) produce fluent language. In contrast, in the non-parallel case, most existing approaches aim to separate content and attributes in the underlying sentence representation. For example, GAN networks, Shen et al. [Shen et al.(2017)] use generative adversarial networks (GANs) to learn this separation, and while these models can effectively generate sentences containing the target style, their high probability of losing and corrupting the original meaning. Therefore, the authors propose

a framework to construct a pseudo-parallel corpus by matching semantically similar sentence subsets in the source and target corpora. The pseudo-parallel corpus effectively alleviates the problem of insufficient data volume in the database used by the authors, and at the same time enables the authors to test and compare the models using supervised learning, which eventually achieves better results.

2.5 Evaluation Matrix

2.5.1 BLEU

The full name of BLEU is Bilingual Evaluation Understudy [Papineni et al.(2002)]. It is a tool for evaluating the quality of machine translation, and the design idea of BLEU is that the closer the result of machine translation is to the result of professional human translation, the better it is. It is based on the concept of n-grams, which are sequences of n words. The BLEU score computes the precision of n-grams in the machine-generated output compared to the reference translations. Since the purpose of BLEU is to assess the degree of similarity between two sentences, it can also be used to assess the degree of similarity between a sentence and other sentences in the generated set. Taking one sentence as a hypothesis and the other sentences as references, we can compute the BLEU score for each generated sentence and define the average BLEU score as the self BLEU of the document. In addition, for parallel datasets, the ref BLEU can be computed using the parallel data in the dataset. In this thesis, we denote the self BLEU score as BLEU-self, and the reference BLEU score as BLEU-ref.

The rules for calculating the BLEU indicator are as follows. Candidate: The cat sat on the mat. Reference: the cat is on the mat. For example, for BLEU-2, for the 5 words in candidate, {the cat, cat sat, sat on, on the, the mat}, find out if they are in reference, and find that there are 3 words in reference, so the percentage is 0.6.

2.5.2 ROUGE

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores [Lin(2004)] are a set of evaluation metrics commonly used in natural language processing and ROUGE evaluates the degree of overlap of n-grams between machine-generated text and reference text.

The rules for calculating the ROUGE indicator are as follows. Candidate: The cat sat on the mat. Reference: the cat is on the mat. For example, for ROUGE-L, we need to calculate the longest common subsequence (LCS). The LCS is the 5-gram "the cat on the mat". ROUGE-L precision is the ratio of the length of the LCS, over the number of unigrams in a candidate. So it's equal to 5/6. ROUGE-L recall is the ratio of the length of the LCS, over the number of unigrams in reference, so it's equal to 5/6. Therefore, the F1-score, which is the ROUGE-L, is equal to

$$\text{ROUGE} = \frac{2(precision \cdot recall)}{(precision + recall)} = 5/6$$

2.5.3 BERT score

BERT [Zhang et al.(2019)] is a pre-trained language model developed by Google that has achieved top-notch performance in a variety of natural language processing tasks. With the corresponding python library, pre-trained contextual embeddings in BERT can be utilized and words in candidate and reference sentences can be matched by cosine similarity. It has been shown to correlate with human judgments on sentence-level and system-level evaluations. In addition, BERTScore computes precision, recall, and F1 metrics, which are useful for evaluating different language generation tasks.

2.5.4 ACC score

Accuracy (ACC) Calculates the success rate of an attribute change. It measures how much the target attribute affects the generation. We generate sentences of a given style and assign labels to the generated sentences using the same classifier used by the authors of the dataset. The accuracy is calculated as the percentage of predictions that match the style.

3 Research Question

The goal of this study is to assess how well existing models and algorithms for style transfer in natural language processing perform when applied to practical style transfer tasks. It is critical to reevaluate the models and algorithms employed in these studies since many writers in the present style transfer study misunderstand what style is. For instance, despite the fact that many works on sentiment transmission may be more about content than style, many of these works have been categorised as style transfer. We intend to verify the usefulness of these models and algorithms through experiments on actual style transfer tasks and to offer guidance on how to create more precise and reliable models for this purpose. In the end, this research may help us grasp what true style transfer comprises and enhance the performance of style transfer models in real-world settings.

To achieve our goal of evaluating the performance of models or algorithms on real style transfer tasks, our main research question is "Does the models that worked on the sentiment transfer are still working on the real style transfer tasks?" Based on the above question, we propose three sub-research questions.

1. "Find and classify datasets of real style transfer."
2. “Find and classify state-of-the-art models for sentiment transfer."
3. "Test these models on the dataset and compare the test results with previous findings on sentiment transfer."

For the first research question, the motivation is to address the issue of defining what

real style transfer entails and to ensure that future research on this topic is based on accurately classified datasets that represent real style transfer.

For the second research question, the motivation is to identify the current state-of-the-art models for sentiment transfer, which are often mistakenly categorized as style transfer. This will allow us to differentiate between sentiment transfer and style transfer and provide insights into how these models can be improved.

For the third research question, the motivation is to evaluate the effectiveness of these state-of-the-art models on real style transfer datasets and compare the results with previous findings on sentiment transfer. This will allow us to determine if these models can be used for style transfer and identify any potential areas for improvement.

These questions are of paramount importance since evaluating the performance of style transfer models is a challenging task that requires careful consideration of various factors, including the choice of evaluation metrics, the selection of appropriate datasets, and the establishment of a reliable benchmark for comparison. By addressing this research question, we aim to contribute to the development of more effective and accurate models for style transfer, which can be used in a variety of practical applications, such as machine translation, text summarization, and content generation.

4 Datasets

We will first look for datasets related to true style transformations in natural language processing. To handle our experiments, we chose five datasets as samples. Which are listed in Table 1.

Origin Papers	Description	Examples	Paralleled
[Rao and Tetreault(2018)]	Paralleled sentences for formality transfer.	I like Rhythm and Blue music. My fav genre of music is R&B.	True
[Atwell et al.(2022)]	Includes over 2,000 offensive comments, as well as hand-rewritten parallel non-offensive comments	Go to therapy – Therapy may help you	True
[Laugier et al.(2021)]	Offensive sentences of 2 Gb with tags	I'm a white woman in my late 60's and believe me, they are not too crazy about me either!! Toxicity Labels: All 0.0	False
[Wu et al.(2020)]	Consists of informal sentences as well as their corresponding formal sentences. Conversations originating from Twitter are composed of two informal sentences, and a manually written formal statement summarizing them.	rubs my beard on your face as we hug - That guy has a lot on his mind.	True
[Madaan et al.(2020)]	Manually labeled and scored for Politeness. The score is a number from 0-1, with larger numbers indicating more politeness.	subject: clickpaper approvals 09/27/00 score:0.2644	False

Table 1: Datasets

Next we'll go into a little more detail about each dataset and how to preprocess the dataset. Also to show why we select these datasets.

4.1 GYAFC Formality transfer dataset

GYAFC is a Formality transfer dataset with informal and formal sentence pairs created by the author [Rao and Tetreault(2018)] from the Yahoo L6 corpus [Napoles et al.(2017)]. After the corpus author removed interrogative sentences, sentences containing URLs, and sentences that were too short or too long, 3000s informal sentences were extracted

as the dev set and 1500s sentences as the test set. The goal of the research in this paper is to evaluate the performance of the model in a real-world style switching task, and the GYAFC dataset will be a suitable choice. This is because Formality transfer is a popular field and the GYAFC dataset is representative of it. The GYAFC dataset has been widely used in style transfer studies and has been recognized and proven valid by a number of previous studies. The GYAFC dataset was used for formality style transfer in papers such as [Luo et al.(2019)], [Zhang et al.(2018)] and [Li et al.(2018)]

The author randomly sample a subset of 53,000 informal sentences each from the Entertainment &Music (E&M) and Family & Relationships (F&R)categories. The dataset details can be seen in Table 2. Since the authors have done detailed data processing, we decided to use the GYAFC dataset directly for Formality transfer training. In addition, the authors provide manual reference sentences for Dev set and Test set, so model performance testing can be easily performed.

However, as a parallel dataset, it possesses the common disadvantage of having a relatively small amount of data and containing a relatively small variety of formality.

		Informal to Formal		Formal to Informal	
	Train	Dev	Test	Dev	Test
E&M	52595	2877	1416	2356	1082
F&R	51967	2788	1332	2247	1019

Table 2: GYAFC dataset statistics

4.2 Offensive Dataset

The datasets for Offensive transfer are from the paper[Atwell et al.(2022)]. The authors gathered data by creating a pipeline that collected and organized a series of offensive comments on Reddit. They then annotated the comments to minimize the offensiveness in the text while retaining the meaning of the original text. The annotation process was done by sociolinguistic experts. The resulting dataset is the first publicly available Offensive parallel corpus.

The dataset authors made a distinction between Global or Local modifications when labeling sentences. When the sentences were successful in removing offensiveness by modifying only a few words, the original structure of the sentence was left as unaltered as possible in order to retain more of the meaning of the sentence itself. This can be observed in table 3.

The dataset provided by the authors has been processed and is ready to be used directly for training, so for Offensive transfer training, this dataset will be used directly without additional data preprocessing. See the table below for detailed information.

Original Comment	Rewritten Comment	Global/ Local	Reason for paraphrasing
You can't do s*** because you're an idiot.	You can't do anything because you're not competent.	Local	Cursing, Insults
So you s**k as person. Got it	So you're not a great person. Got it	Local	Cursing, Insults
What backward b*****k nowhere country do you live in?	What country do you live in?	Local	Xenophobia, Cursing
Keep my phone gallery secrets out your f***** MOUTH	Don't talk about my phone gallery secrets	Global	Cursing, Rudeness
F*** off. Sick of people like you thinking everything is propaganda	Please go away. Tired of people like you thinking everything has a hidden plan	Global	Cursing, Rudeness

Table 3: Examples of applying local and global changes to the comments for different types of offensive speech [Atwell et al.(2022)]

Datasets Attribute	sentence pairs
Train	1,585
Dev	200
Test	199

Table 4: Offensive dataset statistics

4.3 Toxic Transfer Datasets

For the part of the statement Toxic transfer, we chose the dataset from this article [Laugier et al.(2021)]. The dataset used in this thesis is the Civil Comments dataset, which is the largest publicly available toxicity detection dataset to date. It consists of 2 million comments from an independent news site comment plugin that was created between 2015 and 2017 and appears on approximately 50 English-language news sites worldwide. Each comment was annotated for toxicity and toxicity subtypes by a population rater. Comments were grouped into sentences using NLTK's sentence tagger, and a pre-trained BERT toxicity classifier was fine-tuned in the dataset. The statistics for the dataset can be seen in the following table.

Datasets Attribute	Toxic	Civil
Train	90,293	5,653,785
Dev	4,825	308,130
Test	4,878	305,267

Table 5: Toxic transfer dataset statistics

As this dataset is too large, considering the training speed and the feasibility of the training platform, this thesis excludes sentences with more 20 words.

4.4 Twitter Formal datasets

TCFC is a Formality dataset based on GYAFC developed [Wu et al.(2020)]. Unlike the GYAFC dataset, the authors trained the formality classifier on manually labeled 50k text pairs, treating formal sentences as positive examples and informal sentences as negative examples. And use this classifier to construct the dataset by crawling message response pairs from Twitter. Thus the training samples of this dataset are much larger than the GYAFC dataset, with the disadvantage that the training part is not a parallel dataset. For testing purposes, the authors also asked a linguist to manually rewrite some of the statements to create a test set. See the table below for detailed information.

Datasets Attribute	sentence pairs
Train	1,727,251
Dev	980
Test	978

Table 6: TCFC dataset statistics

4.5 Politeness Transfer datasets

The politeness transfer datasets we chose are from this paper [Madaan et al.(2020)]. This dataset consists of a large number of email conversations exchanged by employees of Enron Corporation. The authors assigned politeness scores to the sentences in this corpus using a politeness classifier after pruning them. The authors considered sentences with politeness scores greater than 90% as polite sentences and used them for Politeness transfer training. The authors manually annotated eight hundred sentences for testing purposes. See the table below for detailed information.

Datasets Attribute	sentence pairs
Train	1,048,576
Dev	1,047,798
Test	1,048,510

Table 7: Politeness dataset statistics

4.6 Reference sets of the Datasets

For the datasets GYAFC Dataset, Offensive Dataset, and Twitter Formality Dataset described in Sections 4.1, 4.2, and 4.4 of this paper, the same datasets will be used for the reference metrics tests in the following sections, since they are parallel datasets and the corresponding statements were manually rewritten by linguists as reference.

For the dataset described in section 4.3, since the dataset is non-parallel and the authors did not provide a suitable set of reference sentences, in the subsequent calculation of the metrics, the SELF metrics will be the main metrics to be computed, and for the reference metrics we will utilize the Offensive described in section 4.2 of this paper to do so, since the content of the two datasets is very similar, and the dataset described in section 4.2 is a parallel dataset.

For the dataset described in section 4.5, although it is non-parallel, the authors have provided a small parallel subset, rewritten by hand, so that the relevant tests will be performed using it.

5 Models

In order to find the latest sentiment transfer models to test their performances on different true style transfer tasks, we will look for the latest state-of-the-art sentiment transfer models. We have selected five representative models with different methods, which are word replacement [Li et al.(2018)], encoder-decode rule based model [Shen et al.(2017)], back-translation [Prabhumoye et al.(2018)], GAN-based learning [Zhao et al.(2018)] and reinforcement Learning [Xu et al.(2018)]. All five articles are open source and have good performance models. All five models use the same Yelp dataset, so it is easy for us to make comparisons. Next, we will look deeply into these models.

5.1 Word Replacement Model

For the word replacement method, we chose the model mentioned in this article [Li et al.(2018)]. When doing text style migration, it is important to make the converted sentence do as much as possible of these three things, Attribute Transfer, Content Preservation and Grammaticality. By observing the sentences in the corpus, the authors found that the attributes of the text are generally determined by the specific sentiment words in the sentences. And based on this, they propose the method in this paper, which defines the text sentiment transformation task as changing only the sentiment of the sentence without changing the parts that are not related to the sentiment, for example, "The chicken was delicious - The chicken was bland". Based on this, the author introduced a method to train the model, and there are three parts, delete, retrieve, and generate.

5.1.1 Delete

In this paper, the author defines the relevance of a word u to a sentiment v as

$$s(u,v) = \frac{count(u, D_v) + \lambda}{(\sum_{v' \in V, v' \neq v} count(u, D_{v'})) + \lambda}$$

count is used to count the number of times the word u appears in the sentiment set. λ is the parameter used for smoothing, and when $s(u,v)$ exceeds the threshold λ, it is assumed that u is the part that is relevant to the sentiment.

5.1.2 Retrieve

After deleting the sentiment words from the source sentence, in order to determine what words to insert into c (what is left after deletion), one can go through the set of sentences containing the target sentiment to find sentences that are similar to the original sentence and extract the sentiment words, which is shown in the following formula.

$$x^{tgt} = \arg\min_{x' \in D_v tgt} d(c(x, v^{src}), c(x', v^{tgt}))$$

5.1.3 Generate

The authors use four strategies to generate new sentences (the first two serve as baselines).

RETRIEVEONLY That is, the similar sentences found in the retrieve step are directly output as the result.

Advantage: It will always generate a grammatically sound sentence that contains the target sentiment. Disadvantage: The resulting sentences may differ greatly in content.

TEMPLATEBASED Extract the sentiment words of similar sentences found in retrieve and replace the sentiment words of the original sentence.

Advantages and disadvantages: generally requires the original sentence and the target sentence to be in similar contexts, which is feasible in most cases, but sometimes results in the output sentence being grammatically incoherent.

DELETEONLY The content c obtained after deleting the sentiment words is embedded and fed to an RNN, which splices the output of the last layer of the hidden layer with the learned target sentiment words and feeds it to the RNN decoder to produce a new sentence.

DELETEANDRETRIEVE Similar to DELETEONLY, c is first embedded into one RNN, then the target sentiment words extracted after retrieval are embedded into another RNN, and then spliced and fed into an RNN decoder to produce a new sentence.

The whole model can be seen in the following picture.

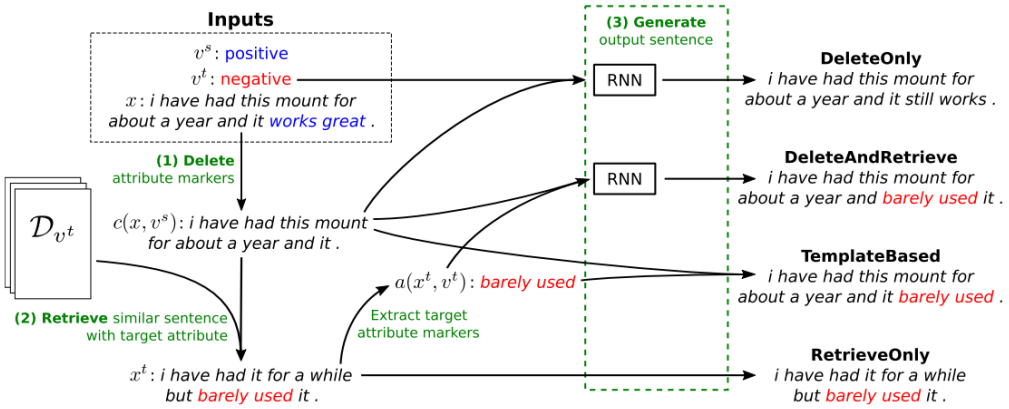

Figure 6: Model structure [Li et al.(2018)]

5.1.4 Training

Now let's see how to train DELETEONLY and DELETEANDRETRIEVE. Because the training set of this paper is some sentences labeled with their belonging sentiment (positive or negative), not just pairs of sentences with different sentiments, the author cannot get v^{tgt} when training DELETEONLY and DELETEANDRETRIEVE models with just the training set. So we can only use the training set for unsupervised learning, so the objective function of DELETEONLY is to maximize the following equation:

$$L(\theta) = \sum_{(x, v^{src}) \in D} \log p(x | c(x, v^{src}), v^{src}); \theta)$$

5.2 Encoder-Decoder Rule-Based Model

In this article [Shen et al.(2017)], the authors introduce a cross-corpus (two corpora have the same content, but non-parallelized data) and precisely alignable representation that, by learning an encoder, can map the input to a style-independent content representation. This is then passed to a style-dependent decoder for decoding. This model is characterized by not using VAE (Variable Auto-Encoder), as it is needed to make the latent expression richer and more natural. In addition, this model has a similar structural model as the style transformation of CV. There are three main tasks: sentiment transfer, word substitution code-breaking, and restoration of word order. Focusing on style migration for non-parallel corpus texts, the authors address its research difficulty: how to separate textual content from attributes, proposing methods that assume the sharing of latent content distributions across different textual corpora and utilize the precise alignment of latent representations to perform style migration.

Methods The authors obtain additional information from the cross-generated (style-converted) sentences to obtain two distributional alignment constraints. For example, affirmative sentences stylistically transformed into negative sentences should, as a whole, match a given negative sentence. This cross-alignment is illustrated in the following figure. The encoder E encodes the sentence and maps it into the hidden space Z to obtain the content representation of the sentence. When the input to the generator G is the original sentence label $y1$, a sentence with the same style as the original is generated (reconstruction, equivalent to Auto-Encoder), which is aligned to the original sentence; when a different style is input to the generator G, a sentence of the target style is generated, which is cross-aligned with the target sentence by one cross.

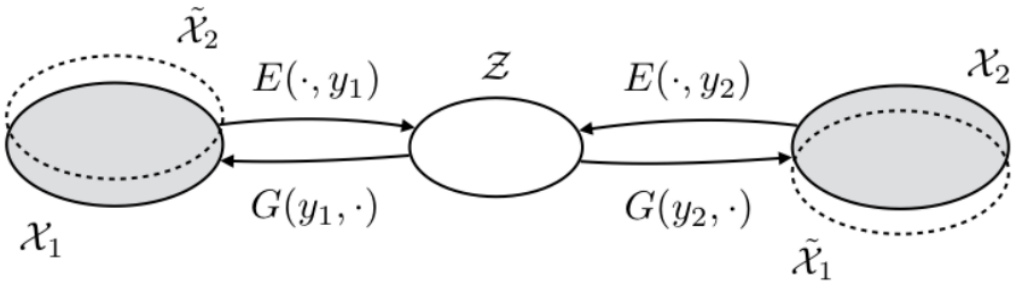

Figure 7: Cross alignment [Shen et al.(2017)]

Formalization For the formalized part, We can focus on the encoder-decoder function.

$$\begin{aligned} p(x_1|x_2 : y_1, y_2) &= \int_z p(x_1, z|x_2 : y_1, y_2)dz \\ &= \int_z p(z|x_2, y_2) \cdot p(x_1|y_1, z)dz \\ &= E_{z\sim p(z|x_2,y_2)}[p(x_1|y_1, z)] \end{aligned}$$

It consists of two main steps:

Step 1: Encoding, mapping into the latent space z. $z \sim p(z|x_2, y_2)$

Step 2: Decoding, and generating the target sentence. $p(x_1|y_1, z)$

Aligned auto-encoder The whole point of the model is to obtain a representation of the latent space z. Omitting the VAEs that make explicit assumptions on $p(z)$ and make the posterior values consistent with $p(z)$, the author align $p_E(z|y_1)$ and $p_E(z|y_2)$ with each other to obtain the following constrained optimization problem.

$$\begin{aligned} \theta &= \arg\min L_{rec}(\theta_E, \theta_G) \\ \text{s.t.} \quad E(x_1, y_1) &= E(x_2, y_2) \quad x_1 \sim X_1, x_2 \sim X_2 \end{aligned}$$

In practice, Lagrangian relaxation, a method of decomposition, is the primal problem rather than optimization. The authors introduce an adversarial discriminator D to calibrate the aggregated posterior distribution for different types of z, which can be seen in the following picture. The authors use single-layer RNNs with GRU units to implement the encoder E and the generator G. E acquires an input sentence x with an initial hidden state y and outputs the last hidden state z as its content representation. g generates a sentence x conditioned on the latent state (y, z). To align the $z_1 = E(x_1, y_1)$ and $z_2 = E(x_2, y_2)$ distributions, the discriminator D is a feedforward network with a single hidden layer and a sigmoid output layer.

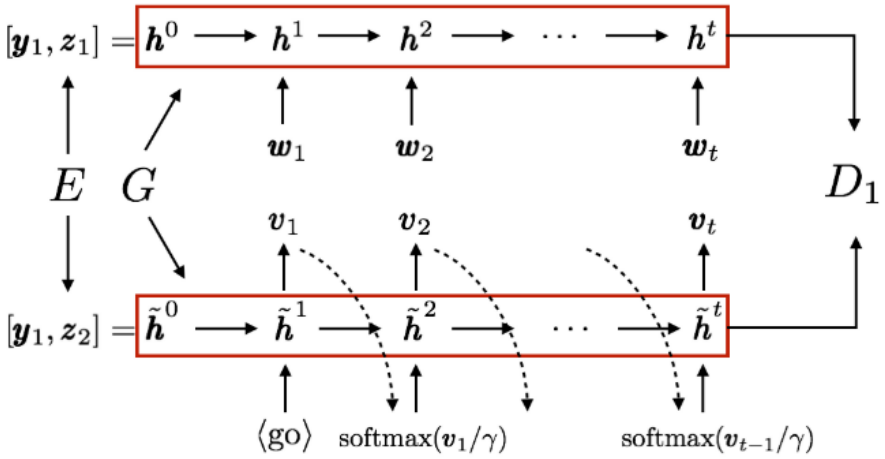

Figure 8: Discriminator [Shen et al.(2017)]

5.3 Back Translation Model

The paper[Prabhumoye et al.(2018)] introduces a new method for automatic style transfer. The authors first learn a latent representation of the input sentence which is grounded in a language translation model in order to better preserve the meaning of the sentence while reducing stylistic properties. Then adversarial generation techniques are used to make the output match the desired style. Compared to two state-of-the-art style transfer modeling techniques they show improvements both in automatic evaluation of style transfer and in manual evaluation of meaning preservation and fluency.

In order to design a style transfer model, the authors refer to the idea of [Shen et al.(2017)] and consider that one of the important steps is to use an excellent method to model a latent content variable z. The latent content variable z is used to assign a specific direction to the generated model. Compared to the cross-alignment approach proposed by [Shen et al.(2017)], the authors propose to design a new latent content variable z, which can be divided into two steps, and the specific structure can be seen in the

following figure.

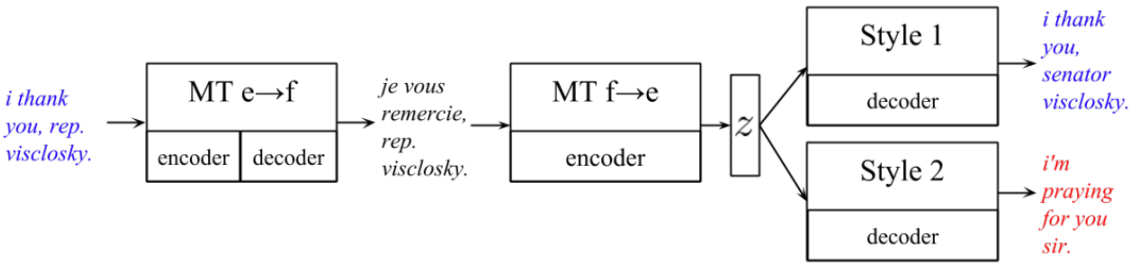

Figure 9: pipeline [Prabhumoye et al.(2018)]

The first step is to represent the meaning of the input sentence in the back-translation. Using the method of back-translation, the meaning of the sentence itself can be effectively preserved and the original style it contains can be eliminated. The authors first train a machine translation model from the source language e to the target language f. Then they also train a back-translation model from f to e. The back-translation model encoder creates an encoder that translates from the source language to the target language. The latent representation created by the encoder of the back-translation model is used as z.

Formally, let θ_E represent the parameters of the encoder of f to e translation system. Then z is given by:

$$z = Encoder(x_f; \theta_E)$$

where x_f is the sentence x in the language f. Specifically, x_f is the output of the e to f translation system. Since z is produced by a non-style-specific process, this encoder is not style-specific.

The second step is to weaken the style attributes of the original sentence itself. The generative model can then generate sentences with the target style according to z. The author trained a bidirectional LSTM to build the decoder model. Words are generated in the following statement.

$$\begin{aligned} \hat{x} \sim z &= p(\hat{x}|z) \\ &= \prod_t p(\hat{x}_t|\hat{x}^{<t}) \end{aligned}$$

5.4 GAN-based learning model

GAN performs well on continuous structures but performs poorly when applied to discrete structures (e.g., text sequences or discrete images), and in this paper [Zhao et al.(2018)],

the authors address this by proposing a flexible approach to deal with this problem.

ARAE combines discrete autoencoders with GAN regularized latent representations and can be seen in the following picture. We will discuss it separately.

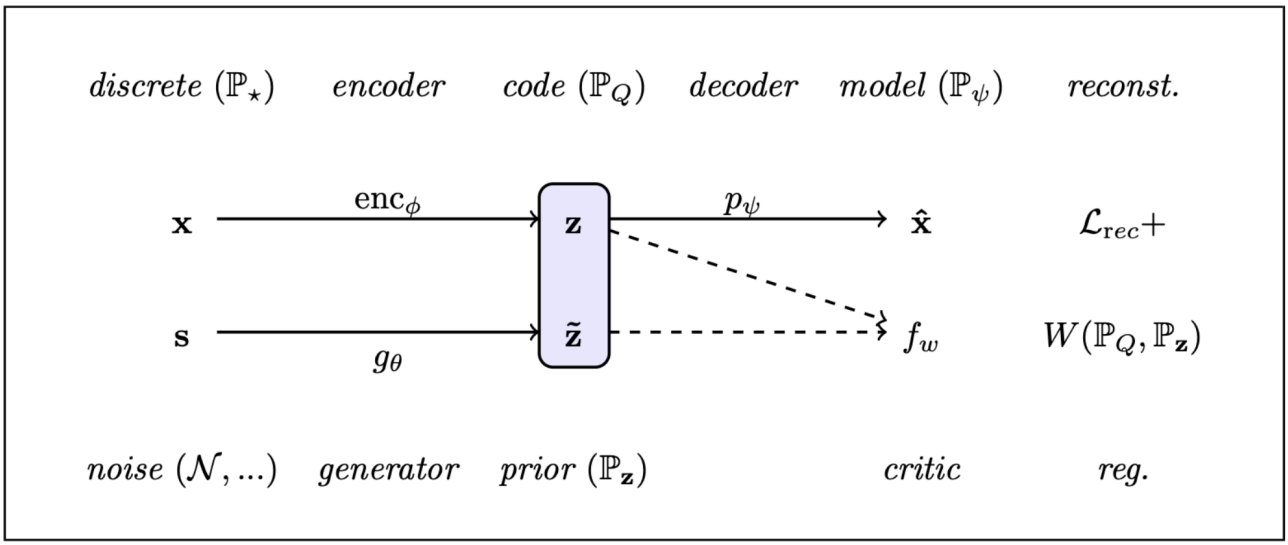

Figure 10: ARAE [Zhao et al.(2018)]

Discrete Autoencoders This part encoder the discrete sequence and then decoder it and discretize it by softmax. The encoder and decoder are problem-specific, and RNNs are generally chosen. The loss function can be seen below.

$$L_{\text{rec}}(\phi, \varphi) = -\log p_\varphi(x|enc_\phi(x))$$

GAN network GANs are types of generative models. Previous conventional GAN models were poor for feature extraction of text, so the authors used a Wasserstein GAN network, WGAN training uses the following min-max optimization for the generator θ and the critic w.

$$\min_\theta \max_{w \in W} E_{Z \sim \text{P}}[f_w(z)] - E_{\tilde{z} \sim P}[f_w(\tilde{z})]$$

ARAE The model structure is actually the combine of the discrete autoencoders and the GAN network. However, it is difficult to optimize the model when it is used directly on discrete structures (e.g., text sequences), so when this model is used for non-aligned textual style migration (sentiment and topic), it needs to be able to separate out the attribute information in the sentence so that it can be contained in y, while the rest of the information is contained in the latent variable z.

$$p_\varphi(x|z, y)$$

This latent space is used to generate new data points that are similar to the input data, as well as to manipulate the input data in various ways. By learning a smooth and structured latent space, the model can perform more effective transformations and generate more realistic outputs.

5.5 Reinforcement Learning

The authors of this article [Xu et al.(2018)] face the same problem as other researchers on the same page, that is, they cannot guarantee that the content will remain the same while the style is migrating. The reason for this is that content and style are in the same hidden vector and it is difficult to interpret all the information mixed together. Because there is no parallel corpus, it is difficult to keep the semantic information unaffected.

This paper proposes a recurrent reinforcement learning model. It includes a neutralization module and an emotionalization module. The neutralization module acts to remove emotion words to extract non-emotional semantic information. The role of the emotionalization module is to add emotion words to make neutral sentences emotional. The core idea is that, in the first step, the neutralization module removes the sentiment first, and then the emotionalization module reconstructs the original sentence based on the original sentiment and semantic content, allowing the emotionalization module to learn to add the sentiment in a supervised manner. In the second step, the sentiment words are reversed so that adding the opposite sentiment words can be realized.

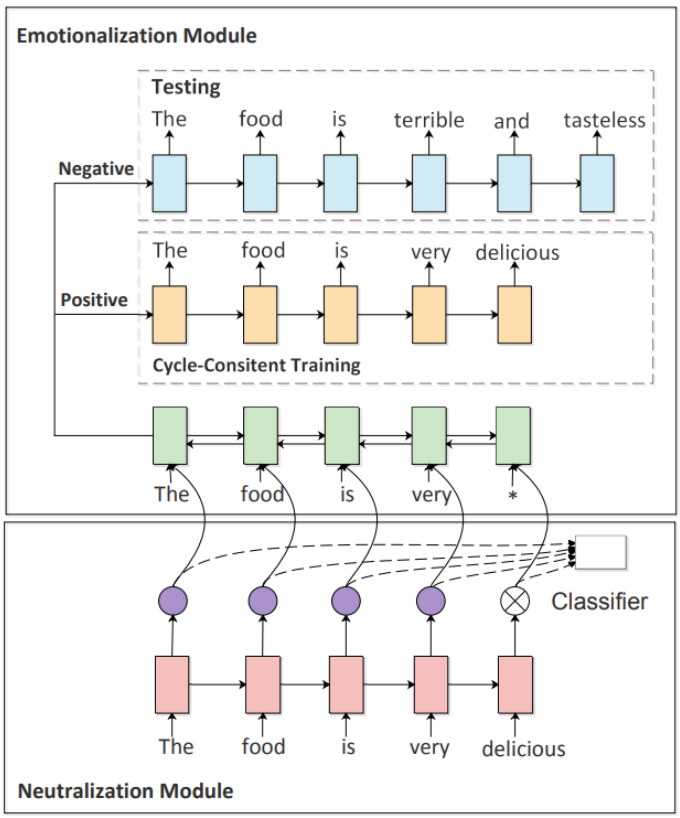

Figure 11: An illustration of the two modules [Xu et al.(2018)]

Neutralization It is an LSTM+Attention sentiment classifier for removing sentiment words. LSTM is used to generate the probability that each word is a neutralizer or polarizer. Recurrent reinforcement learning requires the model to have an initial learning capability, so a pre-training method is proposed to allow the neutralization module to learn to judge non-emotional words. The pre-training uses a sentiment classifier based on a self-attention mechanism that uses the attention weights as a guide. The reason for doing so is that in a trained sentiment classifier model, the attention weights reflect to some extent the contribution of each word to the sentiment. Usually sentiment words are weighted heavily and neutral words are weighted less. The experimental results show that the sentiment classification accuracy reaches 89%-90%, and it can be assumed that the classifier adequately captures the sentiment information of each word. Non-sentiment words are extracted based on the weights, which are discretized into 0 and 1. If the weight of a word is less than the average of the weights of the sentence, its discretization value is 1, otherwise it is 0. Sentiment words are weighted as 1, and non-sentiment words as 0. Putting this result can help to de-sentiment.

Emotionalization Emotionalization module is responsible for adding emotional words. A seq2seq (bi-decoder) model is used, both encoder and decoder are LSTM. there are

two decoders for adding positive and negative sentiment words respectively.

6 Experiments

6.1 Evaluation Protocol

Evaluating a text-to-text task is challenging. In order to properly evaluate the performance of each model, we will use different metrics as criteria.

6.1.1 BLEU score

In this paper we will use BLEU to review the quality of the generated sentences. We will compute the reference BLEU and self BLEU, denoted BLEU-ref and BLEU-self, respectively. Both values are averaged from n-grams with values 1 to 4. Due to space constraints, only the n-gram values of BLEU-ref will be recorded, denoted as BLEU-1, BLEU-2, BLEU-3, and BLEU-4, respectively.

6.1.2 ACC score

We will calculate the ACC using the same classifier that pre-processed the datasets. Specifically, we generate sentences given a style s and assign labels to the generated sentences using the same classifier used by the authors of the dataset. The accuracy is computed as the percentage of predictions that match style s. The accuracy is calculated as the percentage of predictions that match style s.

6.1.3 ROUGE score

In this paper we will mainly compute the ROUGE of Longest Common Subsequence, ROUGE-L. ROUGE-L measures the longest common subsequence between a machine-generated text and a reference text.

6.1.4 BERT score

We will calculate the F1 of the BERT model and use it as BERT score.

6.2 Experimental settings

In order to effectively evaluate the performance of the models described above on the selected datasets, we chose a unified platform for testing.

- Hardware Platform. All experiments were conducted on a server equipped with an NVIDIA RTX 4090 graphics card. The server was equipped with 32 GB of RAM and an AMD 5900x processor.
- Software Environment. Each model uses a separate Anaconda environment with the same version of the Tensorflow or PyTorch frameworks used by the authors.

- Hyperparameter Settings: For each model we select, we perform hyperparameter tuning the same as the authors did in their papers. For example, for the GAN-based learning method model, we set hyperparameters such as the learning rate, batch size, and hidden layer dimension of the generator and discriminator the same as the authors did in their paper.

- Training Process: We train each model on the corresponding dataset.

  The table below shows the details of each model.

Model	batch-size	epochs	optimizer
Word Replacement	256	20	Adadelta [Zeiler(2012)]
Encoder-decoder rule-based	64	20	SGD [Amari(1993)]
Back translation	64	20	CNN
GAN based	64	25	SGD and Adam [Reddi et al.(2019)]
Reinforcement Learning	64	20	Adagrad [Duchi et al.(2011)]

Table 8: Hyperparameter setting

- Testing Procedure: After completing the training, we use the model to transfer styles to the texts on the test set. For different datasets and evaluation metrics, we evaluate the performance of the model.

7 Experimental results

7.1 Matrix scores

The following table 9 shows the results of each model run on the five datasets and table 10 shows the average scores of five datasets for the models.

Model Name	Dataset	BLEU-ref	BLEU-self	ROUGE	ACC	BERT	BLEU1	BLEU2	BLEU3	BLEU4
Encoder-Decoder Rule-Based	GYAFC	7.371475	10.4827	23	31.1	82.52	25.6976	3.2901	0.4663	0.0319
	Politeness	4.2471	6.2981	45.6	39	75.21	11.5232	4.5234	0.9376	0.0042
	Toxic Transfer	5.674425	9.9271	31.4	47	69.23	20.2817	2.3817	0.0342	0.0001
	Twitter Formal	6.2876	7.9171	24.2	46.9	75.3	23.7231	1.4251	0.0021	0.0001
	Offensive	9.3713	14.9827	35.3	36.9	83.54	32.1234	5.2341	0.1234	0.0043
Back Translation	GYAFC	0.380125	8.918	26.7	32.1	63.43	1.4462	0.0742	0.0001	0
	Politeness	0.965525	4.9287	37.3	32.5	64.3	3.6241	0.2314	0.0064	0.0002
	Toxic Transfer	1.917775	2.9181	34.5	28	59.4	5.6927	1.9762	0.0021	0.0001
	Twitter Formal	1.26865	4.8271	22.7	42.1	57.2	4.2392	0.8271	0.0082	0.0001
	Offensive	2.293175	7.8323	24.1	40.1	53.7	7.3421	1.8261	0.0043	0.0002
GAN-based learning	GYAFC	2.7993	5.9372	25.3	28.2	76.4	9.2573	1.9273	0.0123	0.0003
	Politeness	6.9371	9.9271	35.2	52.3	74.2	15.3421	11.0584	1.3421	0.0058
	Toxic Transfer	5.14165	6.9827	22.1	34.2	68.9	18.3826	2.1826	0.0012	0.0002
	Twitter Formal	9.74585	13.9271	46.1	53.2	80.1	34.2346	4.1927	0.532	0.0241
	Offensive	6.291325	10.9271	32.5	58.3	77.2	23.2123	1.9271	0.0234	0.0025
Reinforcement Learning	GYAFC	1.0239	5.9827	42.8	52.8	80.2	3.1632	0.9271	0.0052	0.0001
	Politeness	2.038925	4.8213	34.2	43.1	67.2	7.2341	0.9172	0.0043	0.0001
	Toxic Transfer	3.2589	5.8272	23.1	42.2	86.2	10.8572	2.1722	0.0062	0
	Twitter Formal	5.5385	6.2812	32.5	56.2	64.2	21.1234	1.0283	0.0021	0.0002
	Offensive	5.476275	7.2827	16.3	44.2	77.2	18.9271	2.9372	0.0346	0.0062
Word Replacement	GYAFC	7.0755	8.2971	42.1	35.1	89.8	23.9271	4.2981	0.0726	0.0042
	Politeness	9.277325	19.8271	32.1	32.1	75.5	34.2856	2.8162	0.0074	0.0001
	Toxic Transfer	11.1236675	16.9827	38.2	45.6	65.2	42.2917	2.1826	0.01827	0.0021
	Twitter Formal	6.464625	10.0021	31.1	33.2	86.3	23.9271	1.9271	0.0042	0.0001
	Offensive	9.31195	19.9921	45.4	34.3	62.4	32.9271	4.2912	0.0291	0.0004

Table 9: Results

Model Name	BLEU-avg	ROUGE-avg	ACC-avg	BERT-avg
Encoder-Decoder Rule-Based	6.59038	31.9	40.18	77.16
Back Translation	1.36505	29.06	34.96	59.606
GAN-based learning	6.183045	32.24	45.24	75.36
Reinforcement Learning	3.4673	29.78	47.7	75
Word Replacement	8.6506135	37.78	36.06	75.84

Table 10: Average Results

7.2 Models On YELP datasets

The table below shows the BLEU metrics for each model on the Yelp dataset from the original papers. Although we do not have access to other data that are more indicative of model performance, it can be seen from this table that the five models selected all have far better performance on their own original Sentiment transfer task than on the other style transfer tasks.

Model	BLEU
Word Replacement	17.75
Encoder-decoder rule based	6.8
Back translation	31.15
GAN based	22.46
Reinforcement Learning	29

Table 11: BLEU on Yelp dataset

7.3 Generated examples

These test metrics mentioned above alone do not fully prove our argument, so we will list some of the model-generated sentences as examples so as to continue our argument. Considering the length, we will pick the three models with the highest and lowest ACC, Encoder-Decoder Rule Basd Model, Back Translation Model, and Word Replacement model for analysis. Three models will be listed for each model, and three forward sentences and three reverse sentences will be selected for each model to observe the output.

Datasets	Input sentence	Output sentence
Formal to Informal	It is on their greatest hits albumn.	it is on the <unk>
	I'm not fond of becoming soiled.	it is more on <unk>
	Emma Watson is not very nice looking and not very nice.	plus it is so much is so much as well.
Informal to Formal	Any movie that has vampires I like!	The song is <unk> I am it?
	Familt Guy, The Simpons, Futurama, and South Park!!!!	<unk> <unk> <unk> <unk> <unk> <unk> <unk> <unk> etc.
	funny but I've heard it before	It is what it is fun.
Polite to Not Polite	thanks for visiting with us today .	thanks for us with us today
	the attached has louise 's comments to sections 3.2 and 3.3 .	the attached is thursday 's comments to discuss and nevertheless .
	we met with our database administrator this afternoon .	we met with our group this afternoon .
Not Polite to Polite	florida : you are a laid back person .	: you are going a call back person .
	are we serious about selling the gas ?	are we doing about the gas gas ?
	turn right at this hangar – this is chanute road .	turn right at this is the <unk> – this is weekend .
Offensive to Not Offensive	CUCKOLD Carlson is a problem	Carlson is the <unk>
	shut tf up lmfaoooo you're soooo victimized	shut up <unk> <unk> are <unk>
	This is a cowardly response.	That is what <unk>
Not Offensive to Offensive	Ummm, for obvious reasons?	<unk>, it is for reasons.
	Obviously.	Yes.
	Therapy may help you.	<unk> you help.

Table 12: Encoder- Decoder Rule Basd Mode

Datasets	Input sentence	Output sentence
Formal to Informal	It is on their greatest hits albumn.	it's their greatest <unk> albumn.
	I'm not fond of becoming soiled.	i like becoming <unk>
	Emma Watson is not very nice looking and not very nice.	Emma Watson <unk> <unk> <unk> <unk>
Informal to Formal	Any movie that has vampires I like!	any movie does vampires like it.
	Familt Guy, The Simpons, Futurama, and South Park!!!!	Family Guy, <unk>, <unk>, and <unk>
	funny but I've heard it before	I heard funny before it
Polite to Not Polite	thanks for visiting with us today .	go visit us today
	the attached has louise 's comments to sections 3.2 and 3.3 .	the attached doesn't has comments
	we met with our database administrator this afternoon .	we didn't met her today.
Not Polite to Polite	florida : you are a laid back person .	: <unk>, don' t laid person.
	are we serious about selling the gas ?	we are serious selling these gas.
	turn right at this hangar – this is chanute road .	turn right at this hangar <unk> <unk> line.
Offensive to Not Offensive	CUCKOLD Carlson is a problem	She is problem
	shut tf up lmfaoooo you're soooo victimized	<unk> <unk> <unk> good victimized
	This is a cowardly response.	These are not responses.
Not Offensive to Offensive	Ummm, for obvious reasons?	What is the reasons
	Obviously.	True.
	Therapy may help you.	<unk> can help.

Table 13: Back Translation Model

Datasets	Input sentence	Output sentence
Formal to Informal	It is on their greatest hits albumn.	it's on there greatest hits album
	I'm not fond of becoming soiled.	i don't like to get soiled.
	Emma Watson is not very nice looking and not very nice.	Emma Watson not nice looking not nice.
Informal to Formal	Any movie that has vampires I like!	I like any movie that has vampires.
	Familt Guy, The Simpons, Futurama, and South Park!!!!	Family Guy, The Simpons, Futurama, and South Park
	funny but I've heard it before	It is funny, but I have heard it before.
Polite to Not Polite	thanks for visiting with us today .	visiting us today
	the attached has louise 's comments to sections 3.2 and 3.3 .	the attached has louise 's comments
	we met with our database administrator this afternoon .	we met our database this afternoon .
Not Polite to Polite	florida : you are a laid back person .	: florida : are you laid back person .
	are we serious about selling the gas ?	are we good about selling the gas ?
	turn right at this hangar – this is chanute road .	turn right at this hangar – this is chanute road .
Offensive to Not Offensive	CUCKOLD Carlson is a problem	CUCKOLD Carlson is problem
	shut tf up lmfaoooo you're soooo victimized	you're victimized
	This is a cowardly response.	That is a response
Not Offensive to Offensive	Ummm, for obvious reasons?	for what reasons?
	Obviously.	Obviously.
	Therapy may help you.	Therapy help you well.

Table 14: Word replacement Mode

7.4 Analysis of results

In order to fully compare the performance of each model, we will compare the performance of different models on the same task, the performance of the same model on different tasks, analyze it using various metrics as well as generated sentence samples, and compare it with the model's performance on the Yelp dataset. Offensive transfer is one of the more representative types of style transfer. We will use it to compare the performance of different models.

First we will compare the model with the highest average BLEU score, the Word replacement model. We can observe the metric scores on different datasets in Table 9. For the Word replacement Model, both BLEU-self and BLEU-ref are significantly higher than other models. For BLEU-ref, the avarage for 5 datasets are 8.7, indicating that the model generates text that is more similar to itself and the original text. In the "toxic transfer" dataset, the model has the highest BLEU-ref, which is 11.12, the reason can be inferred by the datasets itself, the aggressiveness of some sentences comes from some specific words. Some data outputs can be observed in Table 15. For example, for the sentence "This is a cowardly response", the model successfully recognizes that the Offensive attribute of the sentence comes from the word "cowardly". Therefore, removing this word will change the style. Thus the advantage of this model in dealing with Toxic transfers may be due to the fact that the core principle of this model is word substitution, and the style of Toxic can be resolved with relative ease by word substitution. This can also be demonstrated by observing the Offensive dataset.

Observing the output of Table 17, it can be noticed that for most of the sentences, the present model tends to modify the style by deleting some of the words. A small portion of them are correct, but most of the rest of the sentences are problematic. Some of them are grammatical errors, for example, "Emma Watson is not very nice looking and not very nice." is transformed into "Emma Watson not nice looking not nice ." The other part of the model did not understand the meaning of the sentence at all and made a wrong transformation, for example, "Therapy may help you." was transformed into "Therapy help you well." The sentence should have been transformed from Not Offensive to Offensive, but it became more polite after the model output. This sentence should be changed from Not Offensive to Offensive, but after modeling, it becomes more polite instead. Also, in terms of metrics, the present model, while having better metrics relative to other models, obtaining a BLEU-ref of 11.12, is still far from the 17.75 BLEU achieved by the model on Yelp. When this model is processing the Yelp dataset, in addition to deleting words, it also accurately replaces them and adds sentences with the target style. For example, in Yelp dataset, "we sit down and we got some really slow and lazy service“ can be easily transferred to "we sit down and we got some great and quick service ." by just substituting two words "slow" and "lazy". This exemplifies the point of the paper, which is that sentiment transfer changes the meaning of the

sentence itself.

The remaining metrics of the present model are also of interest.The ACC of the Word replacement model has an accuracy of only 34.3 on the Offensive transfer task, indicating that the model performs even worse than a coin toss. This proves that the model does not recognize sentences at all, which corroborates with the argument above that the present model cannot correctly handle tasks other than sentiment transfer.

Next, we will compare the model with the lowest average BLEU score, the Back Translation Model. Unlike the word substitution model, this model is a typical encoder-decoder model, which maps input sentences to genre-independent content representations through a learning encoder, and then generates output sentences matching the target genre through a decoder.

Observe Table 10 to learn that BLEU-avg is only 1.36, which is significantly lower than the rest of the models. Also, observe Table 11 to learn that this model performs well on the Yelp dataset with a BLEU score of 31.15. The significant difference that occurs here may be due to the Back Translation architecture of the model. It contains two of the Encoder-Decoder structures, which may have caused a lot of damage to the sentence structure, hence the low BLEU score. In addition, as argued in this paper, the sentiment transfer only changes some of the words and it's not a true style transfer, so the present model has a higher BLEU score on the task for the sentiment transfer and a much lower BLEU score on other tasks.

More details can be seen in the actual output. By looking at Table 13, we find that there are a lot of <unk> in the output data of this model compared to the Word replacement model in Table 14. This is due to the fact that the present model, unlike Word replacement, involves compiling information into Latent space and getting new sentences from it. Therefore there may be some words that the model does not recognize correctly. For example, in the Not Offensive to Offensive task, the sentence "Therapy may help you." is transformed into "<unk> can help.", where the word "Therapy" is not recognized correctly. Meanwhile, we found that none of the three sentences in the example were correctly transformed into the Offensive property in the Not Offensive to Offensive task. However, in the corresponding Offensive to Not Offensive task, although the generated sentences have grammatical errors, some words with obvious Offensive attributes are correctly removed, e.g., the above mentioned "This is a cowardly response" is processed by the model to be For example, "This is a cowardly response" mentioned above is processed by the model as "These are not responses.", removing the word "cowardly". This can be found in the characteristics of the present model, which tends to transfer the style of a sentence more than maintaining its format.

8 Discussion

8.1 Summary of key findings

In this paper, we explore the problem of style transformation in natural language processing with the aim of evaluating the performance of sentiment transformation models in real style transformation tasks.

To answer the first RQ of this paper, "Find and classify datasets of real style transfer", we first searched for datasets related to real text style transformation. These datasets include the GYAFC, Toxic, and Politeness datasets, which cover various stylistic domains such as formality and politeness transfer. The datasets are presented in highly cited articles and have been tested by many authors. By categorizing texts based on tone and content, we prepared the data for our experiments.

To answer the second RQ of this paper, “Find and classify state-of-the-art models for sentiment transfer.", We chose five representative models: word substitution, back translation, GAN-based learning, reinforcement learning, and encoder-decoder rule-based models. All these models are open source and are known for their good performance in sentiment transfer.

To answer the third RQ of this paper, "Test these models on the dataset and compare the test results with previous findings on sentiment transfer." We then run these selected models on selected datasets and compare the test results with previous sentiment transfer studies. In order to evaluate the performance of the models in these tasks, we used standard natural language processing evaluation metrics such as the BLEU, ROUGE, and BERT scores for automated evaluation.

The results of our experiments show that the five models are unable to process the provided data correctly. The generated sentences are very difficult to read and the accuracy of style recognition is very low. Compared to the Yelp dataset originally used by each model, the five selected models generated sentences with very low BLEU metrics, as well as sentences with many grammatical errors and many unrecognizable words. According to the ACC metrics, almost all of the models failed to correctly recognize the style of the sentences in the test set, with less than 50% correctness. Despite their success in sentiment transfer, they were unable to maintain semantic coherence when changing style, resulting in incoherent and meaningless output.

These findings suggest that current state-of-the-art emotion transfer models have significant limitations when applied to real-world style transfer tasks. It emphasizes the need for further research and development to create more effective and reliable models that accurately transform sentence styles while preserving sentence meaning. In addition, it emphasizes the importance of carefully evaluating models in real-world style-switching tasks to ensure their practical applicability and validity.

In summary, our study provides important insights into the challenges and limita-

tions of existing emotion conversion models for textual style conversion. By emphasizing the differences between sentiment conversion and style conversion tasks, we hope to promote further research in this area and contribute to the development of textual style conversion models in natural language processing.

8.2 Limitation

Although our study aimed to find representative style transfer models and evaluate their performance in a real style transfer task, certain limitations should be recognized:

Limited style domains: Our assessment focused on specific stylistic domains such as formality and politeness. However, the concept of style is very broad and includes aspects such as humor, irony, informality, and regional differences. The assessment may not cover all possible stylistic transitions, and there may be other stylistic aspects that our chosen model or algorithm cannot handle effectively.

Method of assessment: Style itself is subjective and may vary from person to person. What one person considers formal, others may consider informal. Thus, assessing the success of a style shift becomes very challenging. Due to resource constraints, there is no way we can hire a manual measurement to judge the model, so we can only assess the performance of the model by using some test metrics and by observing some of the generated data.

Dataset bias: The availability and size of suitable datasets for style transformation can affect the performance of the model.l For example, there is partial overlap between the Offensive dataset used in this paper and the Toxic dataset, but the size of the former is much larger than the latter. If the training data is biased towards certain style domains or lacks style-specific features, this may affect the model's ability to handle these styles effectively.

Computational resources: Evaluating models in real style transformation tasks usually requires significant computational resources. Depending on the complexity of the model and the size of the dataset, the experiments may require a large amount of computational resources, thus limiting the scope of our evaluation.

Despite these limitations, we believe that our study provides valuable insights into style transfer. By being transparent about these limitations, we hope to encourage further research in this area to address these issues and drive the development of more robust and accurate style transfer models.

8.3 Relevance of the thesis topic to the field of COSC

In keeping with the field of computer science (COSC), the focus of this thesis - style transfer - has progressed in an ever-evolving technology. NLP techniques have attempted from their origins to enable machines to understand human language. This

has been a long process of development. By now, it is possible to understand human language using large generative models, such as GPT. The conclusions of this thesis contribute to the gradual development of AI models' understanding of the style of text.

In addition, the research on sentiment style transfer in this thesis explores the boundaries of real style transfer. Through experiments as well as tests, we argue that the two, sentiment transfer and style transfer, are fundamentally different and should not be treated equally.

In conclusion, the thesis theme is a bridge between advances in stylistic transfer and the dynamic field of COSC.

8.4 Future directions and potential improvements

Now Large Language Model (LLM) has begun to gradually come into the limelight. Generative models, led by GPT-4 [Mao et al.(2023)], are gradually changing the direction of NLP research. This has a great impact on the traditional language model chosen in this paper. Compared with the model chosen in this paper, LLM generates sentences of higher quality and can easily handle the style chosen in this paper. However, the model chosen in this paper still has its value in today's world. LLM requires a large training facility and a high cost, which will limit the general researchers to study it. In addition, deploying an LLM requires a large amount of resources, whereas in many cases we only need to deal with a single task, the transfer of a particular class of sentences to a particular style. For example, a style transfer of offensive sentences in the comments section of a forum removes the offense and maintains a friendly conversation. For example, a style transfer of offensive sentences in the comments section of a forum removes the offense and maintains a friendly conversation.

References

[Amari(1993)] Shun-ichi Amari. 1993. Backpropagation and stochastic gradient descent method. *Neurocomputing* 5, 4-5 (1993), 185–196.

[Arsyad and Adila(2018)] Safnil Arsyad and Destiantari Adila. 2018. Using local style when writing in English: the citing behaviour of Indonesian authors in English research article introductions. *Asian Englishes* 20, 2 (2018), 170–185.

[Atwell et al.(2022)] Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations. *arXiv preprint arXiv:2209.08207* (2022).

[Bar-Hillel(1960)] Yehoshua Bar-Hillel. 1960. The present status of automatic translation of languages. *Advances in computers* 1 (1960), 91–163.

[BM([n. d.])] Nechu BM. [n. d.]. Back Propogation. https://towardsdatascience.com/what-is-an-encoder-decoder-model-86b3d57c5e1a.

[Chen(2022)] Guanyi Chen. 2022. *Computational Generation of Chinese Noun Phrases*. Ph. D. Dissertation. Utrecht University.

[Chen et al.(2023)] Guanyi Chen, Fahime Same, and Kees van Deemter. 2023. Neural referential form selection: Generalisability and interpretability. *Computer Speech & Language* 79 (2023), 101466.

[Chen and van Deemter(2020)] Guanyi Chen and Kees van Deemter. 2020. Lessons from Computational Modelling of Reference Production in Mandarin and English. In *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics, Dublin, Ireland, 263–272. https://aclanthology.org/2020.inlg-1.33

[Chen and van Deemter(2023)] Guanyi Chen and Kees van Deemter. 2023. Computational Modelling of Quantifier Use: Corpus, Models, and Evaluation. *Journal of Artificial Intelligence Research* 77 (2023), 167–206.

[Chen et al.(2018)] Guanyi Chen, Kees van Deemter, and Chenghua Lin. 2018. Modelling Pro-drop with the Rational Speech Acts Model. In *Proceedings of the 11th International Conference on Natural Language Generation*. Association for Computational Linguistics, Tilburg University, The Netherlands, 159–164. https://doi.org/10.18653/v1/W18-6519

[Chen et al.(2019)] Guanyi Chen, Kees van Deemter, and Chenghua Lin. 2019. Generating Quantified Descriptions of Abstract Visual Scenes. In *Proceedings of the 12th*

International Conference on Natural Language Generation. Association for Computational Linguistics, Tokyo, Japan, 529–539. https://doi.org/10.18653/v1/W19-8667

[Chen et al.(2020)] Guanyi Chen, Yinhe Zheng, and Yupei Du. 2020. Listener's Social Identity Matters in Personalised Response Generation. In *Proceedings of the 13th International Conference on Natural Language Generation*. Association for Computational Linguistics, Dublin, Ireland, 205–215. https://aclanthology.org/2020.inlg-1.26

[Crocker(1986)] Richard L Crocker. 1986. *A history of musical style*. Courier Corporation.

[Dictionary(2002)] Merriam-Webster Dictionary. 2002. Merriam-webster. *On-line at http://www. mw. com/home. htm* 8, 2 (2002).

[Duchi et al.(2011)] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research* 12, 7 (2011).

[Gatys et al.(2016)] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.

[Goddard(2011)] Cliff Goddard. 2011. *Semantic analysis: A practical introduction*. Oxford University Press.

[Goodman et al.(1992)] Rodney M Goodman, Charles M Higgins, John W Miller, and Padhraic Smyth. 1992. Rule-based neural networks for classification and probability estimation. *Neural computation* 4, 6 (1992), 781–804.

[Hu and He(2021)] Mingxuan Hu and Min He. 2021. Non-parallel text style transfer with domain adaptation and an attention model. *Applied Intelligence* 51, 7 (2021), 4609–4622.

[Jhamtani et al.(2017)] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161* (2017).

[Jin et al.(2019)] Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 2019. IMaT: Unsupervised text attribute transfer via iterative matching and translation. *arXiv preprint arXiv:1901.11333* (2019).

[Khalid and Srinivasan(2022)] Osama Khalid and Padmini Srinivasan. 2022. Smells like Teen Spirit: An Exploration of Sensorial Style in Literary Genres. *arXiv preprint arXiv:2209.12352* (2022).

[Khan et al.(2016)] Wahab Khan, Ali Daud, Jamal A Nasir, and Tehmina Amjad. 2016. A survey on the state-of-the-art machine learning models in the context of NLP. *Kuwait journal of Science* 43, 4 (2016).

[Koehn(2009)] Philipp Koehn. 2009. *Statistical machine translation.* Cambridge University Press.

[Kroll(1986)] Barry M Kroll. 1986. Explaining how to play a game: The development of informative writing skills. *Written communication* 3, 2 (1986), 195–218.

[Lai et al.(2019)] 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)*. 3579–3584.

[Laugier et al.(2021)] Léo Laugier, John Pavlopoulos, Jeffrey Sorensen, and Lucas Dixon. 2021. Civil rephrases of toxic texts with self-supervised transformers. *arXiv preprint arXiv:2102.05456* (2021).

[Li et al.(2018)] 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)*. Association for Computational Linguistics, New Orleans, Louisiana, 1865–1874. https://doi.org/10.18653/v1/N18-1169

[Li et al.(2020)] Xiao Li, Guanyi Chen, Chenghua Lin, and Ruizhe Li. 2020. DGST: a Dual-Generator Network for Text Style Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7131–7136. https://doi.org/10.18653/v1/2020.emnlp-main.578

[Lin(2004)] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[Liu et al.(2017)] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. *Advances in neural information processing systems* 30 (2017).

[Liu et al.(2020)] Yixin Liu, Graham Neubig, and John Wieting. 2020. On learning text style transfer with direct rewards. *arXiv preprint arXiv:2010.12771* (2020).

[Luo et al.(2019)] Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. *arXiv preprint arXiv:1905.10060* (2019).

[Ma et al.(2021)] Yun Ma, Yangbin Chen, Xudong Mao, and Qing Li. 2021. Collaborative learning of bidirectional decoders for unsupervised text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 9250–9266.

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

[Mao et al.(2023)] Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. GPTEval: A Survey on Assessments of ChatGPT and GPT-4. *arXiv:2308.12488* (2023).

[McDonald(2010)] David D McDonald. 2010. Natural language generation. *Handbook of natural language processing* 2 (2010), 121–144.

[McRoy et al.(2003)] Susan W McRoy, Songsak Channarukul, and Syed S Ali. 2003. An augmented template-based approach to text realization. *Natural Language Engineering* 9, 4 (2003), 381–420.

[Medhat et al.(2014)] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal* 5, 4 (2014), 1093–1113.

[Meng et al.(2019)] Yingying Meng, Deqiang Kong, Zhenfeng Zhu, and Yao Zhao. 2019. From night to day: GANs based low quality image enhancement. *Neural Processing Letters* 50 (2019), 799–814.

[Napoles et al.(2017)] Courtney Napoles, Joel Tetreault, Aasish Pappu, Enrica Rosato, and Brian Provenzale. 2017. Finding Good Conversations Online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the 11th Linguistic Annotation Workshop*. Association for Computational Linguistics, Valencia, Spain, 13–23. https://doi.org/10.18653/v1/W17-0802

[O'Connor and McDermott(2001)] Joseph O'Connor and Ian McDermott. 2001. *NLP*. Thorsons.

[Papineni et al.(2002)] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

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

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

[Rao and Tetreault(2018)] Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535* (2018).

[Reddi et al.(2019)] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2019. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237* (2019).

[Saffran et al.(1996)] Jenny R Saffran, Elissa L Newport, and Richard N Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of memory and language* 35, 4 (1996), 606–621.

[Same et al.(2022)] Fahime Same, Guanyi Chen, and Kees Van Deemter. 2022. Non-neural Models Matter: a Re-evaluation of Neural Referring Expression Generation Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 5554–5567. https://doi.org/10.18653/v1/2022.acl-long.380

[Shen et al.(2017)] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems* 30 (2017).

[Straka and Straková(2017)] Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*. 88–99.

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

[Upadhyay et al.(2022)] Bhargav Upadhyay, Akhilesh Sudhakar, and Arjun Maheswaran. 2022. Efficient Reinforcement Learning for Unsupervised Controlled Text Generation. *arXiv preprint arXiv:2204.07696* (2022).

[Van Gompel and Pickering(2007)] Roger PG Van Gompel and Martin J Pickering. 2007. Syntactic parsing. *The Oxford handbook of psycholinguistics* (2007), 289–307.

[Van Noord et al.(2015)] Nanne Van Noord, Ella Hendriks, and Eric Postma. 2015. Toward discovery of the artist's style: Learning to recognize artists by their artworks. *IEEE Signal Processing Magazine* 32, 4 (2015), 46–54.

[Vinuesa et al.(2020)] Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature communications* 11, 1 (2020), 233.

[Voutilainen(2003)] Atro Voutilainen. 2003. Part-of-speech tagging. *The Oxford handbook of computational linguistics* (2003), 219–232.

[Walton(2005)] Douglas Walton. 2005. Deceptive arguments containing persuasive language and persuasive definitions. *Argumentation* 19, 2 (2005), 159–186.

[Wang et al.(2019)] Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality 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)*. 3573–3578.

[Wang et al.(2020)] Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2020. Formality style transfer with shared latent space. In *Proceedings of the 28th International Conference on Computational Linguistics*. 2236–2249.

[Wegmann and Nguyen(2021)] Anna Wegmann and Dong Nguyen. 2021. Does It Capture STEL? A Modular, Similarity-based Linguistic Style Evaluation Framework. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7109–7130. https://doi.org/10.18653/v1/2021.emnlp-main.569

[Wu et al.(2020)] Yu Wu, Yunli Wang, and Shujie Liu. 2020. A dataset for low-resource stylized sequence-to-sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9290–9297.

[Xu et al.(2018)] 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* (2018).

[Xu et al.(2012)] Wei Xu, Alan Ritter, William B Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*. 2899–2914.

[Zeiler(2012)] Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* (2012).

[Zeng et al.(2021)] Chengkun Zeng, Guanyi Chen, Chenghua Lin, Ruizhe Li, and Zhi Chen. 2021. Affective Decoding for Empathetic Response Generation. In *Proceedings of the 14th International Conference on Natural Language Generation*. Association for Computational Linguistics, Aberdeen, Scotland, UK, 331–340. https://aclanthology.org/2021.inlg-1.37

[Zhang et al.(2019)] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[Zhang et al.(2018)] 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* (2018).

[Zhao et al.(2018)] Junbo Zhao, Yoon Kim, Kelly Zhang, Alexander Rush, and Yann LeCun. 2018. Adversarially regularized autoencoders. In *International conference on machine learning*. PMLR, 5902–5911.

[Zheng et al.(2022)] Yinhe Zheng, Guanyi Chen, Xin Liu, and Jian Sun. 2022. MMChat: Multi-Modal Chat Dataset on Social Media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 5778–5786. https://aclanthology.org/2022.lrec-1.621