FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Sourabrata Mukherjee

Text Style Transfer with Neural Language Models

Institute of Formal and Applied Linguistics

Supervisor: Mgr. et Mgr. Ondřej Dušek, Ph.D.

Study Program: Computational Linguistics

Prague 2025

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

Prague, 19. 2. 2025 Sourabrata Mukherjee

Title: Text Style Transfer with Neural Language Models

Author: Sourabrata Mukherjee

Department: Institute of Formal and Applied Linguistics

Supervisor: Mgr. et Mgr. Ondřej Dušek, Ph.D.,
Institute of Formal and Applied Linguistics

Abstract:

Text Style Transfer (TST) aims to modify the stylistic attributes of text while preserving its style-independent content. Despite recent advances, TST faces several challenges: data scarcity, limited multilingual coverage, inconsistent evaluation methods, and under-explored real-world TST applications. This thesis addresses these challenges through five main contributions. First, we develop approaches for effective TST using neural language models in low-resource settings, demonstrating success with both non-parallel and minimal parallel data. Second, we expand TST research into multilingual settings by creating datasets in multiple Indian languages and establishing benchmarks for two popular TST subtasks–sentiment transfer and text detoxification. Third, we introduce improved evaluation metrics that show better correlation with human judgments, including an investigation of large language models as evaluation tools. Fourth, we demonstrate TST's practical utility by developing a polite chatbot that leverages politeness transfer to generate contextually appropriate responses. Finally, we provide a comprehensive analysis of large language models' capabilities in TST tasks across multiple languages, identifying both their potential and limitations. Our findings advance the field by addressing fundamental challenges in TST while establishing new methodologies, evaluation measures, and resources for future research. Future work should focus on culturally-aware multilingual TST, multi-attribute style transfer, and robust evaluation metrics that better align with human judgment in real-world applications.

Keywords: text style transfer (TST), natural language generation (NLG), natural language processing (NLP), low-resource NLG, multilingual NLG, sentiment transfer, politeness transfer, detoxification, large language models (LLMs)

Název práce: Přenos Stylu Textu pomocí Neuronových Jazykových Modelů

Autor: Sourabrata Mukherjee

Pracoviště: Ústav formální a aplikované lingvistiky

Vedoucí práce: Mgr. et Mgr. Ondřej Dušek, Ph.D.,
Ústav formální a aplikované lingvistiky

Abstrakt:

Transfer stylu textu (TST) si klade za cíl změnit stylistické vlastnosti textu při zachování obsahu nezávislého na stylu. I přes nedávný pokrok čelí TST několika výzvám: nedostatku dat, omezenému pokrytí jazyků, nejednotným evaluačním metodám a nedostatečně prozkoumaným praktickým aplikacím. Tato práce se těmto výzvám věnuje prostřednictvím pěti hlavních příspěvků. Nejprve vyvíjíme efektivní metody pro TST s využitím neuronových jazykových modelů v prostředí s omezenými zdroji, kde prokazujeme úspěšnost jak s neparalelními, tak s minimálními paralelními daty. Dále rozšiřujeme výzkum TST do vícejazyčného prostředí vytvořením datových sad v několika indických jazycích a stanovením benchmarku pro dvě populární podúlohy TST – transfer sentimentu a detoxifikaci textu. Poté zavádíme vylepšené evaluační metriky, které vykazují lepší korelaci s lidským hodnocením, přičemž zkoumáme využití velkých jazykových modelů jako evaluačních nástrojů. Praktickou využitelnost TST demonstrujeme vyvinutím zdvořilého chatbota, který využívá transfer zdvořilosti pro generování kontextově vhodných odpovědí. Na závěr podáváme komplexní analýzu schopností velkých jazykových modelů v úlohách TST napříč více jazyky, přičemž identifikujeme jejich potenciál i omezení. Naše zjištění posouvají obor kupředu tím, že řeší základní výzvy v TST a zároveň stanovují nové metodologie, způsoby hodnocení a zdroje pro budoucí výzkum. Budoucí práce by se měla zaměřit na vývoj kulturně uvědomělého TST pro vícejazyčné prostředí, transfer, který současně zpracovává více stylistických aspektů, a na návrh robustních evaluačních metrik, které budou lépe odpovídat hodnocení rodilých mluvčích a efektivně posoudí výkon TST v reálném použití.

Klíčová slova: transfer stylu textu (TST), generování přirozeného jazyka (NLG), zpracování přirozeného jazyka (NLP), NLG s omezenými zdroji, vícejazyčné NLG, transfer sentimentu, transfer zdvořilosti, detoxifikace, velké jazykové modely (LLM)

Acknowledgements

What a journey this has been! Rather than following a predetermined path, each stage unfolded naturally, leading me to the next. My love for coding began during my engineering studies, thanks to the encouragement of my grandfather (Choto Dadu), where the fundamentals of computer science and software development captured my imagination. I have deep respect for Prof. SKP, whose guidance into the coding world further inspired me to contribute to human advancement through technology. While working as a software developer, my introduction to machine learning and artificial intelligence revealed a rapidly evolving field that demanded deeper understanding.

I am deeply grateful to my friend Jibesh, who encouraged me to pursue a PhD and guided me through the application process. My research journey began at TU Darmstadt, where I worked on Natural Language Processing projects. I thank Prof. Iryna Gurevych for the opportunity to work at TU Darmstadt. I am particularly grateful to my colleague Stanley Simoes, who not only motivated me but also introduced me to the opportunity at the Institute of Formal and Applied Linguistics (ÚFAL) at Charles University.

I cannot express enough gratitude to my PhD supervisor, Ondřej Dušek. I could not have wished for a better advisor. From day one, I was in awe of his extraordinary ability to understand and guide - whether we were discussing a nascent idea, diving into technical depths, or exploring unclear concepts, he always grasped the essence of what I was trying to convey. This remarkable understanding created an environment where our discussions flowed naturally and comfortably. Starting my research journey almost from zero, this supportive atmosphere was invaluable in building my confidence as a researcher. Thank you for guiding me how to think and write properly, for our insightful research discussions, for your unwavering support during paper deadlines, and for being there from the very beginning through the completion of this thesis. Your calm demeanor and patient guidance have not only shaped my academic journey but also taught me valuable lessons about mentorship. Without your guidance, I would not be where I am today. I will always be deeply respectful and grateful for the research environment you created, which allowed me to grow from a beginner into a confident researcher.

My six years at ÚFAL have been enriched by wonderful colleagues. I particularly want to thank Vojta and Zdeněk, and Mateusz - for your help and support. ÚFAL brought me more than just academic growth. I am especially thankful to Atul Kr Ojha, who became my mentor and steadfast supporter through both research and personal challenges. Our countless discussions have shaped me as a researcher. My heartfelt thanks go to everyone at ÚFAL who contributed to this journey.

During EMNLP 2023 in Singapore, I had the privilege of meeting Professor Monojit Choudhury, which led to a research internship at MBZUAI, Abu Dhabi, UAE. Thank you for your mentorship and for demonstrating how linguistically nuanced research can be interesting. This experience has fundamentally shaped my understanding of impactful research.

Life outside research has been equally important. I'm thankful to my friends Sunit, Aman, and others for the friendly conversations, sports, and music sessions that made being away from home enjoyable.

With profound emotion, I dedicate this thesis to my late father (Baba). Though you are no longer with us, I know nothing would have made you happier than seeing me complete this doctoral journey. I promise to continue making you proud in everything I do. To my mother (Maa), you are my world when I close my eyes—your boundless love and unwavering support light up every corner of my heart and guide me through life's every challenge. To my Dada, everything I am and will be is because of you. I want the world to know that all my achievements - past, present, and future - are possible because of your unwavering support, along with the loving encouragement of my parents.

This acknowledgment would be incomplete without mentioning my wife. You've been there through every success and failure, becoming an essential part of my life that I cannot imagine being without. A special thank you to my beloved son, my dear nephew, my sister-in-law, and all my extended family members who have enriched my life with their love and encouragement.

My school days were shaped by the unwavering support of my teachers, tutors, and seniors, whose guidance laid the foundation for my academic journey. Studying at RKM instilled in me discipline, curiosity, and a deep love for learning. The encouragement from fellow students, the nurturing environment of my home, and the wisdom of Maharaj Jis all played a crucial role in shaping who I am today. I am deeply grateful to everyone who cared for me and helped me grow, both as a student and as a person.

From my childhood until now, I sincerely acknowledge all the help, support, guidance, and motivation I have received from my friends, beloved ones, and family members - you mean everything to me. Without your help, these achievements would have been impossible. Please keep me in your prayers as I strive to make you proud through dedicated and honest work.

The depth of gratitude I feel cannot be fully captured in these pages. I have received so much from so many that these acknowledgments feel insufficient. I will carry you all in my heart as I move forward on this journey.

The work has been supported by the European Union (ERC, NG-NLG, 101039303) and by the Charles University projects GAUK 392221 and SVV 260575. This work has been using language resources and tools developed and/or stored and/or distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

Contents

1 Introduction

One of the grand goals of Artificial Intelligence (AI) is to simulate human cognition (Turing, 1950), and a crucial aspect of this pursuit is modeling the way humans communicate (Chomsky, 1965). Language is a uniquely human tool that allows for rich, nuanced expression, facilitating the exchange of ideas, emotions, and information (Hauser et al., 2002). Natural Language Generation (NLG), as a core field within AI, aims to produce human-like language from structured or unstructured data inputs (Gatt and Krahmer, 2018). Traditionally, NLG has focused on transforming raw data into coherent, meaningful text that is both grammatically accurate and semantically rich (Gatt and Krahmer, 2018).

As AI models become increasingly sophisticated, the demand for text that resonates on a more personal or context-sensitive level is also rising, leading to the emergence of controllable NLG techniques (Keskar et al., 2019). Controllable NLG seeks to move beyond simply generating text that “makes sense” to generating text that additionally aligns with particular stylistic or contextual goals (Len et al., 2020). This control might target aspects like tone, formality, sentiment, or politeness, making it possible for NLG systems to adapt to diverse social and situational needs (Li et al., 2018; Fu et al., 2018). This branch of NLG, known as style-controlled text generation, lays the foundation for Text Style Transfer (TST), a specific task that focuses on altering the style of an existing text while preserving its core content (Prabhumoye et al., 2018). TST aims to modify aspects such as sentiment, tone, or level of formality in a given text, transforming its style to meet a desired target (Prabhumoye et al., 2020). The ability to retain semantic content while altering stylistic elements is at the heart of TST’s appeal, making it a powerful tool in scenarios where tone, appropriateness, or personal expression is vital (Jin et al., 2022). For instance, consider the sentence "This restaurant's food is absolutely terrible" being transformed to "I regret to inform you that this establishment's cuisine failed to meet expectations" – while the core meaning (negative dining experience) remains intact, the style shifts from informal and direct to formal and polite.

The field of TST traces its origin to earlier explorations in NLG, particularly in machine translation, summarization and dialogue systems, where generating responses that fit a particular style has been a longstanding challenge (Shang et al., 2015; Weston, 2015). Early attempts at TST involved rule-based systems or handcrafted linguistic patterns, which were effective in controlled settings but lacked the flexibility needed for broader applications (Hovy, 1987; Sheikha and Inkpen, 2011; Xu et al., 2012). In recent years, with the advent of deep learning and transformer-based pretrained models, TST has gained new dimensions (Jin et al., 2022).

The rapid development of TST has opened up a range of applications across industries (Hu et al., 2022). For instance, in customer service, TST can help generate responses that adapt to different emotional tones, from apologetic to formal, enhancing customer experience. In social media, TST can make posts more engaging or diplomatic, while in educational settings, it can tailor explanations to match the learning style of students (Hu et al., 2022; Mou and Vechtomova, 2020). Furthermore, TST plays an essential role in creating more inclusive and respectful language by allowing for the removal of biases or offensive terms (Chen et al., 2018).

Despite significant advancements in TST, several challenges remain. This thesis aims to address these challenges by exploring various neural architectures and control mechanisms that enhance the effectiveness of TST. In Section 1.1, we discuss these key challenges in the field. In Section 1.2, we outline the main contributions of this thesis. Finally, in Section 1.3, we provide a chapter-wise overview of the thesis, summarizing the key aspects covered in each chapter.

1.1 Challenges

TST requires generating text that meets stylistic constraints while retaining the original meaning, which often involves balancing multiple, sometimes conflicting, objectives. Addressing these challenges reveals key research questions central to the TST field. These questions guide the research presented in this thesis:

RQ1 **Data Scarcity: How can TST be achieved effectively with limited parallel or non-parallel data?** Neural language models like transformers, though powerful, sometimes struggle to capture subtle stylistic nuances without compromising content, especially in low-resource settings (Tikhonov et al., 2019).

The scarcity of parallel TST datasets intensifies this challenge, as TST often must be performed using either limited parallel data (pairs of sentences with same content in different styles) or non-parallel data (unpaired sentences in different styles) and unsupervised techniques (Jin et al., 2022).

RQ2 **Multilingual TST: How can TST research expand effectively into multilingual settings?** Most TST research focuses on English, with limited exploration in other languages (Hu et al., 2022). Given that stylistic conventions differ significantly across languages, developing cross-linguistic TST methods remains an open challenge (Krishna et al., 2020).

RQ3 **TST Evaluation: How can TST be evaluated in a reliable way?** TST evaluation lacks consistency, as human evaluations are costly and automated metrics often fail to balance style accuracy with content preservation (Mir et al., 2019; Belz et al., 2020). Establishing reliable and reproducible evaluation methods is crucial for advancing the field.

RQ4 **TST Applications: How can TST contribute to building effective downstream applications?** TST has potential applications in dialogue systems, content creation, and personalized communication (Hu et al., 2022), yet application-driven TST development is still underexplored.

RQ5 **TST using LLMs: How do large language models (LLMs) perform on various TST tasks, including multilingual and subtask-specific challenges?** With the rise of LLMs, it is critical to assess their effectiveness in diverse TST tasks and languages, as well as their potential to address existing limitations (Chowdhery et al., 2023; Bommasani et al., 2021).

1.2 Main Contributions

The main contributions of this thesis, addressing the research questions defined above, are as follows:

Ad RQ1 We demonstrate that TST can be effectively achieved using neural language models (LMs) without requiring direct supervision or large parallel datasets, while also being feasible with minimal parallel data (as discussed in Chapter 3). In Section 3.1, we present our approach, which leverages a polarity-aware denoising model structure with two key stages: (1) a pre-trained shared encoder for robust representation learning, and (2) sentiment-controlled generation via separate sentiment-specific decoders.

Furthermore, in Section 3.2, we explore TST in low-resource settings, where we employ multiple adaptation techniques, including data augmentation and self-training along with a novel style reward mechanism, to enhance performance under parallel data scarcity.

Ad RQ2 To address the multilingual gap in TST research, we developed new resources and methodologies (as discussed in Chapter 4). We created a sentiment transfer dataset in nine Indian languages (as discussed in Section 4.1) and a detoxified dataset for both English and Hindi (as discussed in Section 4.2). Our experiments benchmark models in challenging scenarios, including cases where no style-parallel data is available or where training relies solely on machine-translated data from English instead of human-annotated data. Additionally, we conduct a comprehensive multilingual and cross-linguistic analysis, offering insights into effective approaches for multilingual TST.

Ad RQ3 We address the challenge of TST evaluation by systematically analyzing existing text metrics from wider NLP tasks and novel evaluation metrics (as discussed in Chapter 5). To ensure reliability, we conduct a meta-evaluation of the proposed metrics by measuring their correlation with human judgments. Furthermore, we investigate the applicability of LLMs as an evaluation tool. Our findings show that our approaches improve alignment with human judgments over the existing TST evaluation measures.

Ad RQ4 We developed a TST-based application: a polite chatbot (as discussed in Chapter 6). This system leverages a politeness transfer model to generate polite synthetic dialogue pairs, which are used to train the dialogue model, resulting in a chatbot capable of responding in a polite manner across diverse conversational contexts.

Ad RQ5 We conducted a systematic evaluation of LLMs on TST tasks, focusing on sentiment transfer and text detoxification in English, Hindi, and Bengali (as discussed in Chapter 7). Our analysis provides detailed insights into the strengths and limitations of LLMs for these TST tasks, identifying areas for future improvement.

1.3 Thesis Overview

The rest of this thesis is organized into the background chapter (2), the content chapters (3 to 7), and the concluding chapter (8).

Chapter 2 explains the foundation for this research by introducing essential concepts. The subsequent content chapters systematically address the research questions outlined in Section 1.1 and present the contributions discussed in Section 1.2. Chapter 3 explores data scarcity challenges, proposing approaches for TST using non-parallel and minimal parallel data. Chapter 4 extends TST research beyond English to a multilingual setting, while Chapter 5 focuses on evaluation, analyzing existing and proposed metrics and their correlation with human judgments. Chapter 6 introduces a real-world application of TST through a polite chatbot system, and Chapter 7 assesses the capabilities of LLMs in performing TST. Finally, Chapter 8 summarizes the key contributions and overall findings of the thesis.

This thesis builds on ten of the author's publications – eight peer-reviewed conference papers and two preprints. Each chapter or section includes references to the corresponding papers, where the author of thesis is the first author in all cases.

2 Background

The field of Text Style Transfer (TST) has gained significant traction in recent years, driven by advancements in neural language models and their ability to manipulate text attributes while preserving semantic content (Jin et al., 2022). This chapter provides a basic understanding of the technologies underlying TST, focusing on two key areas: the architecture and functionality of neural language models (Section 2.1), and the fundamentals of TST itself (Section 2.2).

In the first section (Section 2.1), we explore the evolution of neural networks with essential concepts such as text representation, Transformer architecture, and pretrained and large language models (LLMs). The second section (Section 2.2) shifts focus to TST. It begins by defining text style and clarifying the distinction between style and content. We then discuss the challenges inherent in TST and examine existing datasets, approaches, and evaluation measures. We also highlight various applications that demonstrate TST's practical relevance in real-world scenarios.

2.1 Neural Language Models

In this section, we explore neural language models, which are central to TST. Neural Language Models are machine learning systems that learn to predict and generate human-like text by modeling the probability distribution of word sequences through neural network architectures (Bengio et al., 2003; Mikolov et al., 2010). We start by discussing the basic concepts of neural networks (Section 2.1.1), and the associated methods for representing text (Section 2.1.2). We then present a comprehensive exploration of the Transformer architecture (Section 2.1.3), leading to pretrained (Section 2.1.4) and LLMs (Section 2.1.5).

2.1.1 Neural Networks

Neural networks emerged from early attempts to mathematically model biological neural systems, they consist of layers of interconnected computational units, or *neurons*, that work collaboratively to process inputs, extract features, and generate outputs (McCulloch and Pitts, 1943). By adjusting the weights of these connections through training, neural networks are capable of learning complex, non-linear mappings between input and output spaces (Goodfellow et al., 2016).

Perceptron The perceptron is a foundational model of a neuron, which computes a weighted sum of inputs, applies a bias, and passes the result through an *activation function* (Mehrotra et al., 1997). The perceptron learns by iteratively adjusting weights and biases based on the error between the predicted and actual output for each training example. Neural networks arrange these perceptrons in interconnected layers, enabling them to learn complex patterns.

Multi-Layer Perceptron (MLP) Building on the perceptron model, multi-layer perceptrons (MLPs) introduce multiple layers of fully connected perceptrons, which consist of an input layer, one or more hidden layers, and an output layer, enabling the learning of complex, non-linear patterns. Each hidden layer allows the network to learn non- linear transformations of the input data. It uses non-linear activation functions (such as the ReLU (Nair and Hinton, 2010)), to transform inputs into richer representations, making MLPs suitable for a wide range of tasks. The final layer in an MLP produces the output, which can represent class probabilities in classification tasks or predicted values in regression tasks.

Limitations of MLP Despite their success in many machine learning tasks, MLPs have limitations when processing text data. MLPs require fixed-length inputs (Elman, 1990), making them poorly suited for handling variable-length sequences like sentences and paragraphs. Additionally, MLPs process each input independently, unable to capture the sequential relationships between words that are crucial for understanding language. These limitations motivated the development of Recurrent Neural Networks (RNNs), which are specifically designed to handle sequential data by maintaining an internal memory of previous inputs (Mikolov et al., 2010).

Recurrent Neural Networks (RNNs) Recurrent Neural Networks (RNNs) enable neural networks to handle sequential data, as a solution to the limitations of MLPs by introducing a simple yet powerful modification: connections that loop back in time (Grossberg, 2013). Unlike MLPs, RNNs maintain an internal memory state that updates with each input token in a sequence, allowing them to capture dependencies between words in natural language. The earliest successful RNNs generally use sigmoid activation functions to process the combination of current input and previous state (Narayan, 1997), enabling them to learn patterns in sequential data like text. This architecture proved particularly effective for language modeling, where the network learns to predict the next word based on all previous words in a sequence (Mikolov et al., 2010).

Limitations of RNNs RNNs theoretically could capture long-range dependencies in text, in practice they struggled with longer sequences due to optimization challenges when propagating gradients through many time steps (Bengio et al., 1994). This *vanishing gradient problem* arises because gradients can diminish as they are propagated backward through many time steps, making it difficult for the network to learn dependencies across long distances in the input (Hochreiter, 1998).

LSTM & GRU To address these issues, advanced RNN architectures like Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRUs) (Cho et al., 2014) introduce gating mechanisms to control the flow of information and maintain a stable gradient. These architectures help mitigate the vanishing gradient problem, enabling the model to capture longer dependencies in the data.

The RNNs and their advanced variants have laid the groundwork for even more powerful architectures, such as the Transformer, which we will explore in Section 2.1.3.

2.1.2 Text Representation

For neural networks to process textual data, it is essential to convert text into numerical representations. Text representation methods serve as the bridge between raw text and the input requirements of neural networks. In this section, we discuss different approaches to text representation, word embeddings, and subword tokenization.

For neural networks to process textual data, it is essential to convert text into numerical representations. Text representation methods serve as the bridge between raw text and the input requirements of neural networks. A basic approach is one-hot encoding, where each word is represented as a sparse binary vector; however, this method is highly inefficient due to its sparse representation, large dimensionality and lack of semantic relationships between words. Instead, more effective techniques such as dense word embeddings and subword tokenization provide compact and meaningful representations (Mikolov et al., 2013). In this section, we discuss these advanced approaches to text representation.

Word Embeddings Word embeddings provide a dense vector representation for each word, capturing semantic relationships between words based on their distributional properties (Bengio et al., 2003; Mikolov et al., 2013). Word embeddings map words to a continuous vector space, where words with similar meanings are closer to each other in terms of cosine or Euclidean distance, enabling more effective processing of text in neural network architectures (Pennington et al., 2014).

Subword Tokenization While word embeddings effectively capture word semantics, a fundamental challenge remains: languages contain an unbounded number of possible words, but neural models require a finite vocabulary with fixed-size embedding matrices (Sennrich et al., 2016a). To address these challenges, subword tokenization techniques were developed to segment words into smaller, reusable units, each assigned its embedding vector. These subword units strike a balance between words and characters—-short enough to handle rare or unseen words but longer than individual characters, preventing sequences from becoming excessively long. Byte Pair Encoding (BPE) emerged as one of the first widely-adopted subword tokenization methods (Sennrich et al., 2016a). BPE iteratively merges the most frequent character sequences into subword tokens until reaching a desired vocabulary size. For example, "unhappiness" might be tokenized as ['un', 'happ', 'iness'], with each subword receiving its embedding.

2.1.3 Transformer Architecture

The Transformer architecture, introduced by Vaswani et al. (2017), represents a major advancement in neural network design for natural language processing (NLP). Unlike earlier models like basic Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks (discussed in Section 2.1.1), Transformers process input sequences entirely in parallel using self-attention mechanisms, overcoming the sequential limitations of RNNs.

High-Level Architecture As illustrated in Figure 2.1, the Transformer model is composed of stacked encoder and decoder blocks, which process input sequences through a combination of self-attention mechanisms and feed-forward layers (see Figure 2.1). Each block follows a consistent structure–the input first passes through a multi-head attention layer which allows the model to capture dependencies across different positions in the sequence, followed by layer normalization. A parallel residual connection ensures stable learning. The output of this step then flows through a feed-forward network that refines the learned representations, again followed by layer normalization and bypassed with another residual connection (detailed mechanisms are described below).

Input and Positional Embeddings The Transformer starts with input embeddings, where each token is represented as a vector to capture semantic properties of the tokens (as discussed in Section 2.1.1). Since the model processes all tokens in parallel, position embeddings are added to the token embeddings to provide sequence order information. While the foundational Transformer used sinusoidal position encodings (Vaswani et al., 2017), later models opted for learned positional embeddings (Kenton and Toutanova, 2019).

Self-Attention Mechanism At the heart of the Transformer is the *self-attention* mechanism, which allows each token in a sequence to attend to all other tokens, capturing contextual dependencies regardless of their distance in the sequence (Vaswani et al., 2017). By replacing sequential processing with parallel attention, self-attention eliminates the inefficiencies of RNNs. The *multi-head attention* mechanism extends self-attention by enabling the model to attend to multiple aspects of the input simultaneously. Each attention head learns a distinct representation, and the outputs of all heads are combined, enhancing the model's capacity to capture diverse relationships between tokens.

Feed-Forward Networks The Transformer processes these attention-weighted representations through feed-forward networks, which apply the same transformation to each position independently.

Residual Connections and Layer Normalization To improve gradient flow during training, Transformers use *residual connections*, where the input to each layer is added to its output, to facilitate gradient flow by providing direct paths between layers. This prevents vanishing gradients and accelerates convergence (He et al., 2016). Additionally, *layer normalization* (Ba, 2016) is applied to stabilize and normalize the activations, further improving training stability and model performance.

Transformer Encoder and Decoder The original Transformer consists of an *encoder-decoder* structure. The encoder in a Transformer consists of a stack of identical layers, each comprising multi-head self-attention and a feed-forward network. The encoder takes the input sequence and generates contextualized representations for each token. The decoder, also built with a stack of identical layers, differs in two ways: (1) it includes *masked self-attention*, which prevents attending to future tokens during autoregressive generation, and (2) it incorporates *encoder-decoder attention*, allowing it to attend to the encoder's output while generating the target sequence.

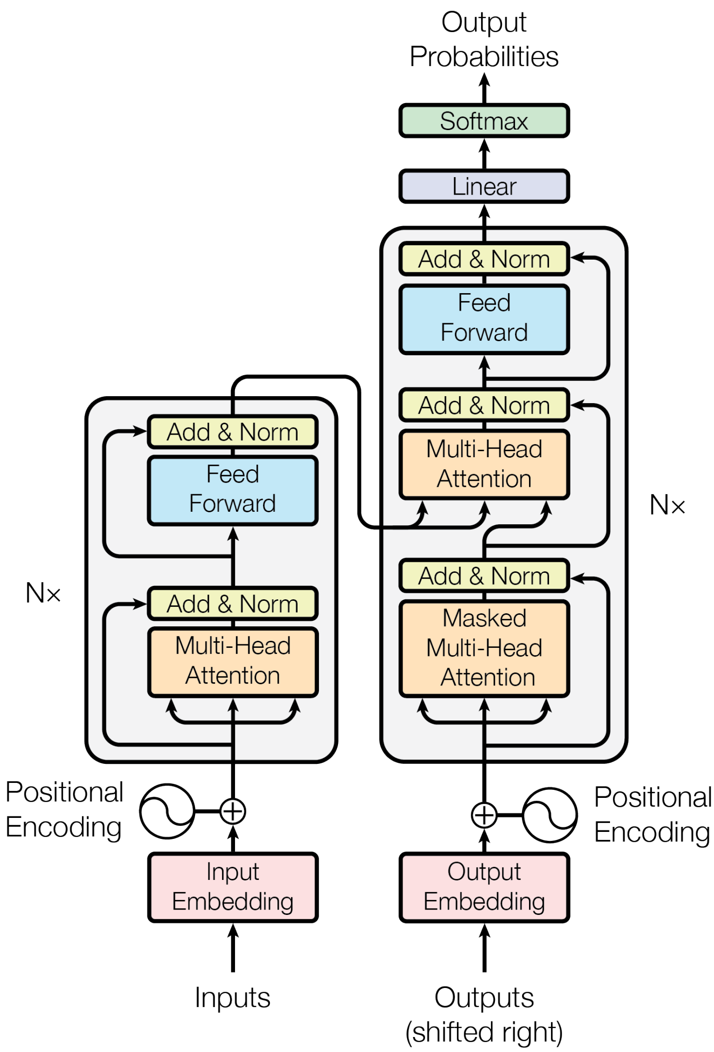


Figure 2.1: Illustration of the Transformer architecture, including the encoder and decoder stacks, with self-attention, encoder-decoder attention, feed-forward networks, and positional encodings. Adapted from Vaswani et al. (2017).

2.1.4 Pretrained Language Models

Building upon the Transformer architecture, Pretrained Language Models (PLMs) advanced NLP by learning language representations from massive text corpora (Kenton and Toutanova, 2019). Rather than training models from scratch for each task, PLMs first undergo a two-stage process: pretraining and fine-tuning. During pretraining, they learn general language understanding through self-supervised learning. This involves predicting missing words in a sentence (masked language modeling–e.g., predicting 'dog' in 'The [MASK] barked') or predicting the next word in a sequence (causal language modeling–e.g., completing 'The dog barked at the…') using large amounts of unlabeled text. After pre-training, these models can be fine-tuned on specific downstream tasks using much smaller amounts of labeled data, leveraging their learned language understanding (Kenton and Toutanova, 2019; Radford et al., 2018).

These models can be broadly categorized based on their architectural choices. Encoder-only models like BERT use bidirectional self-attention to process input text simultaneously in both directions, focusing on understanding tasks by predicting masked tokens using context from both sides (Kenton and Toutanova, 2019). Decoder-only models like GPT employ unidirectional (causal) attention where each token can only attend to previous tokens, making them natural for text generation as they learn to predict what comes next in a sequence (Radford et al., 2018). Encoder-decoder models like BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) combine both approaches: an encoder that processes the full input bidirectionally for comprehensive understanding, followed by a decoder that generates output tokens sequentially while attending to both the encoded input and its own previous predictions, making them particularly effective for text transformation tasks that require both understanding and generation.

2.1.5 Large Language Models (LLMs)

Building upon the success of PLMs, LLMs have pushed the boundaries of neural language modeling through massive scaling in both model size and training data (Zhao et al., 2023). Most LLMs adopt the decoder-only Transformer architecture, trained with causal language modeling, though some use encoder-decoder architectures for specific tasks (as discussed in Section 2.1.4). Research has shown that model performance improves predictably with increased parameters and dataset size, following empirical scaling laws (Kaplan et al., 2020; Chowdhery et al., 2023). A key innovation of LLMs is their ability to perform in-context learning, where they can adapt to new tasks through examples or instructions provided in the prompt, without requiring fine-tuning (Brown et al., 2020). This capability has been further enhanced through instruction tuning, where models are trained to follow natural language instructions for diverse tasks (Sanh et al., 2022).

LLMs have evolved significantly, with both closed-source models (like GPT-4 (Achiam et al., 2023)) and open-source alternatives (like Llama (Touvron et al., 2023a)) demonstrating strong capabilities in language understanding and generation. A key advancement has been the introduction of human feedback techniques: Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022a) and Direct Preference Optimization (DPO) (Rafailov et al., 2023) have significantly improved models' ability to follow instructions and generate more helpful, accurate responses. While closed-source models often lead in performance, open-source LLMs have made these capabilities more accessible to researchers and developers (Hanke et al., 2024).

2.2 Text Style Transfer (TST)

This section is based on the papers *Text style transfer: An introductory overview* (Mukherjee and Dušek, 2024), accepted in the Accepted at 4EU+ International Workshop on Recent Advancements in Artificial Intelligence and *A survey of text style transfer: Applications and ethical implications* (Mukherjee et al., 2024a).

The main goal of NLG is to automatically produce texts that describe, summarize and explain input data or text in a human-like manner. Some of the popular tasks of NLG include data-to-text generation, response generation in dialogue systems, paraphrase generation, and text summarization (Gatt and Krahmer, 2018).

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

TST, a subtask of controllable text generation, is a method that aims to control certain attributes of text while preserving its content as much as possible (examples depicted in Figure 2.2). Style attributes of a text can range from demographic attributes of a person writing the text such as personality (Li et al., 2016), gender (Prabhumoye et al., 2018) to sentiment (Sudhakar et al., 2019), emotion (Zhou et al., 2018), author-style (Tikhonov and Yamshchikov, 2018) or politeness (Niu and Bansal, 2018).

TST has gained significant attention thanks to the rise of deep neural models (Jin et al., 2022). However, the *TST* task still requires deeper attention to the following challenges. First, disentangling content and style in texts has proven to be very hard. Second, TST models can be developed in a supervised way with parallel corpora, i.e. text that comes in pairs with the same content but with different styles (Hu et al., 2022). However, most use cases do not have parallel data available, so TST on non-parallel corpora has become a prolific research area (Hu et al., 2022). Third, research in the area of style transfer for text is currently bottlenecked by a lack of standard evaluation practices (Mir et al., 2019). Effective evaluation of TST requires assessing multiple aspects of the generated output, primarily focusing on *style transfer accuracy*, *content preservation*, and *fluency*. An ideal TST output should successfully adopt the target style, retain the original content's meaning, and be coherent and fluent (Ostheimer et al., 2023).

In this section, we lay out the background for our work, including the distinction between style and content (Section 2.2.1), existing datasets (Section 2.2.2), existing *TST* approaches (Section 2.2.3), evaluation measures–automatic (Section 2.2.4) and human (Section 2.2.5), and applications (Section 2.2.6).

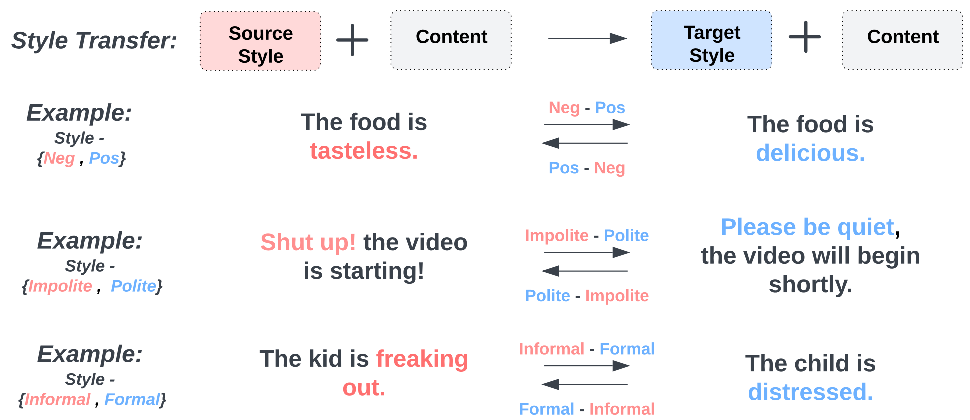


Figure 2.2: Examples of TST: transferring sentiment, politeness, and formality.

2.2.1 Style vs. Content Distinction

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

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

For example, we can consider the *Negative* ⟶ *Positive* instance from Figure 2.2. Here, x is "The food is tasteless" and s_1 is *negative*. The *style-independent content* in x is "The food is". After sentiment transfer, $\hat{x}$ is "The food is delicious" and s_2 is *positive*.

2.2.2 Datasets & Subtasks

Over the years, various datasets have been developed to train and evaluate *TST* models. These datasets support a range of TST subtasks, each focusing on specific style attributes. We discuss some prominent datasets for these subtasks, categorizing them based on their task and characteristics.

- **Politeness Transfer** – The politeness transfer task aims to alter the politeness of text without changing its content. A dataset compiled from the Enron email corpus, which contains instances automatically labeled for politeness, was proposed by Madaan et al. (2020) to study this transformation.

- **Sentiment Transfer** – Sentiment transfer focuses on changing the polarity of a sentence (e.g., from positive to negative) while preserving the content. This task has been explored using popular datasets such as Yelp (Shen et al., 2017), Amazon (He and McAuley, 2016), and IMDb (Dai et al., 2019). Each dataset contains binary sentiment labels (positive or negative), with reviews categorized based on ratings. For instance, in the Yelp dataset, reviews with a rating above 3 are labeled as positive, and those below 3 as negative, with neutral reviews excluded from the dataset.

- **Formality Transfer** – Formality transfer involves converting text between formal and informal styles. Grammarly's Yahoo Answers Formality Corpus (GYAFC) (Rao and Tetreault, 2018) is the largest human-labeled dataset developed for this task. It includes informal sentences from the Yahoo Answers corpus in domains such as Entertainment and Music (E&M) and Family and Relationships (F&R), with each sentence manually rewritten in a formal style. The dataset has been extensively used to evaluate models on tasks requiring nuanced understanding of language formality.

- **Author Imitation** – Author imitation seeks to modify a sentence's style to reflect a particular author. A notable dataset in this category is the Shakespeare-Modern corpus by Xu et al. (2012), which includes line-by-line modern interpretations of 16 of Shakespeare's plays (e.g., *Hamlet*, *Macbeth*) sourced from Sparknotes. This parallel dataset enables models to generate Shakespearean-style text, facilitating exploration in author-specific style transfer. Despite its limited scope, the Shakespeare-Modern dataset is a unique resource in TST, with potential future applications in broader author style imitation tasks.

- **Genre Transfer** – Genre transfer involves changing the stylistic genre of a sentence. Li et al. (2018) introduced a dataset of image captions labeled with genres such as *factual*, *romantic*, and *humorous*. This non-parallel dataset has been valuable for studying models that can shift between different narrative tones and genres, allowing for creative applications in TST.

- **Gender Style Transfer** – Gender-based style transfer aims to modify text to reflect a male or female writing style. Prabhumoye et al. (2018) constructed a gender-labeled dataset using Yelp reviews with gender annotations from prior studies (Reddy and Knight, 2016). Sentences deemed gender-neutral were excluded, resulting in a non-parallel dataset for evaluating gender-related linguistic styles.

- **Political Slant Transfer** – This task seeks to adjust the political stance of a text. Prabhumoye et al. (2018) curated a dataset with Facebook comments from U.S. Congress members' public pages, labeled based on political affiliation (Democrat or Republican). The dataset has been used to examine how language can reflect political perspectives in TST models.

- **Offensive Language Correction** – This task addresses the transformation of offensive language into non-offensive expressions. Nogueira dos Santos et al. (2018) created a dataset from Twitter and Reddit posts, classified as offensive or non-offensive, to support models in moderating or sanitizing online discourse.

A few of these datasets, such as GYAFC and Shakespeare-Modern, are parallel, meaning they provide paired examples in different styles for the same content. Parallel datasets are particularly useful for supervised TST tasks, as they allow models to directly learn mappings between styles. However, most TST datasets, such as Yelp and Amazon for sentiment transfer, are non-parallel, requiring models to perform style transfer without direct supervision.

2.2.3 Overview of Approaches

The TST approaches can be broadly classified into four categories which we discuss in the following paragraphs: parallel supervised, non-parallel supervised, purely unsupervised, and using LLMs.

Parallel Supervised Approaches Parallel supervised TST models are trained using pairs of texts with similar content but different styles. The most common underlying architecture of these models is the sequence-to-sequence architecture, same as for for many other natural language generation tasks (Gatt and Krahmer, 2018). Usually, a sequence-to-sequence model is trained on a parallel corpus wherein the text of the original style is fed into the encoder and the decoder outputs the corresponding text according to the target style. For example, Jhamtani et al. (2017) used this approach to translate between modern and Shakespearean English (where ample training data exists).

Applying the sequence-to-sequence approach to this task is quite challenging due to the unavailability of parallel data (Hu et al., 2022). To handle this, data augmentation methods, such as pseudo-parallel datasets, have been explored (Shang et al., 2019; Jin et al., 2019; Nikolov and Hahnloser, 2018; Liao et al., 2018).

Non-parallel Supervised Approaches An alternative method to using parallel datasets is the non-parallel supervised setting. In this setting, the TST models aim to transfer the style of texts without any knowledge of matching text pairs in different styles. Data for each individual target style is provided, but not aligned with other styles.

To train the TST models in the non-parallel supervised setting, researchers have proposed the strategy of disentangling the style and content in text (Gatys et al., 2015; Yamshchikov et al., 2019). There are three main types of style-content disentanglement strategies:

(1) **Explicit Style-Content Disentanglement** TST models for *explicit style-content disentanglement* adopt a straightforward text replacement approach for generating texts of a target style. For example, Li et al. (2018) found that phrases of a text that are associated with the original style can be replaced with new phrases associated with the target style.

(2) **Implicit Style-Content Disentanglement** For *implicit style-content disentanglement*, TST models first learn the latent representations of the content and the style of the given text. The latent representation of the original content is then combined with the latent representation of the desired target style to generate text in the target style. Techniques such as back-translation and adversarial learning (Shen et al., 2017; Zhao et al., 2018; Fu et al., 2018; Prabhumoye et al., 2018; Hu et al., 2017) have been proposed to disentangle content and style in latent representations.

(3) **No Style-Content Disentanglement** In this setting, the TST models perform TST without disentangling the text's style and content. These approaches include: attribute-controllable generation, where models learn to directly generate text with desired attributes through controlled text generation (Lample et al., 2019); reinforcement learning techniques that use reward signals to guide the model toward the target style (Luo et al., 2019b); and probabilistic modeling approaches that treat style transfer as a conditional generation task, directly modeling the probability distribution of target-style text given the input (He et al., 2020b). These methods avoid the challenges of explicit disentanglement while maintaining effective style transfer through end-to-end training (Dai et al., 2019; Li et al., 2019).

Unsupervised Approaches Works on purely unsupervised TST are comparatively rarer. An early study by Radford et al. (2017) explored the properties of recurrent neural language models at the byte level. They trained an LSTM model on text processed as a sequence of UTF-8 encoded bytes in an unsupervised manner. Interestingly, they found a single neuron within the trained LSTM that directly corresponded to the sentiment. By manipulating this neuron, they were able to transfer sentiments in sentences. Xu et al. (2020) used unsupervised representation learning to separate style and content on a mixed corpus of unspecified styles. They were able to isolate a latent dimension responsible for sentiment and achieved satisfactory results in sentiment transfer. Shen et al. (2020b) computed a "sentiment vector" by averaging latent codes separately for 100 positive and 100 negative sentences in the development set, then calculating the difference between them. Given a test sentence, they changed its sentiment from positive to negative or vice versa by adding or subtracting the sentiment vector.

The performance of these purely unsupervised TST methods is encouraging, although not matching supervised methods. However, it is important to note that these methods have been evaluated primarily on sentiment transfer tasks. More research is needed to determine whether purely unsupervised methods can be generalized to other TST tasks.

Using LLMs Learning complex patterns and structures from vast amounts of text data, LLMs inherently capture various linguistic styles and nuances. This capability is particularly beneficial for TST tasks.

A distinctive feature of LLMs is their ability to perform valuable tasks without fine-tuning, showcasing zero- and few-shot capabilities (Liu et al., 2023a). Reif et al. (2021) frame style transfer as a sentence rewriting task, enhancing LLMs' zero-shot performance for arbitrary TST by using task-related exemplars. Suzgun et al. (2022) proposed a reranking method to select high-quality outputs from multiple candidates generated by the LLM, thereby improving performance. Luo et al. (2023) proposed the use of a prompt-based style classifier to guide the search for word-level edits of a given text. Additionally, Liu et al. (2024a) introduced dynamic prompt generation to guide the language model in producing text in the desired style.

While prompt engineering is a prevalent approach (Brown et al., 2020; Jiang et al., 2020), LLMs are highly sensitive to prompts (Mishra et al., 2021; Zhu et al., 2023) and may not always guarantee optimal performance (Liu et al., 2024a). Furthermore, Ouyang et al. (2022b) indicated that larger LLMs are not always inherently superior in understanding user intent.

2.2.4 Automatic Evaluation

Automatic evaluation offers a cost-effective, reproducible, and scalable means of assessing TST models. Several automated metrics have been proposed to quantify TST effectiveness based on the three criteria of style accuracy, content preservation, and fluency (Pang, 2019a,b; Pang and Gimpel, 2019; Mir et al., 2019).

Style Transfer Accuracy The ability of a model to transfer the intended style is typically measured using a *style transfer accuracy* metric (Hu et al., 2017; Shen et al., 2017; Fu et al., 2018; Luo et al., 2019b; John et al., 2019). This typically involves a pre-trained binary classifier, such as TextCNN (Moschitti et al., 2014), which predicts the style of the generated sentence. The accuracy is then calculated by comparing the classifier's predictions to the target style, providing a quantitative measure of successful style transfer.

For alternative assessments, the *Earth Mover's Distance* can be used to calculate the "distance" between the style distributions of the original and transferred sentences (Mir et al., 2019), offering insight into the intensity of the style change.

Content Preservation This aspect ensures that the original style-independent content is preserved after style modification is critical for TST. Several metrics adapted from other natural language generation (NLG) tasks are used for this purpose:

- *BLEU (Bilingual Evaluation Understudy)* (Papineni et al., 2002): When parallel TST datasets or human references are available, the BLEU score is calculated between the transferred sentence and reference sentences (Tikhonov et al., 2019). It quantifies the overlap of n-grams, serving as a proxy for content preservation.

- *Source-BLEU (sBLEU)*: In non-parallel TST settings, *source-BLEU* evaluates content retention by calculating the n-gram overlap between the transferred sentence and its original source sentence (Madaan et al., 2020). This approach assumes that a high degree of overlap suggests greater content preservation.

- *Cosine Similarity*: The cosine similarity between embeddings of the original and transferred sentences provides another means of evaluating content retention. High similarity indicates that the transformed sentence closely matches the semantic content of the original (Fu et al., 2018; Reimers and Gurevych, 2019).

- *Word Overlap*: John et al. (2019) suggested unigram word overlap, which excludes stop words and style-attributed terms, as a direct metric of content retention. This method counters potential limitations of cosine similarity by focusing on key content words only.

Fluency TST outputs should exhibit natural language fluency. A common fluency measure is *perplexity*, calculated using a pre-trained language model (Briakou et al., 2021b). Lower perplexity indicates higher alignment with natural language, suggesting that the output is syntactically coherent and contextually appropriate. The model is typically trained on a general corpus to compute the perplexity of transferred sentences, assessing their linguistic fluency. However, Pang (2019a); Mir et al. (2019) cautioned against using perplexity (PPL) as a measure of fluency, as it tends to favor awkward sentences with commonly used words.

2.2.5 Human Evaluation

Human evaluation is essential for capturing subjective aspects of style and nuance in content preservation. Human evaluation is particularly valuable due to its flexibility and comprehensive feedback; however, it is also costly, time-consuming, and can be challenging to standardize (Pang, 2019a,b; Mir et al., 2019).

Common human evaluation setups involve *point-wise* scoring or *pairwise* comparisons:

- **Point-wise Scoring** – Human evaluators rate generated sentences on a scale (e.g., 1 to 5) for each criterion (style transfer accuracy, content preservation, and fluency). For instance, a score of 5 in content preservation would imply that the transferred sentence retains the original meaning almost entirely.

- **Pairwise Comparison** – Evaluators compare two or more sentences and determine which version better satisfies a particular criterion or rank multiple outputs accordingly. This approach reduces individual bias, particularly when multiple raters are involved.

Human evaluations offer insights into aspects that automated metrics might overlook, such as nuanced language shifts or subjective style elements. However, they are prone to inconsistencies due to the subjective interpretation of style, making it essential to standardize evaluation guidelines and provide clear scoring rubrics to evaluators (Briakou et al., 2021b).

2.2.6 Applications of TST

This section organizes the diverse applications of TST into five primary types:

User Privacy and Security TST can safeguard user privacy by concealing authorship and removing offensive language or bias. For bias correction, TST methods rephrase narratives to avoid stereotypes and correct social biases (Clark et al., 2018; Sap et al., 2017; Pryzant et al., 2020). TST also helps in concealing authorship by altering writing styles to obscure identity, which is essential in protecting the privacy of individuals such as whistleblowers (Reddy and Knight, 2016; Gröndahl and Asokan, 2020; Zhai et al., 2022). Moreover, it combats offensive language by converting harmful or abusive messages into neutral forms to foster a healthier digital environment (Nogueira dos Santos et al., 2018; Logacheva et al., 2022; Dementieva et al., 2023; Tran et al., 2020).

Creating Personalized Texts TST enhances personalized communication in fields like marketing, content creation, and customer service. In marketing, it can optimize message impact by adapting style to audience profiles (Kaptein et al., 2015; Jin et al., 2020). For text simplification, TST bridges expert and layperson language, making complex information accessible to broader audiences (Saggion, 2017; Cao et al., 2020). TST also supports writing assistants by refining text for varied contexts, ensuring politeness or friendliness in communication (Jin et al., 2022; Yeh et al., 2024). Other applications include generating similes (Chakrabarty et al., 2020; Yang et al., 2023), humor (He et al., 2019; Weller et al., 2020), and empathetic content for settings like creative writing or mental health care (Sharma et al., 2021, 2023; Pérez-Rosas et al., 2023).

Use as Part of Other NLP Tasks TST integrates with various NLP tasks, such as machine translation, where it adapts the formality of translations (Sennrich et al., 2016b; Niu et al., 2017; Wu et al., 2020, 2021; Zou et al., 2024), and image captioning, where TST generates captions in stylized forms that attract more attention (Mathews et al., 2016; Gan et al., 2017; John et al., 2019; Wu et al., 2022). TST is valuable for generating chatbot responses that align with user personalities or conversational contexts, potentially improving user engagement (Kim et al., 2019; Song et al., 2019). Persona-based systems leverage TST to generate responses that consistently reflect predefined persona styles, enhancing interactions (Li et al., 2016; Su et al., 2019; Song et al., 2020). In summarization, TST enables summary generation with specific tones (Chawla et al., 2019; Cao and Wang, 2021; Goyal et al., 2022). In data-to-text generation, TST aids in crafting outputs for structured data in diverse styles (Jing et al., 2023; Tan et al., 2022). It also finds application in software engineering, transforming code styles (Munson et al., 2022; Ting et al., 2023; Chen and Abedjan, 2023).

TST for Data Augmentation Beyond style applications, TST contributes to generating diverse training data for NLP models, enhancing robustness against variability. It has been used to develop counterfactual examples that help in explaining model predictions by generating similar instances with alternative outcomes (Wang et al., 2021, 2023). For adversarial robustness, TST supports creating examples that stress-test model defenses, making it valuable in AI explainability and security (Qi et al., 2021; Pan et al., 2022; Chen and Ji, 2022).

3

Data Scarcity in TST

The scarcity of style-specific parallel data poses a significant challenge in Text Style Transfer (TST), in contrast to Machine Translation (MT) and other NLP tasks where large-scale aligned datasets for specific language pairs or tasks are readily available, such resources difficult to obtain for particular style pairs in TST (Jin et al., 2022; Hu et al., 2022). Due to this limitation, recent TST research has focused on developing alternative methodologies that can leverage non-parallel data, enabling style transfer without the need for direct supervision. Approaches such as latent-space content-style disentanglement (Hu et al., 2017; Shen et al., 2017), prototype editing (Li et al., 2018), and the creation of pseudo-parallel corpora (Zhang et al., 2018; Jin et al., 2019) have been proposed to address this data scarcity (as discussed in Section 2.2.3).

The only previous TST study known to us using parallel data with sequence-to-sequence learning, by Jhamtani et al. (2017), addresses a highly specific task: transforming modern English to Shakespearean style (as discussed in Section 2.2.3). This case benefits from the availability of aligned paraphrases for literary text, a resource that is generally unavailable in other TST tasks. Consequently, most current TST methods rely on non-parallel data and unsupervised learning (Hu et al., 2017; Zhao et al., 2018; Li et al., 2018). While promising, these approaches often incur a performance trade-off and still face challenges due to the limited availability of large, non-parallel style-specific datasets (as discussed in Section 2.2.3).

In this chapter, we present our novel methodologies that address data scarcity in TST by utilizing both non-parallel (see Section 3.1) and low-resource parallel (see Section 3.2) data. We discuss strategies for maximizing the utility of limited parallel data while proposing approaches that effectively leverage non-parallel data, aiming to generalize TST tasks under constrained data conditions. For both approaches, in our experiments, we consider text sentiment transfer which is a substak of TST (as discussed in Section 2.2.2). Section 3.3 concludes the chapter by highlighting its limitations and suggesting directions for future research.

3.1 Our methodology using non-parallel data

This section is based on the paper *Balancing the Style-Content Trade-Off in Sentiment Transfer Using Polarity-Aware Denoising* (Mukherjee et al., 2022), published in the Proceedings of the 25th Text, Speech, and Dialogue Conference (TSD 2022).

A significant challenge in TST arises when utilizing non-parallel datasets, where no direct sentence mappings exist between source and target styles.

As discussed in Section 2.2.3, existing approaches address this challenge through either disentangling text representations into style-independent content and stylistic attributes using generative modeling (Hu et al., 2017; Shen et al., 2017; Prabhumoye et al., 2018) or prototype editing (Li et al., 2018; Fu et al., 2019), which modifies specific style markers to achieve the desired style. However, both methods face limitations, such as the difficulty of achieving disentanglement without inductive biases and errors in content preservation due to unsupervised identification of pivot words (Locatello et al., 2019).

In this work, we integrate both of the above research directions and extend them by incorporating additional supervision. Our approach is structured around two key stages in the sentiment transfer process: improved representation learning using a shared encoder and sentiment-controlled generation with separate sentiment-specific decoders. Initially, sentiment-marker pivot words (identified using a sentiment dictionary) in input sentences are randomly deleted or masked. A shared encoder, pre-trained on a general domain, then generates a latent representation of the text. This representation is subsequently processed by sentiment-specific decoders to alter the sentiment of the original sentence. Empirical results demonstrate that our method outperforms state-of-the-art baselines in terms of content preservation while remaining competitive regarding style transfer accuracy and fluency.

This section is organized as follows. We begin with an overview of our proposed architecture (Section 3.1.1) and a discussion of our model variants (Section 3.1.2). Next, we describe our datasets (Section 3.1.3) and training setup (Section 3.1.4), followed by an overview of the external baselines (Section 3.1.5). We then detail our evaluation methodology, including both automatic metrics (Section 3.1.6) and human evaluation (Section 3.1.7). Finally, we present and discuss our results in Section 3.1.8.

3.1.1 Model Overview

Figure 3.1 shows the overview of our proposed architecture. Following Prabhumoye et al. (2018), we first translate the input text x in the base language to a chosen intermediate language $\bar{x}$ using a translation model.

$$x_{noise} = Noise(\bar{x}; \theta_N). \tag{3.1}$$

We provide x_{noise} to the encoder of the $\bar{x} \rightarrow \hat{x}$ back-translation model (where $\hat{x}$

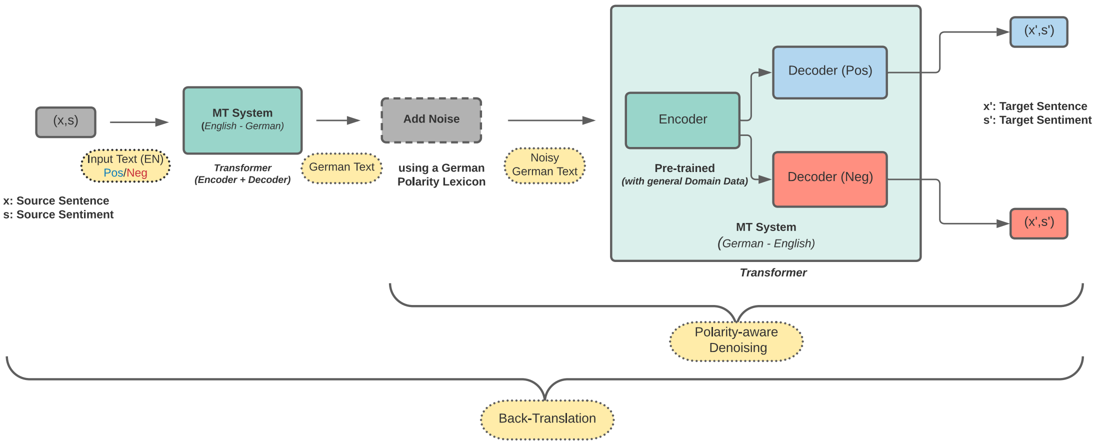


Figure 3.1: Our sentiment transfer pipeline. We (1) *translate* the source sentence from English to German using a transformer-based machine translation (MT) system; (2) *apply noise* on the German sentence using a German polarity lexicon; (3) *encode* the German sentence to latent representation using an encoder of German-to-English translation model; (4) *decode* the shared latent representation using the decoder for the opposite sentiment.

is a text in the base language with changed sentiment polarity). The encoder first converts the text to the latent representation z as follows:

$$z = Encoder(x_{noise}; \theta_E), \tag{3.2}$$

where θ_E represent the parameters of the encoder.

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

$$x_{pos} = Decoder_{pos}(z; \theta_{D_{pos}}) \tag{3.3}$$
$$x_{neg} = Decoder_{neg}(z; \theta_{D_{neg}}). \tag{3.4}$$

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

$$\hat{x}_{neg} = Decoder_{pos}(z; \theta_{D_{pos}}) \tag{3.5}$$

$$\hat{x}_{pos} = Decoder_{neg}(z; \theta_{D_{neg}}) \tag{3.6}$$

where $\hat{x}_{neg}$, $\hat{x}_{pos}$ are the sentences with transferred sentiment conditioned on z and $\theta_{D_{pos}}$ and $\theta_{D_{neg}}$ represent the parameters of the positive and negative decoders, respectively.

3.1.2 Model Variants

In all our experiments in this section, we train sentiment transfer models using back-translation based on the transformer architecture (Vaswani et al., 2017). First, we present baselines for sentiment transfer with simple style conditioning. Next, we propose an approach based on an extended transformer architecture where we use separate modules (either the whole transformer model, or the transformer decoder only) for the respective target sentiment. We further improve upon our approach using polarity-aware denoising, which we propose as a new scheme for pre-training the sentiment transfer models.

Back-translation Back-translation for style transfer was introduced in Prabhumoye et al. (2018). Following their approach, we use translation into German and subsequent encoding in a back-translation model to get a latent text representation for our sentiment transfer task. Prior work has also shown that the process of translating a sentence from a source language to a target language retains the meaning of the sentence but does not preserve the stylistic features related to the author's traits (Rabinovich et al., 2017). A pure back-translation approach (without any specific provisions for sentiment) is referred to as *Back-Translation* in our experiments.[1]

Our Baseline Models In addition to a pure back-translation model, we present several straightforward baselines:

- *Style Tok* is a back-translation model with added sentiment identifiers (*<pos>* or *<neg>*) as output starting tokens. At the time of sentiment transfer, we decode the output with a changed sentiment identifier (*<pos>* $\rightarrow$ *<neg>*, *<neg>* $\rightarrow$ *<pos>*).

[1]We also experimented with an auto-encoder, but we have found that the back-translation model gives better results for sentiment transfer. We hypothesise that it is due to the fact that back-translation prevents the system from sticking to a particular wording, resulting in a more abstract latent representation.

- *Two Sep. transformers:* To get more control over sentiment-specific generation, we train two separate transformer models for positive and negative sentiment, using only sentences of the respective target sentiment. During inference, the model is fed with inputs of the opposite sentiment, which it did not see during training.
- *Shrd Enc + Two Sep Decoders:* We extend the above approach by keeping decoders separate, but using a shared encoder. During training, all examples are passed through the shared encoder, but each decoder is trained to only generate samples of one sentiment. Sentiment transfer is achieved by using the decoder for the opposite sentiment.
- *Pre Training Enc:* Following Gururangan et al. (2020), we introduce a variant where the shared encoder is pretrained for back-translation on general-domain data. The pre-trained encoder is then further fine-tuned during sentiment transfer training.

Sentiment-Aware Denoising We devise a task-specific pre-training scheme (Gururangan et al., 2020) for improving the sentiment transfer abilities of the model. The idea of our pre-training scheme—*sentiment-aware denoising*—is to first introduce noise, i.e. delete or mask a certain proportion of words in the intermediate German input to the back-translation step, then train the model to remove this noise, i.e. produce the original English sentence with no words deleted or masked. To decide which words get deleted or masked, we use automatically obtained sentiment polarity labels (see Section 3.1.4 for implementation details). This effectively adds more supervision to the task on the word level. We apply three different approaches: deleting or masking (1) *general* words (i.e., all the words uniformly), (2) *polarity* words (i.e., only high-polarity words according to a lexicon), or (3) both *general* and *polarity* words (each with a different probability).

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

3.1.3 Datasets

For our back-translation process and model pretraining, we used the WMT14 English-German dataset (1M sentences) from Neidert et al. (2014).

For finetuning and experimental evaluation, we built a new English dataset for sentiment transfer, based on the Amazon Review Dataset (Ni et al., 2019). We have selected Amazon Review because it is more diverse topic-wise (books, electronics, movies, fashion, etc.) than existing sentiment transfer datasets, Yelp (Li et al., 2018) and IMDb (Dai et al., 2019), which are majorly focused on movie and restaurant/business-related reviews. For comparison with previous work, we also evaluate our models on the benchmark IMDb dataset (Dai et al., 2019).

While the Amazon Review data is originally intended for recommendation, it lends itself easily to our task. We have split the reviews into sentences and then used a pre-trained transformer-based sentiment classifier (Wolf et al., 2020) to select sentences with high polarity. Our intuition is that high-polarity sentences are more informative for the sentiment transfer task than neutral sentences. We filter out short sentences (less than 5 words) since it is hard to evaluate content preservation for these sentences. We also ignored sentences with repetitive words (e.g., *"no no no no thanks thanks."*) because these sentences are noisy and do not serve as good examples for the sentiment transfer model. We aim at comparable size to existing datasets (Li et al., 2018): the resulting data has 102k positive and 102k negative examples in total, with 1+1k reserved for validation and testing for each sentiment. The average sentence length in our data is 13.04 words.

3.1.4 Training Setup

In all our experiments, we use a 4-layer transformer (Vaswani et al., 2017) with 8 attention heads per layer. Both embedding and hidden layer size are set to 512. The same model shape is used for both the initial translation into German and the subsequent model handling back-translation, denoising, and sentiment transfer.

We use a German polarity lexicon to automatically identify pivot words for polarity-aware denoising. We developed the German polarity lexicon by translating all vocabulary words from the German training text into English using an off-the-shelf translation system (Košarko et al., 2019), followed by labeling the words with *positive* and *negative* labels using the English NLTK Vader lexicon (Hutto and Gilbert, 2014). We manually validated the results using a small random subset.

We test various combinations of noise settings w.r.t. noise probability, noise type (general or polarity-aware denoising), and noise mode (deleting or masking). Parameter values are pre-selected based on our preliminary experiments with the translation model. The parameters are encoded in the name of the model as used in Table 3.1 (see the table caption for details).

3.1.5 External Baselines

We evaluated and compared our approaches described in Section 3.1.2 with several state-of-the-art systems (Shen et al., 2017; Prabhumoye et al., 2018; Li et al., 2018; Luo et al., 2019a; Wang et al., 2019; He et al., 2020b), as discussed in Chapter 2, using two datasets (see Section 3.1.3).

3.1.6 Automatic Evaluation Metrics

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

Standard Metrics We evaluate our sentiment transfer model using three key metrics: *Style Transfer Accuracy*, *Fluency*, and *Content Preservation*. For style transfer accuracy, we automatically assess the sentiment of transferred sentences using a pre-trained transformer-based sentiment analysis pipeline from Huggingface[2] (Wolf et al., 2020). Fluency is measured indirectly by calculating the negative log-likelihood from the GPT-2 language model (Radford et al., 2019) and additionally by determining the average sentence lengths of the transferred sentences. Content preservation is evaluated by computing the BLEU score (Papineni et al., 2002) between the transferred and source sentences, as well as the cosine similarity score using Sentence BERT (Reimers and Gurevych, 2019) to assess semantic similarity. However, none of these techniques specifically evaluate content preservation in style transfer (Toshevska and Gievska, 2021), as they do not account for the necessity of altering individual words while changing the sentence style, resulting in the penalization of intended differences between source and transferred sentences.

Newly Introduced Metrics for Content Preservation To avoid the problems of the existing metrics, it makes sense in sentiment transfer to evaluate the content and similarity while ignoring sentiment-bearing words (pivot words). Thus, we introduce MaskBLEU and MaskSim scoring methods – these are identical to BLEU and cosine similarity, but they are computed on sentences where pivot words (based on NLTK Vader sentiment dictionary (Hutto and Gilbert, 2014), as discussed in Section 3.1.4) have been masked. This allows measuring content preservation while ignoring the parts of the sentences that need to be altered for sentiment transfer task.

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

Models	Acc	Sim	M/Sim	B	M/B	LM	Len	Avg
Back-Translation Only								
Back-translation only	0.4	0.828	0.768	28.0	45.3	-78.6	11.9	40.9
Our Baseline Models								
Style Tok	13.2	0.536	0.560	4.8	8.6	-52.1	7.6	25.9
Two Sep. transformers	89.3	0.394	0.611	6.8	19.6	-79.0	13.7	56.7
Shrd Enc + Two Sep Decoders	88.1	0.397	0.600	7.3	20.1	-78.0	12.5	56.0
Pre Training Enc	55.3	0.592	0.732	22.6	33.9	-93.3	13.4	54.1
Our Models (with Denoising)								
WG01-AG01-D	71.4	0.517	0.694	17.1	29.8	-88.7	13.7	56.9
WG01-AG01-M	68.0	0.536	0.711	19.4	31.1	-86.3	12.6	56.7
WG03-AG03-D	83.0	0.447	0.648	11.7	24.4	-83.0	13.7	57.4
WG03-AG03-M	78.8	0.481	0.669	14.2	28.2	-82.7	13.0	58.0
WP08-AP08-D	66.9	0.528	0.701	19.5	31.3	-82.8	12.4	56.1
WP08-AP08-M	64.0	0.547	0.726	21.4	34.0	-89.1	12.9	56.9
WP1-AP1-D	58.7	0.570	0.727	22.7	33.1	-87.2	12.2	54.8
WP1-AP1-M	58.9	0.567	0.716	22.2	33.0	-86.5	12.2	54.5
WG03-AG01-D	68.0	0.529	0.697	17.9	30.9	-89.5	13.3	56.2
WG03-AG01-M	80.7	0.473	0.665	13.9	27.5	-82.8	13.1	58.2
WG01-AG03-D (=SCT$_2$)	85.2	0.441	0.646	11.8	25.4	-79.8	13.1	58.4
WG01-AG03-M	70.0	0.534	0.711	19.7	32.3	-84.3	12.4	57.8
WP08-AP1-D	61.6	0.578	0.736	22.5	35.0	-94.4	13.4	56.7
WP08-AP1-M	60.9	0.554	0.724	22.0	33.3	-85.5	12.6	55.6
WP1-AP08-D	68.5	0.525	0.699	19.3	31.1	-84.0	12.4	56.5
WP1-AP08-M	61.1	0.560	0.714	21.5	32.9	-86.0	12.1	55.1
WG03-AP08-D	67.0	0.533	0.697	20.3	31.7	-84.3	12.5	56.1
WG03-AP08-M	65.7	0.546	0.725	21.2	33.5	-85.0	12.5	57.2
WP08-AG03-D	83.3	0.436	0.635	11.0	24.3	-80.5	13.3	57.0
WP08-AG03-M	79.6	0.473	0.665	13.2	26.9	-83.1	13.2	57.6
WG03P08-AG03P08-D	65.5	0.547	0.705	20.3	32.6	-90.4	13.2	56.2
WG03P08-AG03P08-M (=SCT$_1$)	82.0	0.460	0.665	13.7	27.4	-79.6	12.8	58.6
State-of-the-Art Models								
Shen et al. (2017)	88.6	0.346	0.513	3.2	18.3	-74.0	10.9	52.7
Li et al. (2018)	69.9	0.457	0.632	14.7	25.3	-85.1	12.2	52.8
Luo et al. (2019a)	92.4	0.279	0.468	0.0	9.1	-42.0	7.8	49.4
Prabhumoye et al. (2018)	93.5	0.308	0.504	0.9	15.2	-61.0	10.3	53.0
Wang et al. (2019)	79.3	0.385	0.545	10.6	20.3	-116.8	15.1	51.4
He et al. (2020b)	91.5	0.352	0.542	9.5	21.8	-65.9	8.2	55.8

Table 3.1: **Automatic evaluation**. *Accuracy*: Sentiment transfer accuracy. *Sim* and *B*: Cosine similarity and BLEU score between input and sentiment-transferred sentence. *M/Sim* and *M/B*: MaskSim and MaskBLEU (similarity and BLEU with polarity words masked, see Section 3.1.6). *LM*: Average log probability assigned by vanilla GPT-2 language model. *Avg*: Average length of transferred sentences. *Avg*: Average of sentiment transfer accuracy, 100*MaskSim and MaskBLEU. Scores are based on a single run, with identical random seeds. First two sections show our own baselines, third section shows our models with denoising (with the best settings denoted SCT$_1$ and SCT$_2$, see Section 3.1.8). The bottom section shows a comparison with state-of-the-art models. Names of models with denoising reflect settings as follows: *W* denotes WMT pretraining data, *A* denotes Amazon finetuning data; the following tokens denote noise probability values are associated with the respective data. *G/P* represents general/polarity token noising, *D/M* represents noising mode deletion/-masking. E.g. *WG03P08-AG03P08-D*: noise probabilities on WMT and Amazon data are identical, noising by deletion on both general and polarity token noising is applied (with probabilities 0.3 and 0.8, respectively).

Models	Acc	Sim	M/Sim	B	M/B	LM	Len	Avg
Prabhumoye et al. (2018)	87.1	0.345	0.480	2.7	14.3	-63.5	10.0	49.8
Li et al. (2018)	21.0	0.587	0.668	18.3	25.9	-83.6	15.3	37.9
Wang et al. (2019)	84.0	0.357	0.456	9.2	13.2	-63.9	10.8	47.6
He et al. (2020b)	81.7	0.458	0.576	29.0	41.8	-83.6	15.3	60.4
SCT_1 (WG03P08-AG03P08-M)	85.3	0.435	0.612	28.6	42.3	-86.4	15.9	**62.9**
SCT_2 (WG01-AG03-D)	88.2	0.379	0.588	25.8	39.2	-79.6	15.1	62.1

Table 3.2: Automatic evaluation on the IMDb Dataset (see Table 3.1 for metrics explanation).

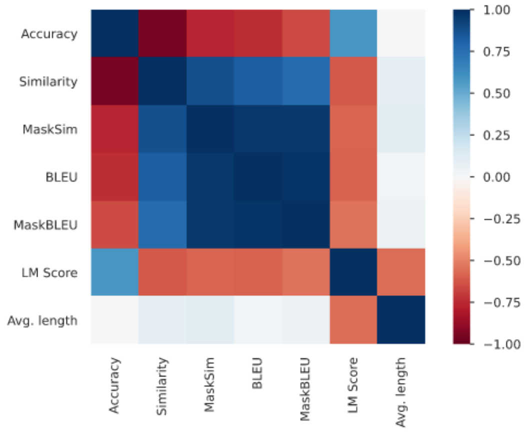


Figure 3.2: Correlations between automatic evaluation metrics on our Amazon Review data: sentiment accuracy is negatively correlated with BLEU, semantic similarity, and their masked variants.

3.1.7 Human Evaluation

As automated metrics for language generation do not correlate well with human judgements (Novikova et al., 2017), we conduct an in-house human evaluation with five expert annotators. We randomly select 100 sentences (50 for each sentiment) from the our Amazon Review test set. The annotators rate model outputs on using a 1-5 Likert scale for style control, content preservation and fluency.

3.1.8 Results

Automatic Metrics results on our Amazon Review data are shown in Table 3.1. Overall, there is clearly a tradeoff between preserving sentiment-independent content and achieving the desired target sentiment. Models which perform very well in sentiment transfer usually achieve worse results on content preservation. This tradeoff is documented by correlations between the automatic metrics (see Figure 3.2). Sentiment accuracy is negatively correlated with BLEU score, similarity measures as well as our newly introduced MaskBLEU and MaskSim scores.

Models	Sentiment	Content	Fluency
Prabhumoye et al. (2018)	3.95	1.19	3.56
Li et al. (2018)	3.35	2.30	3.34
Wang et al. (2019)	3.48	1.67	2.54
He et al. (2020b)	3.69	1.66	3.26
SCT_1 (WG03P08-AG03P08-M)	3.94	**2.61**	3.73
SCT_2 (WG01-AG03-D)	**3.99**	2.56	**3.79**

Table 3.3: Human evaluation of sentiment transfer quality, content preservation, and fluency. Average of 1-5 Likert scale ratings on 100 examples from our Amazon Review data.

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

The main motivation for our work was to find a denoising strategy which offers the best balance between sentiment transfer and content preservation. Our results suggest putting an emphasis on denoising high-polarity words results in the best ratio between the sentiment transfer accuracy and content preservation metrics. Additionally, our models show the ability to produce fluent sentences, as shown by the language model score: our models' scores are similar to the back-translation baseline; other models only reach higher language model scores when producing very short outputs.

Overall, our denoising approaches are able to balance well between sentiment transfer and content preservation. On content preservation, they perform much better than state-of-the-art models, and they stay competitive on style accuracy. We selected two of our model variants – SCT_1=*WG03P08-AG03P08-M* and SCT_2=*WG01-AG03-D* – as the ones giving the best style-content trade-off (SCT) according to the average of sentiment accuracy, masked similarity and MaskBLEU (see Table 3.1).

Automatic metrics on the IMDb dataset (Dai et al., 2019) are shown in Table 3.2, comparing our selected SCT_1 and SCT_2 models with state-of-the-art. Our models outperform the state-of-the-art in terms of sentiment accuracy and reach competitive results in terms of similarity, BLEU, and fluency. Same as on our Amazon Review data, they provide the best style-content trade-off (according to the averaged metric defined in Table 3.1).

Human Evaluation Results: We compare our best SCT_1 and SCT_2 models (selected above) with four state-of-the-art models: two of the most recent models (Wang et al., 2019; He et al., 2020b), and the models with best accuracy (Prabhumoye et al., 2018) and MaskBLEU score (Li et al., 2018).

We have evaluated over 600 model outputs. Results are presented in Table 3.3. The human evaluation results mostly agree with our automatic evaluation results. The results also show that our models are better in content preservation than the competitor models.

We further examined a sample of the outputs in more detail to understand the behavior of different models. We found that state-of-the-art models tend to lose the content of the source sentence, as shown in the example outputs in Table 3.4. On the other hand, our models mostly preserve sentiment-independent content well while successfully transferring the sentiment. We conclude that with our models, there is a good balance between preserving the original sentiment-independent content and dropping the source sentiment, and existing state-of-the-art models tend to sacrifice one or the other.

	Negative → Positive	**Positive → Negative**
Source	**movie was a waste of money : this movie totally sucks .**	**my daughter loves them :)**
Prabhumoye et al. (2018)	stan is always a great place to get the food .	do n't be going here .
Li et al. (2018)	our favorite thing was a movie story : the dream class roll !	my daughter said i was still not acknowledged .
Wang et al. (2019)	movie is a delicious atmosphere of : this movie totally sucks movie !	i should not send dress after me more than she would said not ?
He et al. (2020b)	this theater was a great place , we movie totally amazing .	yup daughter has left ourselves .
SCT_1 (WG03P08-AG03P08-M)	movie : a great deal of money : this movie is absolutely perfect .	my daughter hates it : my daughter .
SCT_2 (WG01-AG03-D)	this movie is a great deal of money.	my daughter hated it .
Source	**nothing truly interesting happens in this book .**	**best fit for my baby : this product is wonderful ! !**
Prabhumoye et al. (2018)	very good for the best .	bad customer service to say the food , and it is n't .
Li et al. (2018)	nothing truly interesting happens in this book .	my mom was annoyed with my health service is no notice .
Wang et al. (2019)	nothing truly interesting happens in this book make it casual and spot .	do not buy my phone : this bad crap was worst than it ?
He et al. (2020b)	haha truly interesting happens in this book .	uninspired .
SCT_1 (WG03P08-AG03P08-M)	in this book is truly a really great book .	not good for my baby : this product is great ! ! ! ! ! ! ! !
SCT_2 (WG01-AG03-D)	in this book is truly awesome .	not happy for my baby : this product is not great ! !

Table 3.4: Example outputs comparison on samples from our Amazon Reviews dataset. Sentiment marker words (pivots) are colored. Note that our models preserve content better than most others.

3.2 Our methodology using parallel data

This section is based on the paper *Leveraging Low-resource Parallel Data for Text Style Transfer* (Mukherjee and Dusek, 2023), published in the Proceedings of the 16th International Natural Language Generation Conference (INLG 2023).

A significant challenge in TST is the limited availability of parallel training data, as acquiring large-scale aligned datasets for specific style pairs is often impractical or unfeasible (Jin et al., 2022; Hu et al., 2022). The only known study utilizing parallel data and sequence-to-sequence learning is by Jhamtani et al. (2017), which focuses on a specific application: converting modern English to Shakespearean style. This unique case benefits from the existence of extensive aligned paraphrases for literary text. In contrast, as discussed in Section 3.1, most TST research employs non-parallel datasets and unsupervised learning methods (Hu et al., 2017; Zhao et al., 2018; Li et al., 2018). While these approaches have shown promising results, they often incur a performance trade-off and cannot fully overcome the data scarcity issue, as large quantities of non-parallel, style-specific data remain difficult to obtain (Li et al., 2022b).

In this work, we address the challenges of TST in low-resource scenarios by proposing methodologies that leverage minimal parallel data. Unlike the approaches discussed in Section 3.1, which employ transformer models trained from scratch, our methodology in this section utilizes pretrained language models to enhance performance in low-resource settings. Our approach integrates multiple low-resource adaptation techniques and introduces a novel style-reward-mechanism. By utilizing a TST system designed for low-resource parallel data, we achieve well-balanced results that surpass previous non-parallel approaches in both automatic and human evaluations.

This section is organized as follows. We begin with an overview of our proposed methodologies (Section 3.2.1). Next, we describe our dataset (Section 3.2.2), parameter settings (Section 3.2.3), and external baselines (Section 3.2.4). Finally, we detail our evaluation and analysis (Section 3.2.5).

3.2.1 Methodologies

We build on transfer learning by finetuning a pretrained BART model on our task (Lewis et al., 2020). We further explore five techniques aimed at this low-resource scenario:

Hyperparameter tuning: As the effectiveness of Transformer models on low-resource data highly depends on hyperparameters (Araabi and Monz, 2020), we adapt our model, focusing on dropout regularization (Sennrich and Zhang, 2019) and label smoothing (Müller et al., 2019).

Prompt-guided generation: To align the style transfer finetuning with pre-training, we adopt using textual prompts, following Li and Liang (2021) and Li et al. (2022a). By adding prompts like *"POS:"* for positive sentences and *"NEG:"* for negative sentences, we provide explicit guidance to the decoder during fine-tuning (this approach builds upon the *Style Tok* model introduced in Section 3.1).

Data augmentation: We use data augmentation by paraphrasing (see Section 3.2.3) to generate more training examples and improve data diversity (Shen et al., 2020a; Qiu et al., 2020).

Self-training: To further expand our data, we use self-training, i.e., training on synthetic data generated by the model itself (He et al., 2020a; Chai et al., 2022). To improve the quality of the synthetic data, we filter them using style classifier accuracy, BLEU, and embedding similarity (cf. Section 3.2.5). We use a geometric mean of all three metrics as a sentence score, then choose a portion of the generated data with the top k highest scores.

Style reward: To make our generator better focus on the target style accuracy, we incorporate rewards from a style classifier into the training loss. We use a simple reward R, which is $+1$ for instances where the generated output matches the target style, and -1 where it does not. We then modify the basic cross-entropy generation loss $\mathcal{L}_{\text{CE}}$ in the following way to get the overall loss $\mathcal{L}$:

$$\mathcal{L} = \alpha \cdot \mathrm{norm}(R) + (1 - \alpha) \cdot \mathcal{L}_{\text{CE}} \tag{3.7}$$

norm denotes normalization (zero mean, unit standard deviation), and α is a weight parameter.

3.2.2 Dataset

We experiment on a small parallel sentiment transfer dataset of Yelp reviews by Li et al. (2018), comprising 500 positive-to-negative and 500 negative-to-positive sentences. The data was intended as an evaluation set only, but we repurpose it as a full low-resource set and split it into 400 examples for training, 100 for development, and 500 for testing. For self-training, we additionally use non-parallel sets of 2000+2000 positive and negative sentences from Li et al. (2018)'s development set.

3.2.3 Parameter Settings

We use BART-base (Lewis et al., 2020) from the HuggingFace library (Wolf et al., 2020).

Hyperparameter tuning: We ran three small-scale random searches for optimal values of individual parameters, resulting in the following changes from the defaults based on development set results: (1) We adjusted the learning rate (*LR*) ($5e-5 \rightarrow 1e-5$) and *batch size* ($8 \rightarrow 3$). (2) We increased the *Dropout* rate ($0.1 \rightarrow 0.15$) and introduced additional attention and activation dropout (both 0.1). (3) We introduced *L2* regularization with a value of 0.01 and *label smoothing* with a value of 0.05.

Prompt-guided generation: This does not have any specific settings; we only add the prompts on the input as described in Section 3.2.1.

Data augmentation: We used the following operations from the NLPAug library (Ma, 2019): substitute words with a *Spelling* mistake from a dictionary, *Insert* or *Substitute* words based on BERT embedding similarity, substitute words with a *Synonym* from WordNet, *Swap* or *Delete* words randomly, *Split* words into two tokens randomly. Additionally, we used *Back-translation* (Sennrich et al., 2016c; Prabhumoye et al., 2018) via German using the online translation tool of Košarko et al. (2019) (building upon the back-translation process we discussed in Section 3.2.1).

We apply an augmentation to each training data example at random with a 50% probability (i.e., roughly 200 additional instances per augmentation type). We also consider an *"All"* setting where we include all augmented data.

Self-training: We generated parallel synthetic data of various sizes up to 2k examples. We further applied our filtering via automatic metrics (see Section 3.2.1) to choose the best 1k out of 2k examples.

Models	Style	Content	Fluency
Li et al.	2.36	1.57	1.58
ChatGPT	4.48	2.75	4.49
Ours	3.98	3.96	4.45

Table 3.5: Human evaluation of 100 randomly selected outputs on style transfer accuracy (Style), Content Preservation (Content), and Fluency .

Style reward: We train a simple BERT-based (Devlin et al., 2019) sentiment classifier for this experiment, only using the same limited training set as for the main task. Its accuracy on our test set is 95.8%. We use this classifier for the style rewards, with a $\alpha = 0.5$, i.e., even split between the base cross-entropy loss and the style rewards.

3.2.4 External baselines

We compare our approaches to well-performing systems for sentiment transfer that utilize large non-parallel datasets, as described in Chapter 2. Our goal is to demonstrate the effectiveness of leveraging low-resource parallel data. As discussed in Section 3.1.5, we evaluate against three baselines[3]: Shen et al. (2017)'s cross-aligned autoencoder with style-specific decoders, Prabhumoye et al. (2018)'s system based on back-translation via French, and Li et al. (2018)'s text-replacement-based approach.

We also compare to state-of-the-art instruction-finetuned large language models (LLMs): ChatGPT[4] and HuggingFace Chat.[5] We prompt them with a task specification and 10 randomly chosen examples from the training set. We only report results for ChatGPT, as HuggingFace Chat did not adhere to the given task, and its outputs were not parsable with our evaluation scripts.

3.2.5 Evaluation & Results

We evaluate three main dimensions: style transfer accuracy, content preservation, and fluency, as outlined in Chapter 2 and utilized for evaluation in Section 3.1.6. We measure sentiment accuracy using DistilBERT (Sanh et al., 2019) finetuned for sentiment analysis on the SST-2 dataset (Socher et al., 2013).[6] Following prior work (Jin et al., 2022; Hu et al., 2022), we evaluate content preservation using BLEU score (Pap-

[3]We faced difficulties when attempting to run some other recent approaches on our data (Xiao et al., 2021; Lee, 2020).

[4]`https://openai.com`, model gpt-3.5-turbo.

[5]`https://huggingface.co/chat/`, model OpenAssist-ant/oasst-sft-6-llama-30b (Köpf et al., 2023).

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

ineni et al., 2002) and embedding similarity (Rahutomo et al., 2012) against the input sentences. We use Sentence-BERT (Reimers and Gurevych, 2019) and cosine similarity for the embedding similarity. We use GPT-2's (Radford et al., 2019) perplexity to estimate fluency.

We also run a small-scale in-house human evaluation on a random sample of 100 sentences from the test set (50 for each direction – positive-to-negative and negative-to-positive). Outputs are rated on a 5-point Likert scale for style transfer accuracy, content preservation, and fluency.

Automatic Metrics Results Table 3.6 shows automatic metrics results. Our base BART model (experiment 01) performs decently in all metrics, but style accuracy is further improved via hyperparameter tuning (02-04), with a slight drop in BLEU score. Adding prompts (05) further increases style accuracy and makes up for the content similarity drop.

Data augmentation (06-14) leads to further improvements, especially for replacing *Synonyms* from WordNet (09), random word *Deletion* (10), and *Back-translation* (11). The best performance is achieved using *All* (14) data augmentation types (which also means a larger number of augmented examples). Augmentation generally leads to a style accuracy increase; perplexity rises, but BLEU and embedding similarity is preserved, indicative of less frequent expressions, but not much change in content. Self-training with synthetic data (15-20) maintained the performance across the board with a slight improvement in BLEU score, but synthesizing too many examples does not lead to further improvements (18-19), likely due to an imbalance between original and synthetic data. The best results are achieved using 1k synthesized instances filtered using automatic metrics (20).

Using style rewards and combining them with data augmentation (21) or self-training (22) brings further improved style accuracy, with other metrics staying approximately the same. Since both experiments 21 and 22 perform very similarly, we choose 22 as the best model for further evaluation because the self-training approach does not require additional tools, unlike the data augmentation toolkit needed for 21.

Compared to unsupervised approaches (23-25), our experiments show similar or better style accuracy while maintaining content preservation and fluency, both of which are very low for unsupervised systems. ChatGPT (26) excelled in style transfer accuracy and fluency, but also lacked in content preservation.

ID	Models	ACC	BLEU	CS	PPL
	Baseline				
01	BART-base	55.4 ± 2.6	33.8 ± 0.2	65.5 ± 0.9	127.7 ± 2.4
	Hyperparameter tuning				
02	01 + LR & batch size	61.7 ± 3.1	33.1 ± 0.2	67.6 ± 1.4	126.4 ± 1.6
03	02 + Dropout	61.1 ± 2.7	33.3 ± 0.3	67.4 ± 1.3	126.1 ± 1.2
04	03 + L2 & label smoothing	61.6 ± 3.1	33.2 ± 0.3	67.6 ± 1.4	126.9 ± 1.4
	Prompt-guided generation				
05	04 + Prompt	67.7 ± 2.6	33.3 ± 0.3	70.1 ± 1.0	126.7 ± 1.8
	Data augmentation				
06	05 + Spelling	71.1 ± 2.5	33.6 ± 0.4	70.0 ± 1.2	132.2 ± 2.2
07	05 + Insert	71.6 ± 2.4	33.1 ± 0.4	70.8 ± 1.4	131.5 ± 0.9
08	05 + Substitute	70.9 ± 3.5	33.2 ± 0.6	69.9 ± 1.2	131.9 ± 1.3
09	05 + Synonym	71.5 ± 2.7	33.5 ± 0.5	71.2 ± 2.1	131.9 ± 0.9
10	05 + Delete	72.0 ± 1.9	33.0 ± 0.5	70.7 ± 1.8	132.6 ± 0.8
11	05 + Back-translation	72.7 ± 2.5	32.9 ± 0.7	70.6 ± 1.3	132.7 ± 1.6
12	05 + Swap	71.1 ± 3.3	33.5 ± 0.1	70.1 ± 1.0	131.9 ± 1.4
13	05 + Split	70.8 ± 4.5	33.5 ± 0.4	70.5 ± 1.4	133.5 ± 0.7
14	05 + All	74.2 ± 3.2	33.2 ± 0.7	70.6 ± 2.7	132.5 ± 1.5
	Self-training				
15	05 + 250	68.4 ± 2.5	33.4 ± 0.2	69.4 ± 1.5	132.5 ± 0.4
16	05 + 500	70.5 ± 5.0	33.6 ± 0.5	71.4 ± 2.3	132.3 ± 2.2
17	05 + 1k	71.5 ± 4.8	34.1 ± 0.4	70.5 ± 2.7	131.0 ± 2.8
18	05 + 1.5k	70.1 ± 5.0	34.2 ± 0.2	70.8 ± 2.8	132.4 ± 1.2
19	05 + 2k	70.0 ± 4.6	34.3 ± 0.2	70.2 ± 2.2	132.4 ± 1.6
20	05 + 1k filtered	72.6 ± 4.4	34.2 ± 0.4	71.5 ± 2.3	132.7 ± 1.3
	Style reward				
21	14 + reward	78.8 ± 2.7	33.1 ± 0.7	72.4 ± 2.4	132.8 ± 1.5
22	20 + reward	78.4 ± 2.9	33.9 ± 0.7	72.2 ± 1.9	132.6 ± 1.2
	External baselines				
23	Shen et al.	64.4	6.7	46.0	338.5
24	Li et al.	71.9	11.6	55.3	366.6
25	Prabhumoye et al.	72.4	3.0	41.7	318.8
26	ChatGPT	95.4	19.4	61.4	115.3

Table 3.6: Automatic evaluation results. We measure the sentiment classifier accuracy (ACC), BLEU score, Content Similarity (CS), and Fluency (PPL). The model names follow a format of experiment ID + Model name, indicating that the current model is built upon a base model from that particular ID. All our models' scores are averages of five runs with different random initializations, with standard deviations shown after "±".

Human Evaluation For the human evaluation, we compared our chosen model (experiment 22) with Li et al. (2018)'s work (24) and ChatGPT (26), chosen for their best automatic metrics results of the external models. The results in Table 3.5 largely confirm the automatic metrics results – the unsupervised system shows relatively poor performance, and while ChatGPT excels in hitting the target style, our approach is best on content preservation. Table 3.7 shows a few illustrative examples, comparing our chosen best model (22) with external baselines (as discussed in Section 3.2.4).

	Negative → Positive	Positive → Negative
Source	terrible menu, high prices, bad customer service .	it 's a much better option than the club scene .
Gold	nice menu , good prices , great service - for both dinner and breakfast !	i would rather go to the club than here .
Shen et al.	fantastic selection of great customer !	it 's a good experience for the whole airport i would .
Li et al.	no nonsense in service .	it 's a much better than the club scene .
Prabhumoye et al.	bad customer service with the food of this location .	she did n't go back with this place .
ChatGPT	marvelous entertainment, budget-friendly choices, exceptional atmosphere.	absolutely disastrous, it's worse than the late-night traffic.
Ours (exp. 22)	great menu, high prices, great customer service .	it's a terrible alternative to the club scene .
Source	the bad news that my vision had deteriorated made the visit even worse .	all of my clothes are returned in sparkling condition !
Gold	the good news that my vision had improved made the visit even better .	all of my clothes are returned in terrible condition !
Shen et al.	the good thing i have the whole nails made my whole gem !	all of my car here are nothing in any room .
Li et al.	the problem was the red deal by handles the night my questions did n't .	all of my clothes are returned in my condition !
Prabhumoye et al.	the worst time i have ever had to get a disappointment .	all of the food is not very good in all .
ChatGPT	the remarkable revelation of my surprise birthday party plans made the visit even more special.	The condition of all my belongings is extremely terrible!
Ours (exp. 22)	the good news that my vision had improved made the visit even better .	all of my clothes are returned in terrible condition !
Source	it's located in a slum scottsdale area and isn't accomodating.	my father has decided to upgrade my mothers engagement ring this xmas .
Gold	it 's located in a great part of scottsdale and was really accommodating .	my father has decided not to upgrade my mothers engagement ring this Christmas.
Shen et al.	cute shop in a sunday area and desert !	my son did to have my whole card to celebrate my appointment off .
Li et al.	no bueno in the north nonsense and not acknowledged a word or anything .	my father has decided to upgrade paint now .
Prabhumoye et al.	minutes later for the food and not worth the food .	my husband ordered me to get the worst service in the food .
ChatGPT	this place is family-owned, but it could greatly benefit from improving their staff.	my father has decided to downgrade my mother's engagement ring.
Ours (exp. 22)	it's located in a slum scottsdale area and is accomodating.	my father has decided not to upgrade my mothers engagement ring this xms.

Table 3.7: Example output comparison on samples from the test set. Sentiment marker words are colored. Note that our model balances well between style transfer accuracy and content preservation, better than others.

3.3 Conclusion

In this chapter, we explore approaches that utilize both non-parallel and parallel data from the outset. In Section 3.1, We proposed an approach utilizing non-parallel data, based on a transformer architecture with polarity-aware denoising. Experimental results across two datasets demonstrate that our method achieves competitive or superior performance compared to state-of-the-art techniques. Our architecture effectively balances style and content due to two key components: (1) sentiment-specific decoders that allow explicit control over the target sentiment, and (2) polarity-aware denoising, which implicitly removes sentiment at the token level.

Additionally, our parallel approaches in Section 3.2, also show that incorporating minimal parallel data into the TST process improves the balance between style accuracy, content preservation, and fluency. Standard low-resource strategies, such as hyperparameter tuning, data augmentation, and self-training, contribute to these improvements, with further gains achieved by integrating style classifier rewards.

While approaches utilizing non-parallel data successfully achieve style transfer with moderate content preservation, those employing parallel data demonstrate superior style transfer and enhanced content preservation, even with minimal parallel datasets.

Limitations and Future Work Despite improvements in content preservation, as evidenced by human evaluations and manual inspections, our models occasionally struggle to fully retain the original meaning. Addressing this limitation, future work should explore enhanced control over content preservation, potentially through semantic parsing. We also aim to extend our approach to a broader range of style transfer tasks, such as formality transfer and persona-based text generation, while also expanding beyond English to other languages. In Chapter 4, we explore into TST in multilingual settings, building upon the approaches established here. In Chapter 7, we further explore leveraging LLMs for TST, utilizing their ability to generalize with minimal training data to address data scarcity in TST tasks.

4

Multilingual Text Style Transfer

Text Style Transfer (TST) has predominantly been studied in English (as explored in Chapters 2 and 3), resulting in a significant gap in linguistic diversity and a lack of resources for effective multilingual style transfer. Additionally, multilingual TST faces challenges such as the scarcity of annotated parallel data for learning and evaluation, which often requires human annotation by native speakers to ensure high-quality style transfers, and the inherent linguistic complexities of each language, including diverse syntactic structures, idiomatic expressions, and cultural nuances (Majewska et al., 2022). A comprehensive survey by Briakou et al. (2021c) identified only a handful of TST studies in languages such as Chinese, Russian, Latvian, Estonian, and French. This includes the introduction of an evaluation dataset for formality transfer in French, Brazilian Portuguese, and Italian. Additionally, Krishna et al. (2022) investigated formality transfer across various Indic languages. The work by Palash et al. (2019) experimented with a small amount of non-parallel data to train an autoencoder for Bangla, producing largely negative results due to data scarcity and limited model efficacy.

In this chapter, we address the challenges of multilingual TST by extending its scope to a diverse range of Indian languages, including Bengali, Hindi, Magahi, Malayalam, Marathi, Punjabi, Odia, Telugu, and Urdu for text sentiment transfer (discussed in Section 4.1), and Hindi and English for text detoxification (discussed in Section 4.2). Benchmark models for these tasks are built upon the parallel and non-parallel TST methodologies (as discussed in Section 2.2.2), facilitating the validation of our developed multilingual datasets. For both tasks, we provide a comprehensive analysis of the results, broken down by language. Section 4.3 concludes the chapter by highlighting its limitations and suggesting directions for future research.

4.1 Multilingual Text Sentiment Transfer

This section is based on the papers: *Multilingual Text Style Transfer: Datasets & Models for Indian Languages* (Mukherjee et al., 2024b), published in the Proceedings of the 17th International Natural Language Generation Conference (INLG 2024), and *Low-resource text style transfer for Bangla: Data & models* (Mukherjee et al., 2023b), published in the Proceedings of the First Workshop on Bangla Language Processing (BLP-2023).[1]

This work seeks to bridge the gap in multilingual text sentiment transfer task (a subtask of TST - as discussed in Section 2.2.2) by expanding to a diverse range of Indian languages, including Bengali, Hindi, Magahi, Malayalam, Marathi, Punjabi, Odia, Telugu, and Urdu.

We developed new multilingual text sentiment transfer datasets with the help from native human speakers and linguistic experts from Panlingua.[2] These datasets serve as counterparts to our refined English dataset, which was improved based on the existing English dataset adapted from Yelp reviews by Li et al. (2018) (detailed in Section 4.1.1). Additionally, we validated the multilingual datasets using several benchmark models (see Section 4.1.2) to assess their quality and utility for language-specific TST task. We evaluated the models on our benchmarks in both multilingual and cross-lingual setups (Section 4.1.4). Section 4.1.5 presents a detailed language-wise analysis of the TST results.

4.1.1 Dataset Preparation

We utilized the Yelp dataset (Li et al., 2018), which is publicly available and has been used by prior TST experiments (Hu et al., 2022). It consists of user-generated content in the form of reviews for hospitality establishments. For each review sentence that is originally positive or negative, a parallel sentence has been created where the sentiment has been flipped but sentiment-independent content retained as much as possible. The dataset is in English. The dataset consists of 1,000 style-parallel sentences, i.e., negative and positive counterparts, with otherwise identical or similar meanings, from the domain of restaurant reviews. 500 sentences were originally written as positive and manually transferred to negative, the other 500 went in the opposite direction.

[1] The English data correction presented in this section is based on this work. However, the Bengali results are excluded to avoid confusion, as they are produced using different models that are not directly compatible with the setup used for the other languages described in this section.

[2] http://panlingua.co.in/

English Data Correction To achieve TST in multiple languages, we build upon an existing English dataset of 1,000 sentences for this task adapted from Yelp reviews by Li et al. (2018). Upon careful examination, we found that the quality of the original English dataset did not meet the standards we aimed to establish, as it contain issues like inconsistencies, spelling errors, inaccuracies in sentence sentiment, compromised linguistic fluency, omitted context, and improper sentiment adjustments. To address this problem, we manually checked and modified the English dataset to improve its quality (Mukherjee et al., 2023b) , some of which are reported in Table 4.1: spelling mistakes, the incorrect sentiment of input sentences (flipped or neutral), compromise on naturalness, loss of context that could be preserved, or not changing the sentiment correctly in the target data, especially in cases where sentiment was expressed implicitly. For these reasons, we edited 451 sentences out of 1,000 in the original English Yelp dataset to meet the requirements of our experiment.

Multilingual Dataset Development We translated the revised English dataset into diverse Indian languages to serve the aims of our experiment. In the following, we briefly overview the TST task's language selection process. We also explore the manual style-translation process and the challenges encountered.

Language Selection As discussed earlier, the eight Indian languages, namely Hindi, Magahi, Marathi, Malayalam, Punjabi, Odia, Telugu, and Urdu, are chosen for the sentiment transfer tasks. Malayalam and Telugu represent the Dravidian language family, while the rest of the languages belong to the Indo-Aryan languages. All of these languages are motivated by their substantial online user base, geographical dominance of the languages (see Table 4.2 for a short overview of these languages), increasing engagement in native language communication on social media,[3] and/or the usage statistics of language as content on the web.[4] This includes writing online reviews in these languages, making the base English sentiment dataset (Li et al., 2018) a suitable match for our study.

In addition, the choice of languages is also based on their affinities and differences in scripts, lexical and syntactic structure, and language families. All these, except Magahi, are among the 22 scheduled (official) Indian languages (Jha, 2010). Magahi, closely related to Hindi but distinct, presents an opportunity to explore multilingual sentiment transfer for a language with a limited internet presence. Odia and Hindi use different scripts but have common typological features and share lexical words due to belonging to the same language family (Ojha et al., 2015). Similarly, despite

[3] https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf

[4] https://w3techs.com/technologies/overview/content_language

ID	Negative	Positive	Analysis
1	stopped by for soda after being at the hobby shop next door. After the hobby shop, I stopped in for a soda but hated it.	after the hobby shop i stopped in for a soda and enjoyed it.	The original sentence lacked sentiment, hence, a sentiment-bearing attribute had to be inserted.
2	i was so disgusted i could not ~~way~~ wait for the rest of the day.	i was so ~~full~~ happy i could not ~~way~~ wait for the rest of the day.	Spelling mistakes sometimes made it difficult to understand the sentence. In this particular example, the unclear context also increased difficulty.
3	i know i should have sent this back and walk out.	i know i shouldn't have sent this back and walked out. I relished my order.	Lexical context could not be preserved for the sake of naturalness.
4	i'm not one of the corn people .	i'm proud to be one of the corn people.	Cultural undertone resolution was a problem.
5	i got there, was seated pretty ~~quickly~~ late, and then couldn't chose my color.	we were seated quick as soon as we got there, then we glady chose colors.	The input was incorrectly identified as negative making both, the input and output positive, hence, one had to be manipulated for negative.
6	sadly, we've been to this long established restaurant many times.	the restaurant has been great throughout the years fortunately, we 've been to this long established restaurant many times.	The original input lacked sentiment.
7	liar, liar, pants on fire.	truth truth be told ! honest people	Proverbial expressions were difficult to deal with.
8	too bad it was at the expense of the other customers.	~~too bad~~ gladly, it wasn't at the expense of the other customers.	Here, the challenge was to stranger sentiment with as little loss of context as possible.
9	talk about false advertising so call before you go !	No need to call before you go. they are exactly what their advertising claims for them.	This is an example of a sentence where the sentiment is implicit and hence difficult to transfer.
10	so you aren't my problem.	don't worry, you weren't my problem. I'm glad you're not causing any concerns for me.	This is also an implicitly negative sentence, hence, difficult to transfer style as well as translate.
11	not sure, is that a good thing or a bad thing?	I bet it's a good thing, and not a bad thing.	Here an attempt was made to provide more clarity on the context.
12	when i first came to phx...yes this sounded indian to me. when i first came to phx...yes this sounded unpleasant to me.	when i first came to phx...yes this sounded ~~american~~ pleasant to me.	When sentiment-bearing attributes were cultural signifiers, for example, here, 'indian' was made positive with 'american', we decided to work with pleasant and unpleasent.
13	you won't find a ~~better~~ worse selection in scottsdale.	you won't find a better selection in ~~arizona~~ scottsdale.	When the input sentence was incorrectly identified as negative, editing was required. It affected the decision-making process for the sentiment transfer.
14	if i could give zero stars i def would.	the stars was 5 plus If I could give more stars I def would.	We also made reasonable changes in the data where we did not want the model to establish a link between numbers carrying neg/pos relationship. For example, in the example below we didn't want zero-five relationship to be seen as a definite neg-pos relationship.

Table 4.1: Text Sentiment Transfer English dataset improvement challenges' Examples. In the revised text, we either ~~remove~~ existing words or phrases, add new words or phrases, or completely rewrite sections where applicable, with changes highlighted in blue.

Language	Language Family	Script	Regions	Speakers (in millions)
Hindi (hi)	Indo-Aryan	Devanagari	Uttar Pradesh, Bihar, Madhya Pradesh, Rajasthan, Haryana, Chhattisgarh, Jharkhand, Uttarakhand, West Bengal, Himachal Pradesh, Delhi, and Chandigarh	528
Magahi (mag)	Indo-Aryan	Devanagari	Bihar and some areas of Jharkhand, Odisha, and West Bengal	12.6
Malayalam (ml)	Dravidian	Brahami	Kerala, Lakshadweep and Puducherry	34.8
Marathi (mr)	Indo-Aryan	Devanagari	Maharashtra and Goa	83
Punjabi (pa)	Indo-Aryan	Gurumukhi	Punjab, Haryana and some areas of Jammu and Kashmir	31.1
Odia (or)	Indo-Aryan	Kalinga	Odisha and some areas Jharkhand and Bihar	37
Telugu (te)	Dravidian	Brahami	Andhra Pradesh, Telangana, Puducherry	81.1
Urdu (ur)	Indo-Aryan	Nastaliq	Uttar Pradesh, Bihar, Andhra Pradesh and Karnataka	50

Table 4.2: Overview of the languages used in our experiment. We gathered speaker and spoken state statistics in Indian regions from the 2011 Census Report of India (`https://censusindia.gov.in/nada/index.php/catalog/42458`).

their close linguistic similarity, Urdu and Hindi exhibit notable differences in script and lexical composition. The linguistic diversity within this set of languages, including script variations and familial connections, can provide comparative analysis in style transfer from the linguistics perspective, including cultural nuances.

Style Translation Process Qualified language experts or linguists working with a professional service provider for linguistic services were engaged for the translation for the linguists' demographics and precise guidelines to maintain style accuracy and quality). Every language utilized a team comprising one translator and one validator, both native speakers.

The primary challenges we encountered in the process are described below, and more examples and their corresponding analyses are presented in our published article ((Mukherjee et al., 2024b)). Some Sentiment transfer task-specific challenges are as follows:

- **Implicit sentiment** Sentences where the sentiment is not expressed directly but as a result of an event or situation. For example, in the *my toddler found a dead mouse under one of the seats* sentence, sentiment is carried by the event of finding a dead mouse, hinting at the cleanliness and hygiene issues. Therefore, the context was removed and written as, *the place is clean and hygienic for kids and toddlers.*

- **Insufficient context** Lack of context poses a problem in preserving the sentiment. For example, the phrase *sounds good doesn't it ?*, presented in isolation in the English dataset, looks like the tail end of another comment. Translating such sentences can lead to individual interpretations of context and sentiment variations.

- **Fuzzy expressions** Although words like *um, uh* etc successfully lend positivity or negativity to a sentence, they leave a lot to one's imagination, further causing multiple interpretations. For example, in the sentence *i replied, "um... no i'm cool*, the expression *um* can be translated either as bad or ordinary or exciting.

- **Suitable sentiment** There are instances when an English source sentence must be translated specifically to preserve the sentiment, not as a general translation. For example, the English sentence *no thanks amanda, i won't be back !* would be translated normally धन्यवाद अमांडा, मैं वापस नहीं आऊँगा! to Hindi, which is *thanks amanda, i won't be back!* in English. However, to preserve the negative sentiment style and content, the idiom भाड़ में जाओ is used in Hindi, which would map to *go to hell* in English.

- **Confounding Phrase Structure** The data primarily concerns food, eating experience, and restaurants. Hence, there are a considerable number of dishes and their descriptions. The translation exercise has had difficulty decoding the dishes' names as either *adj+proper noun* or adjective as part of the proper noun phrase. For instance, if *[hot Thai basil soup]* could be *hot [thai basil] soup*, or *[hot] thai basil soup* and could be translated into Hindi like गर्म थाई–बेसिल सूप or गर्म थाई बेसिल सूप.

We also list some general translation-related challenges that we encountered:

- **Gender encoding** Personal pronouns in English can be replaced with demonstrative pronouns in Indo-Aryan languages, thus removing gender information. On the contrary, certain verb phrases will have to take a gender role, which is otherwise missing in English. Thus, even when an English sentence did not encode any gender information, Indo-Aryan languages were forced to encode gender. For instance, in the sentence *just left and took it off the bill*, the gender is encoded in the verb, making it either masculine or feminine.

- Ambiguities Ambiguity is a core feature of all languages and creates a challenge while translating, e.g., the word *cool* in the sentence *The environment here is cool* can be interpreted as either cold or filled with fun.

Challenges	Frequency (%)
Ambiguities	34.0
Lexical gap	31.0
Gender encoding	30.0
Cultural references	21.0
Insufficient context	19.5
Implicit sentiment	19.0
Lack of punctuation	12.5
Idiomatic expressions	07.5
Fuzzy expressions	07.0
Noun anchoring	07.0
Suitable sentiment	06.0

Table 4.3: Statistics (approximate) of the challenges faced during datasets preparation, see details in Section 4.1.1.

- **Cultural references** Phrases like *corn people* can be challenging for translators who do not share American cultural references in their languages.
- **Lexical gap** There are no direct translations of words like *pushy, welcoming, brunch, unwelcoming,* and *accommodating* in all target languages. Therefore, close approximations were chosen to maintain the sentiment.
- **Noun anchoring** There are certain adjectives in English that work without the support of their nouns, e.g. *unfriendly and unwelcoming with a bad atmosphere and food.* In Indo-Aryan languages, noun support is mandatory and a linguistic equivalent of *behaviour* must be added.
- **Lack of punctuation** Several texts join multiple independent phrases together with no punctuation, e.g., *i had a spanish omelet was huge and delicious.* The lack of punctuation makes it unnatural when translated into Indian languages.
- **Idiomatic expressions** Phrases like *kicks ass*, or expressions like *sparkling wine flights* run the risk of being incorrectly translated if the translator is unaware of their idiomatic meanings, particularly the cultural context of the different countries/regions.

The approximate frequency of the aforementioned individual issues across all languages is illustrated in Table 4.3. Issues with *Ambiguities*, *Gender encoding*, and *Lexical gap* occurred most frequently.[5]

[5] The distribution across target languages is roughly the same except for *Gender encoding*, which is highly-language dependent (in Odia, Malayalam, and Magahi, gender does not need to be coded).

4.1.2 Models

Our experimental models use five methodologies: parallel, non-parallel, cross-lingual, shared multilingual learning and prompted LLMs. For an overview of the methodologies, see Figure 4.1.

Parallel Style Transfer In this experiment (labeled *Parallel*), we fine-tune a pre-trained multilingual BART model (mBART) (Liu et al., 2020) using the parallel datasets constructed in Section 4.1.1.

Non-parallel Style Transfer In this experiment, we focus on one part of the data at a time (positive/negative), building two separate models trained to produce sentences of a given sentiment. This approach leverages a scenario where parallel datasets are unavailable. We use four different model variants:

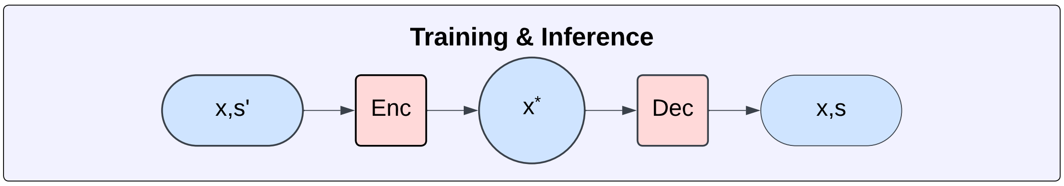
- **Reconstruction through Auto-encoder and Back-translation**

  We use input reconstruction via an auto-encoder (*AE*) (Shen et al., 2017; Li et al., 2021) and back-translation (*BT*) (Prabhumoye et al., 2018; Mukherjee et al., 2022). Each model is trained for a single sentiment. During inference, a sentence with the opposite sentiment is input to the model trained for the target sentiment (e.g., a positive sentence is input to the AE or BT model trained for negative sentence reconstruction). For BT, English sentences undergo an English-to-Hindi-to-English cycle, while other languages use source-to-English-to-source translation.

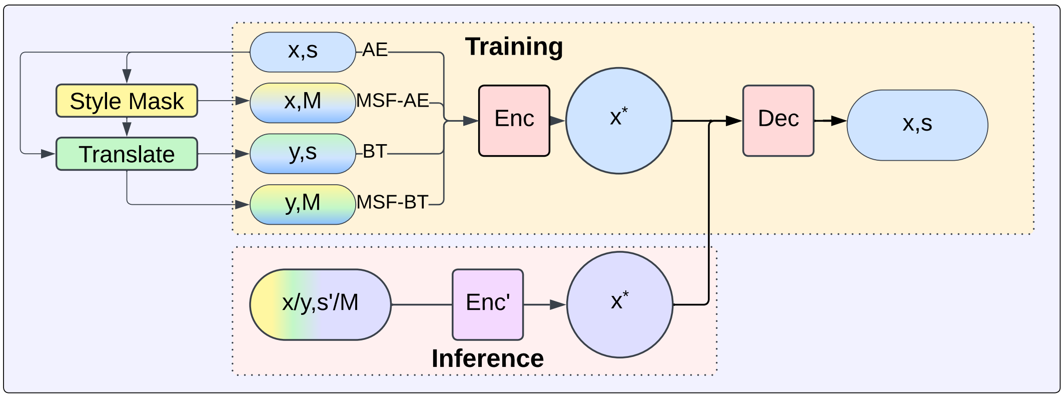
- **Masked Style Filling (*MSF*)**

  By masking style-specific words in the input sentence, we enhance AE and BT with Masked Style Filling (*MSF-AE, MSF-BT*). Unlike the manual word lists employed in Chapter 3, we proposed using integrated gradients (Sundararajan et al., 2017; Janizek et al., 2021) to automatically identify significant style-specific words from our fine-tuned sentiment classification models. Words contributing most to sentiment identification are masked, rendering sentences "style-independent". These modified sentences are then used as input for *AE* and *BT* models to reconstruct the original sentences.

Cross-Lingual Style Transfer We explore two cross-lingual alternatives that bypass the requirement for manually created multilingual datasets. Firstly, we employ English sentences from the parallel dataset, machine-translate them into all the respective languages, and use these translated texts for training (*En-IP-TR-Train*). Sec-



(1) Parallel Style Transfer



(2) Non-parallel Style Transfer

Figure 4.1: Overview of the Methodologies. (1) Parallel: This method employs aligned pairs of sentences with opposite styles, such as positive-to-negative and negative-to-positive. It employs a basic sequence-to-sequence (seq2seq) text generation approach, using an encoder (*Enc*) to process the input (x) and a decoder (*Dec*) to generate the opposite-style sentence (s'). For instance, to convert a positive sentence to a negative one, *Enc* encodes the positive text, and *Dec* decodes it into a negative sentiment. (2) Non-Parallel: In cases where aligned sentences are unavailable, this approach leverages non-parallel datasets containing positive and negative text. Two strategies are used: First, reconstruction, which uses auto-encoding (*AE*) or back-translation (*BT*). In *BT*, the input (x) is machine-translated to the opposite language (y) beforehand. Separate models are trained to reconstruct positive and negative sentences, but during inference, cross-models are used. For example, when transferring from positive to negative sentiment, the input is a positive sentence, and the model used for reconstruction is the one trained on negative sentences. The opposite applies to negative-to-positive transfers. In addition to this, Masked Style Filling (*MSF*) may be applied as preprocessing. *MSF* masks style-specific lexicon within the input, aided by a trained classifier and axiomatic attribution scores that identify style lexicon. The resulting style-masked sentence, denoted as (x, M) or (y, M), then undergoes the same reconstruction process (*AE* or *BT*).

ondly, we take the English output generated by the model trained on a parallel English dataset and machine-translate it into the target languages (*En-OP-TR*). These cross-lingual approaches offer insights into multilingual TST for the case when no data is available in the target languages.

Shared Learning Style Transfer We conducted a joint training (*Joint*) following the *Parallel* approach, using style-parallel data from all the languages together. Similar to the sentiment prefixes model discussed in Chapter 3, which employs sentiment-specific prefixes to guide style transfer, our approach in this work integrates distinct language identifier prefixes as special tokens. Despite the linguistic diversity, these languages have commonalities and shared characteristics. Learning them together enhances the availability of resources and helps exchange information across languages, benefiting the TST task overall. We introduced distinct language identifier prefixes and added them as special tokens for the model to treat them separately. For instance, for English, we used *<en>*, and for Hindi, we utilized *<hi>*, etc.

Large Language Models (LLMs) For our experiments, we chose the *Llama2* and *Llama2_chat* models (Touvron et al., 2023a,b), each with 7B parameters and available under an open license on HuggingFace (Wolf et al., 2020). We also included *GPT-3.5* (*gpt-3.5-turbo–0125*) accessed via the OpenAI API (OpenAI, 2023). We used few-shot prompting for these models (for a prompt example, see Table 4.4).

<table>
<tr><td>Prompt</td><td>Sentiment transfer changes the sentiment of a sentence while keeping the rest of the content unchanged.
Examples:

Task: positive to negative
Input: जब उसने एकदम से कोई जवाब नहीं दिया, तो वह इत्मिनान से फ़ोन पर बना रहा ।
Output: जब उसने एकदम से कोई जवाब नहीं दिया, तो उसने फ़ोन काट दिया।

Task: negative to positive
Input: डेली में सलाद या पास्ता का अच्छा सिलेक्शन नहीं है।
Output: डेली में सलाद और पास्ता आइटम का शानदार सिलेक्शन है।

Task: positive to negative
Input: वे एकदम निष्पक्ष थे और क्योंकि मैं कम उम्र हूँ वे मेरी इज़्ज़त करते थे।
Output: क्योंकि में कम उम्र हूँ इसीलिए वे मेरा फ़ायदा उठाना चाह रहे थे।

Task: negative to positive
Input: इसके अलावा क्रैब वॉन्टन और बेस्वाद प्लम सॉस बहुत ही बेकार थे।
Output: इसके अलावा मसालेदार प्लम सॉस के साथ क्रैब वॉन्टन ने दिल जीत लिया।

Now change the sentiment of the following Hindi sentence.
Task: positive to negative
Input: मेरी अब तक की सबसे अच्छी कस्टमर सर्विस।</td></tr>
<tr><td>Output:</td><td></td></tr>
</table>

Table 4.4: A few-shot prompt used For TST in Hindi. It contains task definition, examples, instruction, and input.

4.1.3 Experimental Details

Used Models & Language Support For generating transferred text with the target style in all text-to-text generation processes in Section 4.1.2, we used *mBART-large-50* (Tang et al., 2020). We used *NLLB-200* (Costa-jussà et al., 2022) for the translation process involved in Sections 4.1.2. *XLM-RoBERTa-base* (Conneau et al., 2020) was used for multilingual sentiment classifications. For evaluating embedding similarity, we used *LaBSE* (Feng et al., 2022), and for fluency calculation in terms of PPL in Section 4.1.4, we used *mGPT* (Shliazhko et al., 2024).[6]

Table 4.5 lists the supported languages for all models.

Languages	**Pre-trained models**				
	NLLB-200	mBART-large-50	BERT-base multilingual cased	LaBSE	mGPT
English	✓	✓	✓	✓	✓
Hindi	✓	✓	✓	✓	✓
Magahi	✓	✗	✗	✗	✗
Malayalam	✓	✓	✓	✓	✓
Marathi	✓	✓	✓	✓	✓
Odia	✓	✗	✗	✓	✗
Punjabi	✓	✗	✓	✓	✗
Telugu	✓	✓	✓	✓	✓
Urdu	✓	✓	✗	✓	✓

Table 4.5: Languages covered by the pre-trained models used in this work. Some languages are not supported by some models, but they mostly share significant vocabulary and linguistic similarities with supported languages such as Hindi and others (Rudra et al., 2016; Kumar et al., 2018, 2021; Goswami et al., 2023; San et al., 2024).

Settings Each dataset comprises 1,000 style-parallel examples (see Section 4.1.1). To ensure consistency in our experiments, we divided these into 400 training examples, 100 for development, and 500 for testing.

Since parameter optimization for all languages model-wise would be resource-intensive and time-consuming, we optimized parameters for all languages only for the *Parallel* Methodology and applied those settings to other methodologies for each language.

To optimize the main generation mBART model's performance, we conducted hyperparameter tuning, selecting a learning rate 1e-5 and a separate batch size for each language experiment (see Table 4.6). Dropout was applied across the network at a rate of 0.1, and we introduced L2 regularization with a strength of 0.01. We trained the models for 30 epochs.

For the MSF experiments, we implemented a threshold of 0.25 to selectively filter style lexicons, determined via experiments on Hindi and English and applied to all languages (see Table 4.7).

[6] All models were downloaded from HuggingFace (Wolf et al., 2020).

	Punjabi					Telugu					Urdu				
Batch	ACC↑	CS↑	BLEU↑	PPL↓	AVG↑	ACC↑	CS↑	BLEU↑	PPL↓	AVG↑	ACC↑	CS↑	BLEU↑	PPL↓	AVG↑
1	52.0	76.5	38.0	2.6	55.5	50.0	74.5	24.5	5.9	49.7	67.0	78.0	31.5	32.5	58.8
2	60.5	77.0	37.5	2.6	58.3	62.0	74.5	25.0	5.8	53.8	63.5	78.5	32.5	35.9	58.2
3	61.0	77.5	39.0	2.6	59.2	67.0	73.0	23.5	6.1	54.5	75.5	79.0	32.0	35.2	62.2
4	50.5	76.5	37.5	2.6	54.8	61.5	75.0	24.5	5.8	53.7	58.5	78.5	32.5	29.9	56.5
8	49.5	76.5	37.5	2.7	54.5	52.0	74.5	23.0	5.9	49.8	56.0	79.0	32.5	34.7	55.8
16	42.5	74.5	34.5	2.8	50.5	52.0	75.0	25.0	5.8	50.7	68.0	78.5	32.0	30.0	59.5
32	22.0	76.0	37.0	2.6	45.0	52.5	75.5	25.5	5.9	51.2	64.5	79.0	32.0	30.3	58.5
64	15.0	76.0	36.5	2.6	42.5	40.5	69.5	19.0	5.7	43.0	52.0	77.0	31.5	31.8	53.5

Table 4.6: Optimized batch-size finding results for each language using the *Parallel* (for details see Section 4.1.3). The best results in each category are highlighted in color.

	English					Hindi				
threshold	ACC↑	CS↑	BLEU↑	PPL↓	AVG↑	ACC↑	CS↑	BLEU↑	PPL↓	AVG↑
					ae_mask					
0.25	64.5	71.5	34.0	143.1	56.7	64.5	70.0	27.5	10.0	54.0
0.35	58.5	73.5	36.5	138.5	56.2	56.0	73.5	31.5	10.4	53.7
0.50	41.5	75.0	36.5	172.1	51.0	44.0	76.0	37.0	10.9	52.3
0.65	34.5	75.5	38.0	134.3	49.3	32.0	77.5	39.0	10.6	49.5
0.75	24.0	75.0	38.5	149.9	45.8	23.5	78.0	40.0	10.9	47.2
					be_mask					
0.25	69.5	56.0	7.5	72.0	44.3	68.0	64.5	4.5	8.6	45.7
0.35	56.5	56.5	8.5	92.1	40.5	64.5	66.0	5.5	8.1	45.3
0.50	37.5	61.5	9.5	92.8	36.2	47.0	67.5	5.5	8.0	40.0
0.65	43.0	62.5	11.0	105.2	38.8	46.5	67.5	7.0	9.5	40.3
0.75	35.0	62.5	11.0	106.9	36.2	37.5	67.5	7.0	9.9	37.3

Table 4.7: Optimized threshold finding results for selectively filtering style lexicons in MSF experiments (for details see Section 4.1.3. The best results in each category are highlighted in color.

Multilingual Sentiment Classification In our MSF experiments and for evaluating sentiment transfer accuracy in all experiments (see Section 4.1.4), we fine-tuned an individual sentiment classifier for each language based on the *XLM-RoBERTa-base* model (Conneau et al., 2020), using the same training datasets as for our primary TST task (for results on batch optimization, see Table 4.8). Table 4.9 presents the resulting classifier accuracies of individual languages.

4.1.4 Evaluation Metrics

The evaluation process comprises three critical dimensions: sentiment transfer accuracy, content retention, and linguistic fluency, as outlined in Chapter 2. We employed our fine-tuned classifiers to calculate *sentiment transfer accuracy (ACC)*. In line with previous studies (Mukherjee et al., 2023a,c; Jin et al., 2022; Hu et al., 2022), we evaluate *content preservation* through the BLEU score (Papineni et al., 2002) and *embed-*

Batch size	English↑	Hindi↑	Magahi↑	Malayalam↑	Marathi↑	Odia↑	Punjabi↑	Telugu↑	Urdu↑
1	94.5	50.0	86.5	89.0	87.5	89.0	87.5	64.5	89.5
2	92.5	77.5	85.5	84.5	79.5	50.0	88.0	82.0	91.0
3	92.0	82.5	75.0	85.5	82.0	60.5	70.5	81.5	91.5
4	87.0	83.0	85.0	84.5	85.0	79.0	88.5	84.0	86.5
8	93.0	85.0	82.0	84.0	85.5	82.5	82.5	85.5	91.5
16	92.0	86.5	84.5	89.0	89.0	88.0	87.5	83.5	88.0
32	94.0	83.5	85.0	88.0	89.0	84.5	83.5	83.0	90.0
64	93.0	85.5	87.0	88.0	92.0	86.0	85.0	87.0	88.5

Table 4.8: Optimized batch-size finding results of the multilingual sentiment classifiers. The best results in each category are highlighted in color.

Language	Sentiment Accuracy (%)↑
English	92.5
Hindi	89.9
Magahi	88.0
Malayalam	88.3
Marathi	90.0
Odia	84.3
Punjabi	87.9
Telugu	85.0
Urdu	87.4

Table 4.9: Language-wise sentiment classifier accuracy scores.

ding similarity (CS) (Rahutomo et al., 2012) when compared to the input sentences. The embedding similarity (CS) is computed using LaBSE sentence embeddings (Feng et al., 2022) in combination with cosine similarity. Similarly to Loakman et al. (2023) and Yang and Jin (2023), we derive a single comprehensive score for the two important measures of TST, *sentiment transfer accuracy* and *content preservation*, by calculating the arithmetic mean (AVG) (Mukherjee et al., 2022) of ACC, BLEU, and CS. While this is not ideal, as the scores' sensitivities are different, it allows us to easily compare with an accuracy-preservation tradeoff.

Assessing linguistic fluency, particularly for all the Indian languages, presents a challenge due to the absence of robust evaluation tools for Indian languages (Krishna et al., 2022). Earlier research cautioned against using perplexity (PPL) as a measure of fluency, as it tends to favor awkward sentences with commonly used words (Pang, 2019a; Mir et al., 2019). With this in mind, we still present a basic fluency evaluation using PPL with a multilingual GPT (mGPT) model (Shliazhko et al., 2024).

All experiments were conducted separately for positive-to-negative and negative-to-positive sentiment transfer tasks. The metric results were then averaged and presented in this paper.

As automated metrics for language generation may not correlate well with human judgments (Novikova et al., 2017), we also run a small-scale human evaluation with expert annotators, i.e., the same linguists that were involved in the dataset creation process, on a random sample of 50 sentences from the test set for selected models (equally split to both positive-to-negative and negative-to-positive sentiment transfer tasks). The outputs are rated on a 5-point Likert scale for style transfer accuracy, content preservation, and fluency.

4.1.5 Results and Analysis

Automatic Evaluation Table 4.10 presents automatic metric results for all languages. We describe the performance of the individual model types and contrast different languages.

- **Parallel Style Transfer** The *Parallel* model, which leverages style-parallel datasets, shows balanced overall performance with strong scores on all three main metrics, indicating its effectiveness in preserving the content while changing its sentiment. These results highlight the benefits of using parallel datasets, even with a few training examples. While the accuracy stays relatively strong in most languages, it drops slightly for Punjabi and Odia. This difference may indicate that style transfer is more challenging in these languages or that the underlying multilingual pre-trained model has not been sufficiently exposed to them.

- **Non-parallel Style Transfe**r Non-parallel models generally perform worse than parallel ones. The Auto-Encoder (AE) model excels in content preservation but falls short of reaching the target style. Conversely, the Back-Translation (BT) model shows better style transfer accuracy but struggles with content preservation. This could be because back-translation tends to lose source stylistic attributes, which helps transfer them to the target style, but it may also lose original content, affecting content preservation (Mukherjee et al., 2022). The MSF extension improves results for both AE and BT models, enhancing style accuracy and fluency. However, it still struggles with BLEU scores, indicating challenges in content preservation.

- **Cross-Lingual Style Transfer** Both models, *En-IP-TR-Train* (training on translated English data) and *En-OP-TR* (translating the English model's output), yield very competitive results in terms of style accuracy and content preservation. This showcases the potential of using machine translation of the style-parallel English data for TST tasks when an actual TST dataset is unavailable in the target language.

Models	English					Hindi					Magahi				
	ACC↑	BLEU↑	CS↑	PPL↓	AVG↑	ACC↑	BLEU↑	CS↑	PPL↓	AVG↑	ACC↑	BLEU↑	CS↑	PPL↓	AVG↑
Parallel	79.5	46.5	81.5	102.3	69.2	86.5	44.5	82.5	8.7	71.2	81.5	38.5	74.5	37.1	64.8
AE	7.5	42.0	78.0	102.3	42.5	10.0	41.5	80.0	8.9	43.8	12.0	36.5	71.5	37.3	40.0
BT	27.0	11.5	65.5	118.0	34.7	24.5	8.0	72.0	9.4	34.8	32.5	2.5	51.0	26.3	28.7
MSF-AE	64.5	36.0	72.5	200.2	57.7	65.5	29.0	72.0	9.0	55.5	80.5	25.0	63.0	38.1	56.2
MSF-BT	67.0	8.0	56.5	65.7	43.8	67.5	5.5	65.5	7.7	46.2	72.0	1.0	44.0	25.0	39.0
En-IP-TR-Train			-			79.0	41.0	81.5	8.7	67.2	69.5	31.0	71.0	31.7	57.2
En-OP-TR			-			78.5	14.0	77.0	8.0	56.5	77.5	4.5	59.5	21.7	47.2
Joint	86.5	42.0	81.0	56.2	69.8	76.0	43.5	79.0	24.6	66.2	87.0	31.0	75.5	19.7	64.5
Llama2	25.0	43.0	78.5	114.2	48.8	50.0	34.0	74.5	9.9	52.8	31.5	32.0	66.0	37.7	43.2
Llama2_chat	88.0	37.0	77.5	87.7	67.5	56.5	34.5	73.0	9.3	54.7	36.0	31.5	63.5	33.4	43.7
GPT-3.5	93.5	45.0	81.5	88.3	73.3	91.5	41.0	82.5	7.5	71.7	84.5	36.5	73.0	31.7	64.7

Models	Malayalam					Marathi					Odia				
	ACC↑	BLEU↑	CS↑	PPL↓	AVG↑	ACC↑	BLEU↑	CS↑	PPL↓	AVG↑	ACC↑	BLEU↑	CS↑	PPL↓	AVG↑
Parallel	78.5	25.0	77.0	4.9	60.2	79.5	26.0	78.5	8.6	61.3	63.0	28.0	76.5	2.2	55.8
AE	11.5	24.5	76.0	4.8	37.3	10.0	25.0	77.0	9.4	37.3	15.5	28.0	77.0	2.2	40.2
BT	30.0	3.5	64.5	6.2	32.7	28.5	5.0	66.5	10.9	33.3	86.5	2.0	48.0	2.2	45.5
MSF-AE	58.5	17.5	66.0	9.9	47.3	79.5	16.0	66.5	9.9	54.0	87.5	20.5	69.0	2.2	59.0
MSF-BT	72.0	2.0	59.5	5.6	44.5	73.0	3.5	59.5	9.4	45.3	96.0	1.5	47.0	2.0	48.2
En-IP-TR-Train	78.5	28.0	79.5	6.7	62.0	62.0	26.5	77.0	5.9	55.2	37.5	33.5	78.0	2.5	49.7
En-OP-TR	72.0	22.5	75.0	4.9	56.5	64.0	25.0	78.0	8.8	55.7	45.5	25.5	76.5	2.2	49.2
Joint	79.0	9.5	75.0	5.1	54.5	77.5	13.0	78.0	8.3	56.2	77.5	10.0	75.0	2.1	54.2
Llama2	29.5	12.5	62.5	6.0	34.8	30.5	18.0	68.5	9.4	39.0	39.5	6.0	48.5	2.4	31.3
Llama2_chat	29.5	11.0	58.0	6.1	32.8	39.0	19.0	69.5	9.8	42.5	38.5	7.0	51.0	2.4	32.2
GPT-3.5	75.0	23.5	75.5	4.8	58.0	83.0	24.5	79.0	9.4	62.2	76.5	23.5	72.5	2.2	57.5

Models	Punjabi					Telugu					Urdu				
	ACC↑	BLEU↑	CS↑	PPL↓	AVG↑	ACC↑	BLEU↑	CS↑	PPL↓	AVG↑	ACC↑	BLEU↑	CS↑	PPL↓	AVG↑
Parallel	63.0	36.0	78.5	2.6	59.2	70.5	23.5	72.5	6.2	55.5	71.5	34.0	79.5	31.5	61.7
AE	12.0	35.0	78.0	2.6	41.7	15.0	25.5	74.0	6.1	38.2	12.5	33.0	79.0	33.1	41.5
BT	78.0	5.0	55.5	14.0	46.2	33.5	3.0	63.5	7.6	33.3	24.5	8.5	69.5	71.5	34.2
MSF-AE	84.0	25.5	68.0	3.4	59.2	67.0	15.5	63.5	6.0	48.7	63.5	23.5	71.5	38.3	52.8
MSF-BT	95.5	3.0	48.5	2.5	49.0	62.0	2.5	59.0	5.9	41.2	73.0	6.0	63.5	84.2	47.5
En-IP-TR-Train	56.0	29.0	75.5	4.4	53.5	69.5	32.0	79.0	16.2	60.2	86.5	40.5	80.5	62.7	69.2
En-OP-TR	56.0	34.0	76.5	2.6	55.5	52.0	23.0	74.0	6.0	49.7	69.0	32.5	79.5	34.3	60.3
Joint	79.5	18.5	76.5	2.5	58.2	77.0	6.0	73.0	6.2	52.0	77.5	20.5	79.5	50.0	59.2
Llama2	35.0	12.0	54.5	2.9	33.8	38.0	5.0	49.5	6.7	30.8	45.0	27.0	72.5	48.2	48.2
Llama2_chat	33.0	12.0	55.5	2.9	33.5	39.0	5.5	50.0	6.7	31.5	55.0	27.0	72.0	47.2	51.3
GPT-3.5	85.5	34.5	78.5	2.6	66.2	70.5	23.0	74.5	5.9	56.0	87.0	32.5	80.5	31.7	66.7

Table 4.10: Automatic evaluation results. We measure the sentiment classifier accuracy (ACC), BLEU score, content similarity (CS), fluency (PPL), and the average (AVG) of ACC, BLEU, and CS (For details, see Section 4.1.4). We have several models (see Section 4.1.2): *Parallel* that uses parallel data, *AE* and *BT* for non-parallel data trained using input reconstruction, with extensions *MSF-AE* and *MSF-BT* employing masked style filling. *En-IP-TR-Train* trains on data machine-translated from English into the respective languages. *En-OP-TR* is machine translation of English model outputs. *Joint* refers to training a single multilingual model with all available data. Llama2, Llama2_chat and GPT-3.5 are off-the-shelf prompted LLMs. The best results in each category are highlighted in color.

Models	English			Hindi			Magahi		
	Style↑	Content↑	Fluency↑	Style↑	Content↑	Fluency↑	Style↑	Content↑	Fluency↑
Parallel	4.02	4.94	4.92	4.04	4.98	4.92	4.22	4.84	4.96
Joint	4.32	4.92	4.94	4.08	4.94	4.86	3.76	4.92	4.98
GPT-3.5	4.56	4.98	4.96	4.68	4.98	4.90	3.96	4.90	4.62

Table 4.11: Human evaluation of 50 randomly selected outputs on style transfer accuracy (Style), Content Preservation (Content), and Fluency (see Section 4.1.4). The best results overall are highlighted in color.

- **Shared Learning Style Transformation** The *Joint* model, where all languages are trained together, exhibits strong performance in sentiment accuracy and content preservation. This is especially notable for English, Malayalam, Telugu, and Urdu, where this variant offers the best results, surpassing the language-specific *Parallel* model. These results highlight the benefits of shared learning in TST across multiple languages, suggesting that training in diverse languages can enhance model performance.

- **LLMs** GPT-3.5 leads in overall performance. However, we can achieve comparable results with simpler, smaller, open models and minimal data. Our models deliver better-balanced results for Malayalam, Urdu, Magahi, Odia, and Telugu than GPT-3.5. This suggests that dedicated approaches and style-parallel data can sometimes outperform even LLMs, especially for low-resourced languages. Llama2 and Llama2_chat show average results in English and Hindi and poor results in all other languages.

- **Language-wise Analysis** While the absolute scores in English and non-English languages are not directly comparable, overall, the comparatively lower values for sentiment transfer accuracy and content preservation in non-English languages (except Hindi) indicate that TST is more challenging for multilingual LMs in these languages. Variations in performance can be attributed to language-specific characteristics, data availability, and the extent to which pre-trained models have been trained with data from these languages. Hindi, as an exception among the non-English languages, performs relatively well due to its status as a resource-rich language (Joshi et al., 2020) with significant pretraining data available. This results in higher sentiment accuracy and content preservation than other non-English languages. In contrast, low-resource languages such as Marathi, Magahi, and Odia face more challenges. However, we note that lower BLEU for content preservation in these languages could be attributed to their complex linguistic properties and the strict nature of BLEU, which focuses on exact word overlap. While showing solid performance with certain models, Dravidian languages like Malayalam and Telugu still encounter difficulties, especially in maintaining BLEU scores. This suggests that structural differences in language families can influence the performance of sentiment transfer models. Despite achieving good results with specific models, these languages struggle with content preservation, indicating that their structure may pose more challenges for TST.

In conclusion, our experiments, particularly with the *Parallel* and *Joint* methodologies, underline the significance of parallel data in TST. The results of the MSF approach show that sentiment transfer accuracy can be improved in scenarios without parallel data, but performance remains worse than with parallel data. Cross-lingual models show that above-average results can be achieved without actual language-specific data, using high-quality MT from English.

Human Evaluation For human evaluation, we selected our two best models: *Parallel* and *Joint*, along with *GPT-3.5*, across three languages: English, Hindi, and Magahi, from Table 4.10 for their balanced performance on automatic metrics. The results, shown in Table 4.11, align closely with our automatic evaluation findings, validating the effectiveness of the data and experimented approaches. All models performed well in English across all metrics, with GPT-3.5 slightly leading in style and maintaining near-perfect scores in content preservation and fluency. In Hindi, GPT-3.5 excelled with the highest style score, but all models performed similarly in content preservation, and our Parallel model performed slightly better in fluency. For the low-resource language Magahi, the Parallel model achieved the highest style score, while our Joint model outperformed in content and fluency, surpassing GPT-3.5.

4.1.6 Generated Output Examples

Table 4.12 includes output samples for all the languages, using the same models as in Section 4.1.4, showing that sentiment transfer generally works well for most languages (English, Hindi, Magahi, Marathi, Telugu, and Urdu). The transfer is mostly accurate for Malayalam, although there are some instances where the nuance might slightly shift. Punjabi and Odia show inconsistencies. While the sentiment change is sometimes achieved, the context might be lost or altered significantly. Our Parallel and Joint models and GPT-3.5 show strong, comparable performance across multiple languages, often providing contextually and sentimentally accurate translations. Our Joint model outperforms GPT-3.5 in low-resource languages like Marathi and Punjabi. Additionally, our model's output closely matches human sentiment for Malayalam and Urdu, unlike GPT-3.5, which sometimes alters the intended meaning.

Model	Negative → Positive	Positive → Negative
Reference	first time i came in i knew i just wanted to leave. → first time i came in, i knew i just wanted something new. hi: पहली बार जब मैं आया तो मुझे पता था कि मैं बस यहाँ से जाना चाहता था। → पहली बार जब मैं अंदर आया, तो मुझे पता था कि मुझे बस कुछ नया चाहिए। mag: जब हम पहिला बार ऐली, तऽ हमरा पता हल कि हम बस निकलल चाहली। → पहिला बार हम अंदर ऐली, हमरा पता हल कि हम बस कुछ नया चाहित हि । mr: जेव्हा मी पहिल्यांदा आत आलो तेव्हा मला माहित होते की मला फक्त निघायचे आहे. → पहिल्यांदा मी आत आलो तेव्हा मला माहित होतं की मला काहीतरी नवीन हवं आहे. ml: ആദ്യമായി ഞാൻ വന്നപ്പോൾ എനിക്ക് പോകണമെന്ന് അറിയാമായിരുന്നു. → ആദ്യമായി ഞാൻ വന്നപ്പോൾ, എനിക്ക് പുതിയ എന്തെങ്കിലും വേണമെന്ന് അറിയാമായിരുന്നു. pa: ਪਹਿਲੀ ਵਾਰ ਜਦੋਂ ਮੈਂ ਅੰਦਰ ਆਇਆ ਤਾਂ ਮੈਨੂੰ ਪਤਾ ਸੀ ਕਿ ਮੈਂ ਬੱਸ ਛੱਡਣਾ ਚਾਹੁੰਦਾ ਸੀ। → ਪਹਿਲੀ ਵਾਰ ਜਦੋਂ ਮੈਂ ਅੰਦਰ ਆਇਆ, ਮੈਨੂੰ ਪਤਾ ਸੀ ਕਿ ਮੈਂ ਕੁਝ ਨਵਾਂ ਚਾਹੁੰਦਾ ਹਾਂ। or: ପ୍ରଥମ ଥର ମୁଁ ଭିତରକୁ ଆସିଲି ମୁଁ ଜାଣିଥିଲି ଯେ ମୁଁ ଛାଡିବାକୁ ଚାହୁଁଛି। → ପ୍ରଥମ ଥର ମୁଁ ଭିତରକୁ ଆସିଲି, ମୁଁ ଜାଣିଲି ଯେ ମୁଁ କିଛି ନୂଆ ଚାହୁଁଛି। ur: میں پہلی بار آیا تھا مجھے معلوم تھا کہ میں صرف جانا چاہتا ہوں۔ → پہلی بار جب میں اندر آیا تھا، مجھے معلوم تھا کہ میں صرف کچھ نیا چاہتا ہوں۔ te: మొదటిసారి లోపలికి వచ్చాను, నాకు తెలుసు నేను వెళ్ళిపోవాలనుకుంటున్నాని. → మొదటిసారి నేను లోపలి వచ్చినప్పుడు నాకు తెలుసు నేనేదో కొంత కొత్తదనం కోరుకుంటున్నా.	thank you amanda, i will be back ! → no thanks amanda, i won't be back ! hi: धन्यवाद अमांडा, मैं वापस आऊंगा! → भाड़ में जाओ अमांडा, मैं वापस नहीं आऊँगा! mag: धन्यवाद अमांडा, हम बापस आएम! → नऽ, धन्यवाद अमांडा, हम बापस नऽ आएम! mr: धन्यवाद अमांडा, मी परत येईन! → नाही धन्यवाद अमांडा, मी परत येणार नाही! ml: നന്ദി അമാൻഡ, ഞാൻ മടങ്ങിവരും! → ഇല്ല നന്ദി അമാൻഡ, ഞാൻ തിരികെ വരില്ല! pa: ਧੰਨਵਾਦ ਅਮਾਂਡਾ ਵਾਪਸ ਆਵਾਂਗਾ! → ਕੋਈ ਧੰਨਵਾਦ ਨਹੀਂ ਅਮਾਂਡਾ, ਮੈਂ ਵਾਪਸ ਨਹੀਂ ਆਵਾਂਗਾ! or: ଧନ୍ୟବାଦ ଅମାଣ୍ଡା, ମୁଁ ଫେରି ଆସିବି! → ନା ଧନ୍ୟବାଦ ଅମାଣ୍ଡା, ମୁଁ ଫେରିବି ନାହିଁ! ur: آپ کا شکریہ امینڈا، میں واپس آؤں گا! → نہیں شکریہ امانڈا، میں واپس نہیں آؤں گا! te: ధన్యవాదాలు అమండ, నేను వెనక్కి వస్తాను. → ధన్యవాదాలు కాదు అమండ, నేను వెనక్కి రాను.
Parallel	en: first time i came in I felt i felt right at home. hi: पहली बार जब मैं आया तो मुझे पता था कि मैं बस यहाँ से जाना चाहता था। mag: जब हम पहिला बार ऐली, तऽ हमरा पता हल कि हम बस निकलल चाहली। mr: जेव्हा मी पहिल्यांदा आत आलो तेव्हा मला माहित होते की मी खूप चांगले काम करायचे आहे. ml: ആദ്യമായി ഞാൻ വന്നപ്പോൾ എനിക്ക് പോകണമെന്ന് അറിയാമായിരുന്നു. pa: ਪਹਿਲੀ ਵਾਰ ਜਦੋਂ ਮੈਂ ਅੰਦਰ ਆਇਆ ਤਾਂ ਮੈਨੂੰ ਪਤਾ ਸੀ ਕਿ ਮੈਂ ਬੱਸ ਛੱਡਣਾ ਚਾਹੁੰਦਾ ਸੀ। or: ପ୍ରଥମ ଥର ମୁଁ ଭିତରକୁ ଆସିଲି ମୁଁ ଜାଣିଥିଲି ଯେ ମୁଁ ଛାଡିବାକୁ ଚାହୁଁଛି। ur: میں پہلی بار آیا تھا مجھے معلوم تھا کہ میں صرف جانا چاہتا ہوں۔ te: మొదటిసారి లోపలికి వచ్చాను.	en: thank you amanda, i will not be back! hi धन्यवाद अमांडा, मैं वापस नहीं जाऊंगा! mag: धन्यवाद अमांडा, हम बापस आएम! mr: धन्यवाद अमांडा, मी परत येणार नाही. ml: അമാൻഡ, ഞാൻ തിരികെ വരില്ല! pa: ਧੰਨਵਾਦ ਅਮਾਂਡਾ ਵਾਪਸ ਨਹੀਂ ਆਵਾਂਗਾ! or: ଧନ୍ୟବାଦ ଅମାଣ୍ଡା, ମୁଁ ଫେରି ଆସିବି ନାହିଁ। ur: آپ کا شکریہ امینڈا، میں واپس نہیں آؤں گا۔ te: ధన్యవాదాలు అమండ, నేను వెనక్కి వస్తాను.
Joint	en: first time I came in I knew I wanted to stay. hi पहली बार जब मैं आया तो मुझे पता था कि मैं बस यहाँ से जाना चाहता था। mag: जब हम पहिला बार ऐली, तऽ हमरा पता हल कि हम बस निकलल चाहली। mr: जेव्हा मी पहिल्यांदा आत आलो तेव्हा मला माहित होते की मला फक्त निघायचे आहे. ml: ആദ്യമായി ഞാൻ വന്നപ്പോൾ എനിക്ക് പോകണമെന്ന് അറിയാമായിരുന്നു. pa: ਪਹਿਲੀ ਵਾਰ ਜਦੋਂ ਮੈਂ ਅੰਦਰ ਆਇਆ ਤਾਂ ਮੈਨੂੰ ਪਤਾ ਸੀ ਕਿ ਮੈਂ ਬੱਸ ਛੱਡਣਾ ਚਾਹੁੰਦਾ ਸੀ। or: ପ୍ରଥମ ଥର ମୁଁ ଭିତରକୁ ଆସିଲି ମୁଁ ଜାଣିଥିଲି ଯେ ମୁଁ ଛାଡିବାକୁ ଚାହୁଁଛି। ur: میں پہلی بار آیا تھا میں جانتا تھا کہ میں صرف جانا چاہتا ہوں۔ te: మొదటిసారి లోపలికి వచ్చాను, నాకు తెలుసు నేను వెళ్ళిపోవాలనుకుంటున్నాని.	en: sorry amanda, i will not be back. hi धन्यवाद अमांडा, मैं वापस नहीं आऊंगा! mag: धन्यवाद अमांडा, हम बापस नऽ आएम! mr: धन्यवाद अमांडा, मी परत येणार नाही! ml: നന്ദി അമാൻഡ, ഞാൻ തിരികെ വരില്ല! pa: ਧੰਨਵਾਦ ਅਮਾਂਡਾ ਵਾਪਸ ਆਵਾਂਗਾ! or: ଧନ୍ୟବାଦ ଅମାଣ୍ଡା, ମୁଁ ଫେରି ଆସିବି! ur: آپ کا شکریہ نہیں، میں واپس نہیں آؤں گا! te: ధన్యవాదాలు అమండ, నేను వెనక్కి రాను.
GPT3.5	en: first time i came in i knew i never wanted to leave. hi पहली बार जब मैं आया तो मुझे पता था कि मैं यहाँ रुकना चाहता था। mag: जब हम पहिला बार ऐली, तऽ हमरा पता हल कि हम बस निकलल चाहली। mr: जेव्हा मी पहिल्यांदा आत आलो तेव्हा मला माहित होते की मला आवडतंय आहे. ml: ആദ്യമായി ഞാൻ വന്നപ്പോൾ എനിക്ക് പോകണമെന്ന് അറിയാമായിരുന്നു. pa: ਪਹਿਲੀ ਵਾਰ ਜਦੋਂ ਮੈਂ ਅੰਦਰ ਆਇਆ ਤਾਂ ਮੈਨੂੰ ਪਤਾ ਸੀ ਕਿ ਮੈਂ ਬੱਸ ਆਗੇ ਬਢ਼ਣਾ ਚਾਹੁੰਦਾ ਸੀ। or: ପ୍ରଥମ ଥର ମୁଁ ଭିତରକୁ ଆସିଲି ମୁଁ ଜାଣିଥିଲି ଯେ ମୁଁ ଛାଡିବାକ ur: میں پہلی بار آیا تھا اور مجھے معلوم ہوا کہ میں اس جگہ کو بہت پسند کروں گا۔ te: మొదటిసారి లోపలికి వచ్చాను, నాకు తెలుసు నేను వెళ్ళిపోవాలనుకుంటున్నాని.	en: thank you amanda, i will never be back ! hi धन्यवाद अमांडा, मैं वापस नहीं आऊंगा। mag: धन्यवाद अमांडा, हम बापस नऽ आएम। mr: धन्यवाद अमांडा, मी परत येईन! ml: നന്ദി അമാൻഡ, ഞാൻ മടങ്ങിവരുന്നില്ല! pa: ਖੇਦ ਅਮਾਂਡਾ ਵਾਪਸ ਆਵਾਂਗਾ! or: ଧନ୍ୟବାଦ ଅମାଣ୍ଡା, ମୁଁ ଫେରି ଆସିବି ନାହିଁ। ur: آپ کا شکریہ امینڈا، میں واپس نہیں آؤں گا! te: ధన్యవాదాలు అమండ, నేను వెనక్కి రాను.

Table 4.12: Sample dataset references (on top, shown as input $→$ target) and corresponding outputs generated by selected models for all languages in the dataset.

4.2 Multilingual Text Detoxification

This section is based on the paper *Text detoxification as style transfer in English and Hindi* (Mukherjee et al., 2023a), published in the Proceedings of the 20th International Conference on Natural Language Processing (ICON 2023).

Content warning: *This section contains examples that are toxic, offensive, and/or sexist in nature.*

Text detoxification, as discussed in Section 2.2.2, is a subtask of TST (Dale et al., 2021). Existing methods often rely on rule-based removal of toxic words or phrases (Dementieva et al., 2022), which may result in unnatural sentences and fail to preserve the original meaning. Additionally, limited resources in detoxification datasets make simple sequence-to-sequence training insufficient (Dementieva et al., 2021).

To address these challenges, we propose three methods (Section 4.2.2): (i) knowledge transfer from related tasks, (ii) multi-task learning with sequence-to-sequence modeling and classification tasks, and (iii) a delete-and-reconstruct approach. These techniques enhance text transformation quality and outperform basic sequence-to-sequence methods. Leveraging the detoxification dataset by Dementieva et al. (2021), we collaborated with linguistic experts from Panlingua to curate high-quality data, pairing toxic sentences with their most appropriate non-toxic counterparts. Additionally, we developed a novel multilingual dataset with 500 parallel toxic and non-toxic sentences in Hindi, aligned with English (Section 4.2.1), offering a valuable resource for future multilingual research. Our methodologies demonstrate enhanced detoxification performance in low-resource settings, achieving results comparable to external benchmarks (Section 4.2.6).

4.2.1 Dataset Development

The original dataset (Dementieva et al., 2021) is a collection of user-generated comments which are toxic in nature. The features of the dataset can be summarised as follows:

(i) The utterances are a mix of miscellaneous domains ranging from political to personal to religious, one cannot infer a specific domain/topic from their structure and content.

(ii) The source data is user-generated (comments and reviews) in real-time, hence it does not consist of well-formed sentences. Instead, the data includes of numerous typos, grammatical inaccuracies, and fragmented speech-like structures. The non-toxic/civil has been cleared of typos. But the fragmented speech-like structure remains unchanged, as annotators must preserve the content.

Style Conversion Methodology Based on observation of the original data (Dementieva et al., 2021), we are of the opinion that there are largely two ways utilized to transform toxic utterances into civil utterances:

(i) *Replacement* of abusive words or phrases with synonyms that soften the blow without changing the meaning. For example, *holy shit* can be expressed as *oh goodness.* The element of excitement remains as-is, but the utterance sounds more civil. The problem with this approach is that one size does not fit all. Context and usage must be examined before choosing a suitable replacement.

(ii) Truncation: In certain cases, an abusive word or an abusive phrase is inserted in the utterance. Such phrases can be removed and the remaining sentence can be retained. For example, in the utterance *shit, I might have it.* simply removing *shit* makes the sentence non-toxic but retains the meaning.

Optimising the Original Dataset In order to conduct our experiments and prepare for the Hindi translation, we performed several cleanup exercises:

(i) Of the 2,778 sentences in the data, only 1,108 are unique toxic sentences (Dementieva et al., 2021). The data comprised of 1-5 paraphrases for all the sentences. We manually selected the most suitable response from the available variations, prioritizing those that were free of toxicity while preserving the original intent as closely as possible.

(ii) The data has been cleansed of numerical values. The authors have replaced numbers with *DIGIT*, *number*, or *<number>*. For example: (a) *no fucking way number bricks fits in a car , this is bs .* (b) *DIGIT year olds can be little shits too, doesn t mean you fight them .* There was no way we could have located the original numbers and we did not create fictionalised numbers either. This feature has been carried forward in the translation as well.

(iii) There were certain foreign or non-English texts in the civil part of the database. We replaced such utterances with suitable English equivalents. For example, toxic: *dont be such a hypocrite*, civil: *não seja mentiroso*, corrected civil: *dont be such an unfair person.*

Toxic to non-toxic parallel Hindi dataset A random selection of 500 sentences was manually translated into Hindi by qualified native language experts from Panlingua (as discussed in Section 4.1). We kept the guidelines simple: maximum preservation of content and natural use of language. The following observations were made while translating this data.

(i) Direct and Implicit Toxicity: Most of the utterances contain direct toxicity with the use of abusive words. Implicit toxicity can also be seen in the dataset where the actions and wishes of the user make it toxic. For example, कैंसर पीड़ित हो जाओ और जितनी जल्दी हो सके मर जाओ। *(catch cancer and die as soon as possible).* Here the wish for someone to die after contracting a fatal disease is implicitly toxic whereas explicit abusive words are lacking.

(ii) Choice of pronouns: Hindi pronouns can carry the weight of politeness or impoliteness and this linguistic feature has style transfer more pronounced. The toxic sentences, wherever possible, use the pronoun तू *(tu)*, whereas non-toxic would use आप *(aap).*

(iii) Lack of style transfer: In certain cases, there is hardly any difference between toxic and civil due to the kind of abusive language used. For example, there is not much difference between *awful* and *god awful* in English, and the difference is lost completely when translated into Hindi like भयंकर *(bhayankar) or* भयानक *(bhayanak).*

A few more such examples are documented in Table 4.13.

ID	**Toxic**	**Non-Toxic (Civil)**
1	move to europe if you want to live with stupid . अगर तू बेवकूफ के साथ रहना चाहता है तो यूरोप चला जा।	I don't think it is a good idea to move to Europe. मुझे नहीं लगता कि यूरोप में जाना एक अच्छा विचार है।
2	ya unlike the dems who haven t done shit about climate change. हां, उन डेम्स के विपरीत जिन्होंने जलवायु परिवर्तन के बारे में घंटा कुछ नही किया है।	ya unlike the dems who haven t done anything about climate change . हां, उन डेम्स के विपरीत जिन्होंने जलवायु परिवर्तन के बारे में कुछ नही किया है।
3	thats a great fucking point . एक बहुत ही कमाल का तर्क है।	That is a fantastic point. यह एक शानदार तर्क है।
4	go fuck yourself , and not in a kindly matter at all . अपनी गाँड मार ले, और हल्के से मत मारीयो।	mind your own stuff अपने काम से काम रख।

Table 4.13: A set of examples from English and Hindi detoxification dataset.

4.2.2 Detoxification Models

We first introduce the seq2seq baseline (*seq2seq*), then present three distinct extension approaches: Knowledge Transfer from a Similar Task (*kt*), Multitask Learning, and Delete and Reconstruct (*del_recon*) for the text detoxification task.

Sequence-to-Sequence Baseline *(seq2seq)* The baseline is a sequence-to-sequence learning approach using mBART (Liu et al., 2020) with parallel data. We use the cross-entropy loss for the sequence-to-sequence task, defined as:

$$L_{seq2seq} = -\sum_{t=1}^{T}\sum_{k=1}^{K}\log(P(y_{t,k}|X)), \tag{4.1}$$

where $L_{seq2seq}$ represents the sequence-to-sequence loss, X is the input text, T is the length of the sequence, K is the vocabulary size, and $P(y_{t,k}|X)$ is the predicted probability of the k-th token at time step t given the input X.

Knowledge Transfer from a Similar Task *(kt)* In scenarios with limited resources, leveraging knowledge from a related task can enhance our approach. To achieve this, we employ a two-step process. First, we fine-tune a model to perform the negative-to-positive text sentiment transfer task using the text sentiment transfer yelp dataset provided by Li et al. (2018). Subsequently, we transfer the acquired knowledge in the form of model weights to further fine-tune the model using our toxic-to-non-toxic data, in the same fashion as the *seq2seq* baseline.

Multitask Learning In this approach, we employ a multitask learning setup to transfer toxic attributes in text to non-toxic attributes. This involves learning multiple tasks simultaneously. For an overview of this methodology, see Figure 4.2.

We introduce several classification tasks, each of which works in conjunction with the primary sequence-to-sequence task (*seq2seq*):

(i) **Classification of Input Text *(cls_ip)*:** This task aims to classify the input text (in the encoder) as toxic or non-toxic. The associated loss, combined with the sequence-to-sequence loss, is defined as:

$$L_{cls\_ip} = -\sum_{i=1}^{N}[t_i\log(P_{cls\_ip}(x_i))+ \\ (1-t_i)\log(1-P_{cls\_ip}(x_i))], \tag{4.2}$$

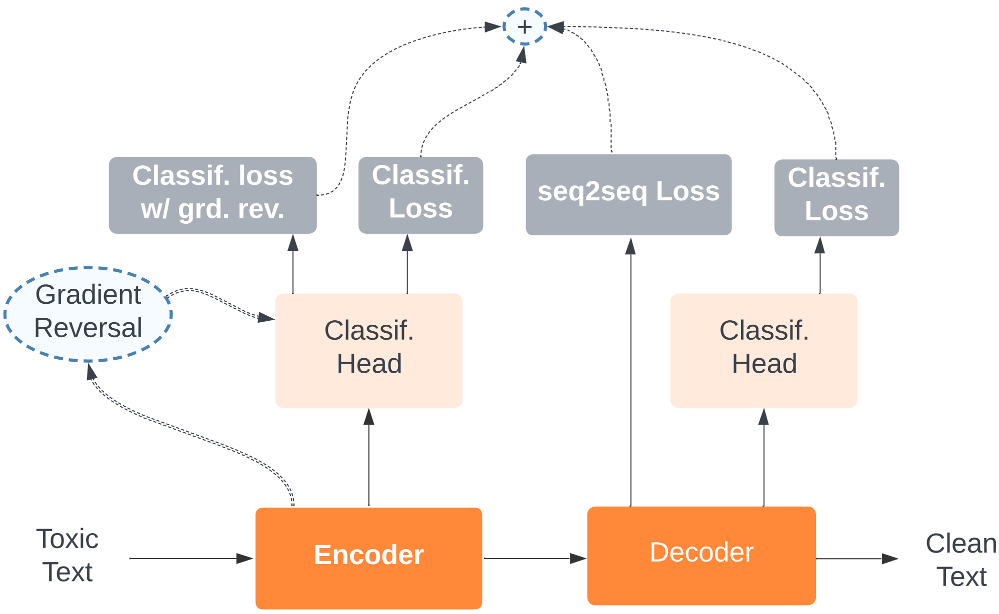


Figure 4.2: Overview of the Multitask Learning Methodology.

where L_{cls_ip} is the toxicity classification loss, N is the number of training samples, x_i is the input text sample, t_i is the corresponding toxicity label (0 for toxic, 1 for non-toxic), and $P_{cls\_ip}(x_i)$ is the predicted probability of toxicity for input x_i.

(ii) **Classification with a Gradient Reversal Layer *(cls_gr_ip)*:** Similar to the classification of input text, in addition, this task includes a gradient reversal layer before the classification head. The gradient reversal layer effectively scales the gradient flowing to the encoders by a factor of $-\lambda$, which should help keep representations of similar-meaning toxic and non-toxic texts similar, focusing on content preservation.

$$Grad_{rev} = -\lambda \cdot \nabla J \tag{4.3}$$

where $-\lambda$ represents a scaling factor that multiplies the gradient. ∇J (`grad_output`) is the gradient flowing through the network.

Then the classification loss is defined as:

$$L_{cls\_gr\_ip} = -\sum_{i=1}^{N}[t_i \log(P_{cls\_gr\_ip}(x_i)) + (1 - t_i)\log(1 - P_{cls\_gr\_ip}(x_i))], \tag{4.4}$$

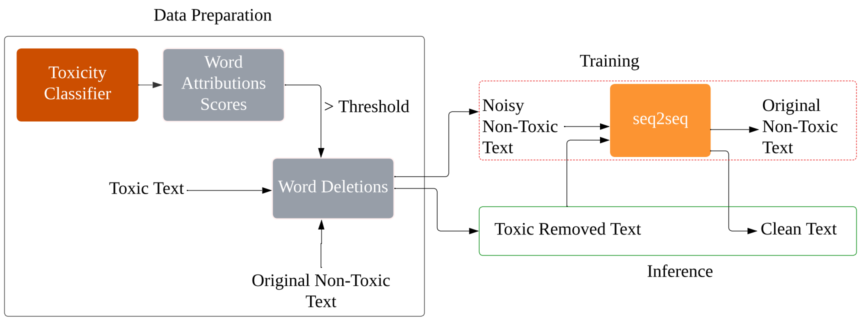


Figure 4.3: Overview of the Delete and Reconstruct Methodology.

where $L_{cls\_gr\_ip}$ is the classification with gradient reversal layer, and $P_{cls\_gr\_ip}(x_i)$ is the predicted probability after applying the gradient reversal layer (Equation 4.3).

(iii) **Classification of Generated Output Text *(cls_op)*:** This task focuses on detecting whether the generated output text (in the decoder) is toxic or non-toxic. The loss is defined similarly to the previous classification tasks:

$$L_{cls\_op} = -\sum_{i=1}^{N}[d_i \log(P_{cls\_op}(y_i)) + (1 - d_i)\log(1 - P_{cls\_op}(y_i))], \quad (4.5)$$

where L_{cls_op} is the loss for detecting generated output text toxic or non-toxic, y_i is the generated output text sample, d_i is the corresponding target toxicity label (0 for toxic, 1 for non-toxic), and $P_{cls\_op}(y_i)$ is the predicted probability of non-toxicity.

Delete and Reconstruct *(del_recon)* This approach is shown in Figure 4.3. We start with a toxicity classifier trained to differentiate between toxic (1) and non-toxic (0) sentences, using the training portion of our dataset (see Section 4.2.4). Leveraging this same classifier, we calculate word attributions for all sentences, encompassing both toxic and non-toxic examples. We then selectively remove words with attributions exceeding a threshold of 0.5.

In the training phase, we fed the sentences after eliminating toxic words or phrases into mBART (Liu et al., 2020) with non-toxic text from the dataset serving as the target output (see Equation 4.6):

Language	Classifiers Accuracy (%)
English	91.7
Hindi	59.8

Table 4.14: English and Hindi accuracy scores for Toxicity classifiers.

$$L_{\text{rec}} = \sum_{i=1}^{N}[y_i \log(P_{\text{rec}}(x_i)) + (1 - y_i)\log(1 - P_{\text{rec}}(x_i))] \quad (4.6)$$

The ultimate goal is to preserve non-toxic content while generating natural and clean text through this process. The loss measures the difference between the reconstructed and original sentences. In Equation 4.6), L_{rec} is the reconstruction loss, $P_{\text{rec}}(x_i)$ is the predicted probability of the reconstructed sentence, y_i is the original non-toxic sentence, and x_i represents the input sentence (where toxic words have been deleted).

4.2.3 Experimental Settings

To ensure consistency in our experiments, we partitioned the modified English dataset into 508 examples for training, 100 for development, and 500 for testing. For the Hindi dataset, we created training and development sets of the same size as the English dataset through machine translation. We utilized the Facebook NLLB-200-3.3B model (Costa-jussà et al., 2022) available from HuggingFace. For evaluation in Hindi, we employed our translated dataset consisting of 500 samples (see Section 4.1.1). We employed the mBART-large-50 model (Tang et al., 2020) from the HuggingFace library (Wolf et al., 2020) for both English and Hindi. To optimize model performance, we conducted hyperparameter tuning, leading to the selection of a learning rate of 1e-5 and a batch size of 3. Throughout the network, dropout was applied with a rate of 0.1. Additionally, we introduced L2 regularization with a strength of 0.01. The training was executed over 5 epochs.

4.2.4 Evaluation Metrics

The evaluation process involves three primary aspects: accuracy of the toxic to civil text transfer (detoxification accuracy), content preservation, and fluency, as outlined in Chapter 2. Detoxification accuracy is assessed using our fine-tuned mBERT classifier, which used the same training set for finetuning as our primary TST task (see Section 4.1.1). Classifier accuracies of toxic and non-toxic text in English and Hindi languages are shown in Table 4.14. The rather low accuracy in Hindi might be a result of the fact that the classifier is finetuned on synthetic training and development sets created by machine English-to-Hindi translation, while it is evaluated using manually translated data.[7] Content preservation is evaluated through BLEU score (Papineni et al., 2002) and embedding similarity (Rahutomo et al., 2012) compared against the input sentences, where embedding similarity is determined using language-agnostic BERT sentence embeddings (LaBSE) (Feng et al., 2022) in conjunction with cosine similarity. Evaluating fluency, particularly for Hindi, poses a challenge due to the limited availability of assessment tools for Indic languages (Krishna et al., 2022). We include a basic fluency assessment using perplexity (PPL) measured with a multilingual GPT model (Shliazhko et al., 2024).

As automated metrics for language generation may not correlate well with human judgments (Novikova et al., 2017), we also run a small-scale human evaluation with language expert annotators on a random sample of 50 sentences from the test set for each language. Outputs are rated on a 5-point Likert scale for detoxification accuracy, content preservation, and fluency (as discussed in Chapter 2 and 3).

4.2.5 External Baselines

Dementieva et al. (2022) provided two of their *RuT5* and *Delete* detoxification baseline methods publicly. We could not use them directly for a result comparison as they are only designed for the Russian language. Therefore, we adapted the *RuT5* model,[8] which is based on the Russian language, using *t5-base* (Raffel et al., 2020) for English and mt5-small (Xue et al., 2021) for Hindi. For the *Delete* method, Dementieva et al. used a dictionary of toxic words and/or phrases. To generate non-toxic sentences, they simply deleted from toxic sentences all toxic words and phrases contained in their dictionary. To adopt this method, we translated their dictionary from Russian to English and then English to Hindi using *Google Translate*[9] and then applied the same technique.

[7] We observed that some of the Hindi machine translation outputs for English toxic inputs are less toxic or not toxic at all. While this may be a generally desired outcome of machine translation, it makes our task of achieving clear toxic and non-toxic classification more challenging.

[8] `https://huggingface.co/ai-forever/ruT5-base`

[9] `https://translate.google.com/`

	English				Hindi			
Models	**ACC**	**BLEU**	**CS**	**PPL**	**ACC**	**BLEU**	**CS**	**PPL**
Our Baseline								
seq2seq	67.4	43.1	76.8	221.4	68.4	39.6	77.2	8.5
Our Methodology - Knowledge Transfer								
kt	71.0	45.6	77.5	237.9	92.0	42.0	78.6	9.3
Our Methodology - Multitask Learning								
seq2seq + cls_ip	64.0	43.7	75.6	202.4	77.2	38.5	76.8	8.3
seq2seq + cls_gr_ip	95.6	0.2	16.3	20.4	75.2	36.2	72.6	8.2
seq2seq + cls_op	75.8	44.2	76.6	348.3	79.8	39.8	78.2	8.2
Our Methodology - Delete and Reconstruct								
del_recon	80.6	44.5	76.9	304.6	94.0	41.2	78.9	8.2
External Baselines (see Section 4.2.5)								
Delete	68.6	41.0	74.4	599.7	92.8	40.4	76.8	11.4
T5	59.2	44.7	77.1	221.6	99.6	1.2	38.7	64.7

Table 4.15: Automatic evaluation results. We measure the detoxification accuracy (ACC) using a toxicity classifier, BLEU score, Content Similarity (CS), and fluency using perplexity (PPL) (see Section 4.1.4). Model names follow the conventions introduced in Section 4.2.2.

4.2.6 Results and Analysis

Automatic Evaluation Automatic evaluation results are presented in Table 4.15.

- Performance of Our Methodologies: The *seq2seq* baseline model showed moderate performance across all metrics, indicating its basic capability in text detoxification. Our knowledge transfer methodology (*kt*) exhibited a substantial improvement in text detoxification accuracy and content preservation over the baseline, thanks to the knowledge transferred from a similar task. In the Multitask Learning setup, *seq2seq + cls_ip* model achieved average results in text detoxification accuracy, content preservation, and fluency. The *seq2seq + cls_gr_ip* approach displayed exceptional text detoxification accuracy but at the cost of fluency and content preservation. The *seq2seq + cls_op* model demonstrated good overall performance with balanced results in text detoxification accuracy, content preservation, and fluency. Our *del_recon* methodology also demonstrates good performance in terms of detoxification accuracy (ACC) and content preservation scores (BLEU and CS), which are the highest overall, but at the cost of poor fluency.

Models	English			Hindi		
	Accuracy	Content	Fluency	Accuracy	Content	Fluency
Our Methodologies						
kt	3.0	4.9	4.9	1.9	4.8	4.8
seq2seq + cls_op	3.2	4.9	4.8	2.0	4.9	4.5
del_recon	3.3	4.7	3.9	2.0	4.9	4.7
External Baseline						
Delete	2.6	4.9	3.0	1.9	4.9	3.9

Table 4.16: Human evaluation of 50 randomly selected outputs on toxic to non-toxic transfer accuracy (Accuracy), Content Preservation (Content), and Fluency (see Section 4.2.6).

- Language-wise Analysis: The results in English and Hindi exhibit mostly the same trends, with methodologies that perform well in one language tending to perform well in the other. However, it is worth noting that while our models maintain mostly good performance in both languages, English text detoxification consistently demonstrates slightly better results. The performance in Hindi is a little inconsistent, possibly due to the use of synthetic data for training (cf. Section 4.2.4).

- Comparison with External Baselines: Comparing our methodologies with external baselines, it is evident that our models outperform both the *Delete* and *T5* baselines in most aspects. Our models generally exhibit more balanced results across multiple metrics.

In summary, our knowledge transfer methodology (*kt*) exhibits notable advancements in text detoxification, with balanced results across different metrics. The multitask learning approaches show promise, with variations that excel in specific aspects. These findings underscore the potential of our methodologies to enhance text detoxification tasks, fostering safer and more respectful online communication.

Human Evaluation A group of in-house experts performed the human evaluation exercise, for which we chose the following four models: (1) *seq2seq+cls_op* as one of our best models from the multitask learning experiment, (2) *kt*, (3) *del_recon*, and (4) *Delete* as the best external baseline. The choice was based on the overall balanced results on automatic metrics (see Section 4.2.6). The results, shown in Table 4.16, mostly align with our automatic evaluation findings, validating the effectiveness of the data and our proposed methodologies.

Sample Output The sample outputs in Table 4.17 provide an overview of the performance of all the systems used in the human evaluation across English and Hindi.

Source → **Target**	**kt** *(Ours)*	**seq2seq + cls_op** *(Ours)*	**del_recon** *(Ours)*	**Delete** *(External baseline)*
and telling nato to fuck off . → and telling nato to leave. और नाटो को चलता करना। → और नाटो को जाने के लिए कहना।	and telling nato to stop. नाटो को चलता करना	and telling nato to stop और नाटो को चलता करना.	and telling nato to off. और नाटो को चलता करना।	and telling nato to off . और नाटो को चलता करना।
ours is pretty fucked up too . → ours isnt good either. हमारे वाले की भी भयंकर लगी पड़ी है। → हमारा वाला भी अच्छा नही है।	ours is pretty down too हमारे वाले की भी भयंकर लगी है।	ours is pretty bad too हमारे वाले की भी भयंकर लगी पड़ी है।	ours is pretty up too. हमारे वाले की भी भयंकर लगी पड़ी है।	ours is up too . हमारे वाले की भी भयंकर लगी पड़ी है।
oh shit you just blew their minds .→ Oh you just blew their minds तेरी माँ की आँख, तूने उनके दिमाग़ की दही कर दी। → ओह, तुमने उनको अचंभित कर दिया।	oh shit you just blew their minds तेरी माँ की आँख, उनके दिमाग़ की दही कर दी।	Oh no you just blew their minds तेरी माँ की आँख, तूने उनके दिमाग़ की दही कर दी।	oh you just blew their minds. तूने उनके दिमाग़ की दही कर दी।	oh you just blew their minds . तेरी माँ की आँख, तूने उनके दिमाग़ की दही कर दी।
he is an autocratic bully . → he is a bully वह एक निरंकुश धमकाने वाला है। → वह एक बदमाश है	he is an autocratic person वह एक निरंकुश धमकाने वाला है	He is not good वह एक निरंकुश धमकाने वाला है	he is an arrogant leader. वह एक निरंकुश धमकाने वाला है।	he is an autocratic . वह एक धमकाने वाला है।
you are a veritable idiot ! → I disagree with you तू एक सत्यनिष्ठ बेवकूफ है! → मैं आप से असहमत हूं	You are unbelievable! तू एक सत्यनिष्ठ व्यक्ति है!	You are not good आप एक सत्यनिष्ठ व्यक्ति है।	you are a veritable person तू एक है?	you are a veritable ! तू एक सत्यनिष्ठ है!

Table 4.17: Selected models' sample outputs in English and Hindi.

The *Delete* baseline's output is inaccurate as compared to *seq2seq + cls_op*. While the *Delete* baseline's deletes the abusive lexical unit's, this makes the sentence syntactically incorrect. If we look closely at the examples presented here, we notice that the adjective or the adverb previously attached to an abusive noun remains syntactically disconnected, as the noun has not been replaced with any other non-abusive lexical unit. Similarly, when we compare the outputs of *kt* and *del_recon*, we find that the former replaces the abusive word with another noun in most cases, whereas *del_recon* simply truncates the abusive word, creating a syntactic inconsistency in the sentence. The results for Hindi are not on par with English for any model.

4.3 Conclusion

In this chapter, we addressed the challenge of multilingual TST, focusing on the subtasks of text sentiment transfer and text detoxification in Indian languages. For text sentiment transfer (in Section 4.1), we contributed valuable resources in nine Indian languages, explored various benchmark models, and analyzed experimental results across these languages. We also extended the task of text detoxification (in Section 4.2) to English and Hindi, providing annotated datasets and benchmark models validated through comprehensive analysis.

Our datasets are both style-parallel and parallel across languages, ensuring consistency and comparability for the TST task, which supports cross-linguistic style transfer research and facilitates broader multilingual applications.

Limitations and Future Work The effectiveness of our approach may vary across different languages and dialects. In future work, we aim to explore a broader range of style attributes and incorporate additional languages and dialects. Building on our existing methodologies and framework, we plan to adapt our approach to any style attribute, given the availability of parallel data, to further advance multilingual TST research. Additionally, in Chapter 7, we explore leveraging LLMs for both the TST tasks: sentiment transfer and detoxification.

5

Text Style Transfer Evaluation

This chapter is based on the paper *Evaluating Text Style Transfer Evaluation: Are There Any Reliable Metrics?* (Mukherjee et al., 2025), under review at NAACL-SRW (2025)

Evaluating the quality of machine-generated text is inherently challenging, and this complexity only increases in the context of Text Style Transfer (TST) (Hu et al., 2022). TST evaluation is a task requiring consideration of multiple dimensions, including style transfer accuracy, content preservation, and overall fluency (as discussed in Section 2.2). Human evaluation is widely regarded as the gold standard for assessing the quality of TST outputs due to its ability to provide nuanced judgments (Briakou et al., 2021b). However, it is also expensive, time-consuming, and lacks reproducibility (Belz et al., 2023). Consequently, automated metrics have become a proxy for human judgment, but there is a notable lack of standardization and consensus on which metrics best capture style transfer accuracy, content preservation, and overall naturalness (Mir et al., 2019; Briakou et al., 2021a; Ostheimer et al., 2023). In addition, large language models (LLMs) could serve as alternatives to traditional human evaluation and automated metrics for TST evaluation (Ostheimer et al., 2024). However, the rapid evolution of LLMs, particularly for closed-source models, raises concerns about reproducibility (Gao et al., 2024; Chen et al., 2023).

In this chapter, we examine various existing and novel evaluation metrics for two popular TST subtasks: *sentiment transfer* (as discussed in Section 4.1) and *detoxification* (as discussed in Section 4.2). Our experiments span a multilingual setting, covering English, Hindi, and Bengali (based on datasets produced in Chapter 4), to investigate the utility of these metrics across diverse linguistic contexts. We conduct a meta-evaluation of all the metrics by measuring their correlation with human judgments. To further explore the potential of automated metrics, we also combine them in ensembles, experimentally creating hybrid scores. Additionally, we investigate the applicability of LLMs as an alternative evaluation tool. Our results show that newly applied text metrics, hybrid approaches, and LLMs can improve correlation with human evaluations over existing TST metrics, offering a more robust and comprehensive assessment of TST outputs.

This chapter is organized as follows. Section 5.1 presents all the evaluation metrics explored in our study, including existing TST metrics, other NLP text metrics previously used in other contexts, and newly proposed metrics, along with LLMs used as evaluators. Section 5.2 details the experimental setup for the metric meta-evaluation, followed by a discussion of the results in Section 5.3. Finally, Section 5.4 concludes the chapter by highlighting its limitations and suggesting directions for future research.

5.1 Metrics Compared

Traditional TST metrics can be grouped under style transfer accuracy, content preservation, and fluency (as discussed in Section 2.2). Following this, we conduct evaluations in two scenarios: (1) *reference-based*, where metrics are computed against a reference text (when available), and (2) *reference-free*, where metrics directly compare the generated text against the source text (measuring similarity or distance from the original), without requiring a reference.

Previously Used TST Metrics For style transfer accuracy, we include *Sentence Accuracy* based on a fine-tuned *XLM-RoBERTa-base* (Conneau et al., 2020) classifier (as discussed in Section 7.1.4) (Prabhumoye et al., 2018), and *WMD* (Kusner et al., 2015; Wei et al., 2023; Mir et al., 2019). For content preservation: *BLEU* (Papineni et al., 2002; Tikhonov et al., 2019), *Cosine Similarity* (Rahutomo et al., 2012; Reimers and Gurevych, 2019), *Masked BLEU and Masked Cosine Similarity* (as discussed in Section 3.1.6), *ROUGE-2* and *ROUGE-L* (Lin and Hovy, 2003; Lin, 2004; Lin and Och, 2004; Yamshchikov et al., 2021). For fluency, we use *Perplexity* of *GPT-2* (Radford et al., 2019; Briakou et al., 2021b) and *MGPT* (Shliazhko et al., 2024).

Newly Applied Text Metrics We expand the TST evaluation by incorporating additional metrics from related NLP tasks, categorizing them into trainable and non-trainable metrics as well as word-overlap-based and embedding-based measures.

For style transfer accuracy, we utilize non-trainable statistical measures such as *Earth Mover's Distance (EMD)* (Rubner et al., 2000), *KL Divergence* (Kullback, 1997), *Cosine Similarity* (Rahutomo et al., 2012), and *Jensen-Shannon Divergence* (Lin, 1991), which quantify the distributional shift between source and generated text. Additionally, we incorporate a trainable *Classifier Confidence* score, derived from the Sentence Accuracy classifier described earlier.

For content preservation, we include both word-overlap-based and embedding-based metrics. The word-overlap-based metrics include *PINC* (Chen and Dolan, 2011), which measures the proportion of n-grams in the generated text that do not appear in the source text (higher values indicate greater lexical divergence), *METEOR* (Banerjee and Lavie, 2005), which accounts for synonymy and stemming, and *Translation Edit Rate (TER)* (Snover et al., 2006), which evaluates the number of edits required to transform the generated text into the reference. Embedding-based measures include *Word Mover's Distance (WMD)* (Kusner et al., 2015; Wei et al., 2023), *BERTScore* (Zhang et al., 2020a), *S^3BERT* (Opitz and Frank, 2022), and *BLEURT* (Sellam et al., 2020), all of which assess content similarity based on contextualized vector representations. Additionally, we introduce *Tree Edit Distance (TED)* (Zhang and Shasha, 1989), which measures structural similarity by computing the minimum number of tree edit operations (insertion, deletion, substitution) required to transform one syntactic tree into another. This metric is particularly useful in evaluating syntactic shifts in generated text.

For fluency evaluation, we employ language model perplexity, using *Finetuned GPT-2* and *Finetuned MGPT* trained on target styles (see fine-tuning details in Appendix C of our paper (Mukherjee et al., 2025)). Lower perplexity scores indicate higher fluency, as they reflect the model's confidence in the generated text.

Novel Metrics We analyze the structural similarity between the source/reference and the system-generated outputs by parsing them into two Abstract Meaning Representation (AMR) (Banarescu et al., 2013) and Syntactic Dependency Trees (Straka and Straková, 2017). AMR provides a semantic abstraction of sentences by capturing their core meaning as directed graphs, while Syntactic Dependencies represent the grammatical relationships between words in tree form. To measure structural similarity, we first convert syntactic dependency trees into AMR-style structure trees, ensuring both syntactic and semantic representations are in a comparable graph format. We then compute Smatch similarity (Cai and Knight, 2013) for both AMR graphs and the synthetic trees translated to AMR-style trees. Smatch (a graph-matching metric) computes the F-score between AMR graphs by aligning their nodes and edges optimally, regardless of differences in variable naming or graph representation. A higher Smatch score, i.e. a higher AMR graph and syntactic tree similarity, indicates greater preservation of meaning and syntactic structure in the transformed text.

LLM Prompting Following Ostheimer et al. (2024), we use LLM as TST evaluator and extend their method to newer models, TST tasks, and additional languages. We used GPT-4 (Achiam et al., 2023) and Llama-3.1 8B (Dubey et al., 2024), to assess the TST tasks. We employed a Likert-scale-based approach (as discussed in Section 4.1.4) to evaluate style transfer accuracy, content preservation, and fluency. To facilitate direct comparison with *Sentence Accuracy*, we also conducted a binary evaluation for style transfer accuracy (*GPT4-bin-acc, Llama-bin-acc*). Detailed prompt instructions are provided in Appendix D of our paper (Mukherjee et al., 2025).

Hybrid We propose two ensemble-based oracle metrics – *Hybrid-Simulation* and *Hybrid-Learned* – to show the potential of integrating multiple evaluation metrics.[1] In *Hybrid-Simulation*, we first select the top three metrics (based on correlation with human judgments) for each task and language from Tables 5.1 and 5.2. We then conduct a simulation to determine the selected metrics' relative weights by tuning them on human-labeled target data and compute their geometric average to form the final ensemble score. In *Hybrid-Learned*, we train a *RandomForestRegressor* (Liaw, 2002) using all available metrics as features and human ratings as the target labels. The model assigns importance scores to each metric, and we select the top three metrics with the highest normalized importance scores. Their geometric average, weighted by these importance scores, is used to generate the ensemble result. For details on the selected metrics and their respective weights, see Tables 4 and 5 in Appendix A of our paper (Mukherjee et al., 2025).

Overall Score Following Loakman et al. (2023) and Yang and Jin (2023), we adopt the geometric mean of style transfer accuracy, content preservation, and fluency as a single aggregated score for comparison. We again aim to show the potential of this approach by producing oracle metrics. Based on the Pearson correlation results from our experiments (Tables 5.1, 5.2 and, 5.3), we select the best-performing metrics for these three dimensions from existing, LLM-based, and newly proposed methods

[1] These metrics are considered "oracle", since the approach learns optimal weights based on the target data

(excluding hybrid approaches), creating *Ours1* score. We then extend *Ours1* by incorporating the top-performing metrics from our proposed approaches, including hybrids, to construct *Ours2*. Table 7 in Appendix A from our paper (Mukherjee et al., 2025) details the metrics selected for each language and task.

5.2 Experiment Setup

Evaluation Data: Tasks, Languages, and Model Outputs We evaluate our methods on the outputs of TST models and human annotations, which were produced in our LLM experiments described in Chapter 7. The style transfer data used have been described in Chapter 4. This comprises two TST tasks – sentiment transfer (positive to negative statements and vice versa), where data is available for English, Hindi, and Bengali, and detoxification (toxic to clean text), with English and Hindi data. Model outputs for all tasks were produced by GPT-3.5 (OpenAI, 2023), Llama-2-7B-Chat (Touvron et al., 2023b) and Mistral-7B-Instruct (Jiang et al., 2023), as well as previous finetuned BART models which were discussed in Section 4.1.4.

Meta-Evaluation Approach We follow common practice for meta-evaluation (Kilickaya et al., 2017; Zhang et al., 2020a; Liu et al., 2023b) and compute all metrics' correlation with human judgment using Pearson (PC), Spearman (SC), and Kendall's Tau (KC) Correlations (Schober et al., 2018; Puka, 2011).

5.3 Results Analysis

Since we found that reference-based metrics generally underperform their reference-free variants, we focus on the reference-free setting in the analysis. We include reference-based results in Appendix B of our paper (Mukherjee et al., 2025).

5.3.1 Style Transfer Accuracy

The results for style transfer accuracy in the reference-free setting are shown in Table 5.1.

Previously Used *Sentence Accuracy* generally achieves moderate to good correlation with human judgments, suggesting that direct style classification accuracy can be a reliable indicator of style transfer quality. Meanwhile, *EMD* demonstrates a moderate degree of alignment, implying that capturing distributional shifts of stylistic cues correlates moderately with human perceptions.

Metrics	Sentiment Transfer (reference free) English PC	SC	KC	Hindi PC	SC	KC	Bengali PC	SC	KC	Detoxification (reference free) English PC	SC	KC	Hindi PC	SC	KC
Previously used & LLMs															
Sentence Accuracy	**0.51**	0.49	0.48	**0.61**	0.61	0.59	**0.57**	0.57	0.54	**0.36**	0.36	0.35	0.38	0.37	0.36
EMD	0.27	0.24	0.20	0.36	0.43	0.34	0.50	0.52	0.40	0.29	0.21	0.17	**0.47**	0.53	0.43
GPT4	**0.92**	0.81	0.79	**0.87**	0.84	0.79	**0.82**	0.83	0.77	**0.74**	0.72	0.65	**0.74**	0.74	0.68
GPT4-bin-acc	0.89	0.78	0.77	0.84	0.83	0.80	0.77	0.78	0.74	0.61	0.61	0.59	0.60	0.61	0.59
Llama	0.16	0.17	0.15	-0.11	-0.10	-0.09	-0.17	-0.15	-0.13	0.20	0.18	0.17	0.20	0.16	0.15
Llama-bin-acc	0.49	0.44	0.43	0.50	0.51	0.49	0.31	0.31	0.30	0.24	0.24	0.23	0.27	0.27	0.27
Newly applied & Novel															
Classifier Confidence	0.51	0.43	0.35	0.66	0.57	0.46	0.59	0.52	0.40	0.39	0.32	0.25	0.41	0.38	0.30
KL Divergence	0.59	0.31	0.24	0.66	0.66	0.54	**0.62**	0.62	0.50	**0.46**	0.46	0.36	0.51	0.60	0.49
Cosine Similarity	-0.55	-0.44	-0.36	-0.66	-0.67	-0.54	-0.53	-0.59	-0.46	-0.43	-0.40	-0.32	-0.48	-0.58	-0.47
Jensen-Shannon Divergence	**0.67**	0.40	0.32	**0.69**	0.67	0.55	**0.62**	0.64	0.51	0.41	0.50	0.39	**0.53**	0.60	0.50
Hybrid-Simulation	**0.69**	0.40	0.32	**0.71**	0.67	0.54	**0.62**	0.64	0.51	**0.44**	0.47	0.37	0.53	0.61	0.49
Hybrid-Learned	0.67	0.37	0.30	0.70	0.63	0.50	0.61	0.62	0.49	0.43	0.47	0.37	**0.55**	0.62	0.50

Table 5.1: Style transfer quality (reference-free). Pearson (PC), Spearman (SC) and Kendall's Tau (KC) correlations.

Newly Applied: *Classifier Confidence*, *Cosine Similarity*, *KL Divergence*, and *Jensen-Shannon Divergence* generally exhibit stronger alignment with human judgments compared to existing metrics, highlighting the effectiveness of distributional measures for style intensity comparisons.

LLMs: *GPT-4* exhibits consistently high correlations, whereas *Llama* performs notably worse, although a binarized version (*Llama-bin-acc*) shows some moderate improvements.

Hybrid: *Hybrid-Simulation* demonstrates strong alignment with human ratings by combining multiple signals into a single score, while *Hybrid-Learned* performs comparably, though it may fall marginally below its simulation-based counterpart in certain cases.

Direct classification metrics reliably capture style accuracy, while distribution-based and LLM-based evaluations enhance overall alignment with human judgments, especially when integrated in hybrid frameworks. In English tasks, approaches like GPT-4 and hybrid methods achieve particularly high correlations, whereas in Hindi and Bengali, top metrics (e.g., KL, JS Divergence, and hybrid approaches) remain strong but show more pronounced performance gaps, potentially due to greater linguistic complexity.

5.3.2 Content Preservation

We present the meta-evaluation of content preservation metrics in a reference-free setting in Table 5.2.

	Sentiment Transfer (reference free)									Detoxification (reference free)					
	English			Hindi			Bengali			English			Hindi		
Metrics	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**
Previously used & LLMs															
BLEU	0.24	0.22	0.18	0.24	0.19	0.15	0.32	0.31	0.25	0.14	0.13	0.11	0.45	0.37	0.31
Cosine Similarity	**0.54**	0.27	0.22	**0.33**	0.24	0.20	**0.43**	0.40	0.32	**0.28**	0.19	0.15	**0.59**	0.45	0.38
Masked BLEU	0.21	0.21	0.17	0.15	0.12	0.10	0.23	0.24	0.19	0.15	0.15	0.12	0.45	0.39	0.32
Masked Cosine Similarity	0.36	0.17	0.14	0.19	0.13	0.11	0.28	0.29	0.23	0.23	0.15	0.12	0.56	0.45	0.37
METEOR	0.38	0.25	0.21	0.20	0.18	0.14	0.33	0.27	0.22	0.16	0.10	0.08	0.54	0.34	0.28
ROUGE-2	0.24	0.19	0.16	0.19	0.20	0.16	0.28	0.30	0.24	0.17	0.11	0.09	0.41	0.37	0.31
ROUGE-L	0.39	0.25	0.21	0.26	0.23	0.19	0.28	0.32	0.25	0.22	0.12	0.10	0.46	0.39	0.33
GPT4	**0.42**	0.36	0.35	**0.39**	0.41	0.39	**0.51**	0.54	0.48	**0.46**	0.31	0.30	**0.46**	0.42	0.40
Llama	0.24	0.26	0.24	0.32	0.28	0.26	0.32	0.38	0.35	0.25	0.11	0.10	0.28	0.16	0.16
Newly applied & Novel															
PINC	-0.18	-0.17	-0.15	-0.16	-0.12	-0.10	-0.27	-0.28	-0.23	-0.12	-0.12	-0.10	-0.41	-0.36	-0.30
WMD	0.35	0.28	0.23	0.27	0.24	0.20	0.34	0.35	0.28	0.15	0.14	0.11	0.41	0.38	0.32
BERTScore	**0.50**	0.31	0.26	**0.45**	0.33	0.27	0.49	0.44	0.36	0.21	0.19	0.15	**0.62**	0.38	0.31
Smatch (Dependency Trees)	0.25	0.24	0.20	0.18	0.20	0.17	0.26	0.30	0.25	0.16	0.15	0.12	0.34	0.31	0.26
Smatch (AMR)	0.38	0.25	0.20	0.22	0.20	0.17	0.32	0.32	0.26	0.19	0.13	0.11	0.37	0.34	0.28
S3BERT	0.46	0.23	0.19	0.30	0.18	0.14	0.30	0.30	0.24	0.22	0.20	0.16	0.49	0.38	0.31
BLEURT	0.47	0.30	0.25	0.41	0.35	0.29	**0.56**	0.53	0.42	0.18	0.17	0.14	**0.62**	0.43	0.35
TER	0.42	0.26	0.22	**0.45**	0.28	0.24	0.34	0.33	0.27	0.21	0.17	0.14	0.58	0.37	0.31
TED	0.43	0.24	0.22	0.42	0.29	0.25	0.20	0.28	0.24	**0.48**	0.21	0.18	0.48	0.36	0.30
Hybrid-Simulation	**0.57**	0.32	0.26	**0.48**	0.33	0.27	**0.57**	0.53	0.43	**0.28**	0.19	0.15	**0.68**	0.43	0.35
Hybrid-Learned	0.56	0.32	0.26	0.47	0.35	0.29	0.56	0.53	0.43	0.19	0.15	0.12	0.64	0.38	0.31

Table 5.2: Content preservation (reference-free). Pearson (PC), Spearman (SC) and Kendall's Tau (KC) correlations.

	Sentiment Transfer									Detoxification					
	English			Hindi			Bengali			English			Hindi		
Metrics	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**
Previously used & LLMs															
Perplexity (GPT-2)	**0.13**	0.13	0.11	-0.11	-0.10	-0.08	-0.11	-0.07	-0.05	**0.06**	0.00	0.00	**0.17**	-0.13	-0.11
Perplexity (MGPT)	0.08	0.19	0.15	0.00	0.07	0.05	0.16	0.19	0.15	0.05	0.00	0.00	0.11	0.03	0.03
GPT4	**0.43**	0.40	0.37	**0.39**	0.39	0.35	**0.37**	0.40	0.36	**0.16**	0.13	0.12	**0.17**	0.17	0.16
Llama	0.17	0.18	0.17	0.15	0.17	0.15	0.08	0.06	0.06	**0.16**	0.13	0.12	-0.01	-0.02	-0.01
Newly applied															
Perplexity (Finetuned GPT-2)	**0.14**	0.16	0.13	0.08	0.14	0.11	0.02	0.05	0.04	**0.14**	0.00	0.00	0.11	-0.06	-0.05
Perplexity (Finetuned MGPT)	0.04	0.08	0.07	**0.17**	0.15	0.12	**0.23**	0.21	0.16	0.00	0.03	0.03	**0.23**	0.04	0.03

Table 5.3: Fluency (reference-free). Pearson (PC), Spearman (SC) and Kendall's Tau (KC) correlations.

	Sentiment Transfer									Detoxification					
	English			Hindi			Bengali			English			Hindi		
Metrics	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**	**PC**	**SC**	**KC**
Existing	0.32	0.02	0.02	0.11	-0.02	-0.01	0.25	0.18	0.13	-0.04	-0.18	-0.14	0.07	-0.19	-0.14
Human	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GPT4	**0.73**	0.62	0.54	**0.78**	0.75	0.61	**0.78**	0.77	0.63	**0.65**	0.62	0.51	0.62	0.59	0.46
Llama	0.08	0.16	0.13	0.02	0.01	0.01	0.01	0.01	0.00	0.18	0.14	0.11	0.27	0.23	0.19
Ours1	0.57	0.33	0.26	0.59	0.54	0.43	0.54	0.57	0.42	0.38	0.44	0.34	0.47	0.43	0.32
Ours2	0.68	0.40	0.31	0.72	0.68	0.53	0.59	0.59	0.42	0.41	0.38	0.29	**0.63**	0.57	0.43

Table 5.4: Overall results (reference-free). Pearson (PC), Spearman (SC) and Kendall's Tau (KC) correlations.

Previously Used: *BLEU* generally shows low alignment with human judgments, while *Cosine Similarity* exhibits better performance in several tasks. *Masked BLEU* and *Masked Cosine Similarity* offer slight improvements over their unmasked counterparts, yet they still lag behind more recent methods. *ROUGE-2* and *ROUGE-L* provide moderate correlations but do not consistently outperform newer metrics.

Newly Applied: *BLEURT* remains consistently reliable, while *BERTScore* also proves robust across various styles and languages. *TER* and *TED* offer competitive results, particularly for certain language-specific tasks. In contrast, PINC shows weak correlations, indicating its limited effectiveness in capturing content preservation.

Novel: *Smatch (Dependency Trees)* and *Smatch (AMR)* outperform or at least match the performance of traditional metrics, though it generally falls behind the newly introduced text-based methods and LLM-driven approaches on average.

LLMs: *GPT-4* achieves higher correlations than traditional metrics across different styles and languages, demonstrating its strong ability to capture human-like judgments of text transformations. In contrast, *Llama* tends to underperform, indicating considerable variability in how well different LLMs reflect stylistic and content-based shifts.

Hybrid: *Hybrid-Simulation* achieves robust alignment with human ratings by unifying multiple signals into a single score, whereas *Hybrid-Learned* shows comparable performance, albeit slightly trailing the simulation-based approach in some scenarios.

5.3.3 Fluency

Table 5.3 presents fluency evaluation results. *GPT-2 Perplexity* displays limited correlations with human judgments, while *Finetuned GPT-2 Perplexity* yields only marginal gains. *MGPT Perplexity* and *Finetuned MGPT Perplexity* provide moderate improvements under fine-tuning, underscoring the importance of multilingual modeling and style-specific training for better alignment with human fluency assessments. *GPT-4* demonstrates relatively strong correlations with human assessments of fluency for sentiment-related tasks, suggesting it captures fluidity and coherence more effectively when the stylistic shift involves changing sentiment. However, for detoxification tasks, its alignment with human judgments diminishes, indicating that removing toxicity poses different challenges for GPT-4. In contrast, *Llama* exhibits generally weaker correlations and struggles in various settings, implying that its evaluations of fluency do not consistently match human perceptions.

Moreover, differences across languages persist; English often yields slightly better correlations, while Hindi and Bengali results vary more substantially.

5.3.4 Overall Score

Table 5.4 shows the single aggregate score.

Previously Used: Aggregating traditional metrics often yields near-zero or negative correlations across various languages and tasks, indicating that simply merging these measures fails to capture the overall quality.

LLM: *GPT-4* consistently aligns well with human assessments of overall quality in both Sentiment Transfer and Detoxification. *Llama*, however, shows weaker correlations, indicating that not all LLMs possess the same evaluative capabilities.

Ours: Our approaches (*Ours1* and *Ours2*) provide noticeable improvements over existing methods. Although these do not surpass GPT-4, they clearly outperform many traditional and alternative measures.

5.4 Conclusion

We presented a comprehensive evaluation of existing and newly proposed metrics for two TST subtasks—*Sentiment Transfer* and *Text Detoxification*—in English, Hindi, and Bengali. Our findings demonstrate that traditional word-overlap-based metrics like BLEU and ROUGE often show limited correlation with human judgments, whereas our proposed experimental metrics and prompted LLM-based evaluations provide significantly stronger alignment. Moreover, our oracle hybrid ensemble and combined approaches show an even greater potential of merging multiple metrics.

Limitations and Future Work Our study is constrained to two specific tasks and three languages, raising open questions about the generalizability of these metrics to other styles, languages, and domains. Additionally, while oracle ensemble metrics provide valuable insights, further research is needed to develop fully generalizable evaluation methods that do not rely on target-specific tuning. Future work could explore adapting these approaches across a broader range of stylistic transformations and diverse languages (as discussed in Chapter 4) to assess their robustness and applicability in real-world scenarios.

6 Application of Text Style Transfer: Polite Chatbot

This chapter is based on the paper *Polite Chatbot: A Text Style Transfer Application* (Mukherjee et al., 2024c), published in the Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop (EACL-SRW 2023).

The ability to modify text style without changing the underlying intent has led to numerous applications across various domains (as discussed in Section 2.2.6), notably in stylized dialogue generation has attracted significant attention in recent years (Gao et al., 2019; Zheng et al., 2021; Zeng and Nie, 2021), with prior works focusing on personalized (Li et al., 2016; Luan et al., 2017; Su et al., 2019), polite (Niu and Bansal, 2018), or emotional (Zhou et al., 2018) dialogues.

In this chapter, we focus on one such application: the development of a polite chatbot. Politeness is a text style attribute closely tied to social interactions, enabling smooth communication in conversations such as emails or memos (Coppock, 2005). Importantly, politeness can be decoupled from the content (Kang and Hovy, 2019), making it an ideal candidate for style transfer tasks. The ability to generate polite responses is crucial for dialogue systems aiming to provide engaging and socially appropriate interactions. Traditionally, polite chatbot responses have been accomplished through manual dialogue design, where predefined rules or templates generate responses based on certain keywords or scenarios (André et al., 2004; Gupta et al., 2007; de Jong et al., 2008). This approach has limitations such as requiring extensive human effort, being domain-specific, and lacking flexibility or diversity (Firdaus et al., 2022). However, one of the major challenges in training a polite chatbot is the

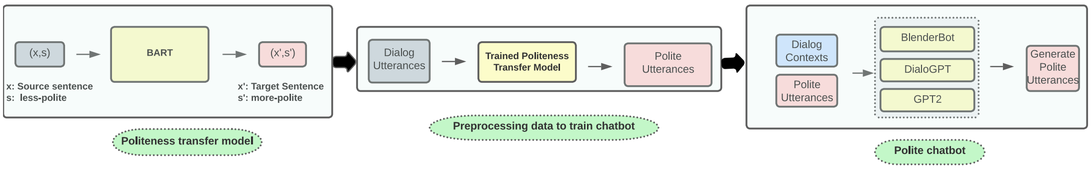


Figure 6.1: Our method: We (1) train the politeness transfer model; (2) generate synthetic training data by applying the transfer model to neutral utterances; (3) train the dialogue models using the synthetic data.

lack of parallel data, as discussed in Chapter 3. The stylistic features we aim to capture are embedded in unpaired texts, which cannot be directly utilized by supervised models (Gao et al., 2019). This limitation necessitates alternative approaches, such as unsupervised or weakly supervised methods, to effectively learn politeness transformations. This typically leads stylized chatbot models to employ complex, multi-step setups, potentially involving reinforcement learning (Niu and Bansal, 2018; Sun et al., 2022; Firdaus et al., 2022). Zheng et al. (2021) argue that it is challenging to produce coherent and style-specific responses by relying solely on reinforcement learning rewards. Our approach first employs a politeness transfer model (Madaan et al., 2020) to generate synthetic dialogue pairs consisting of contexts and polite utterances using TST methodologies. These synthetic pairs are then utilized to train a dialogue model, resulting in an end-to-end system. This straightforward training procedure simplifies the process and improves the coherence and politeness of the responses generated.

This chapter is structured as follows. In Section 6.1, we present our methodological approach. Section 6.2 details the experimental setup and procedures. The evaluation process and corresponding results are discussed in Section 6.3. Finally, Section 6.4 concludes the chapter by addressing the study's limitations and outlining directions for future work.

6.1 Methodology

Our method consists of three steps (Figure 6.1). First, we train a politeness transfer model. Our goal here is to train a model that takes as input a neutral sentence x and outputs a sentence $\hat{x}$ that retains the content while increasing politeness. Second, we apply this politeness transfer model to generate synthetic polite chat data. Finally, we use the corpus $\hat{\mathcal{D}}$ to train a dialogue model.

Politeness Transfer Model Although we do not have parallel corpora available for politeness transfer, our transfer model is trained in a supervised fashion on synthetic input-output pairs. These are obtained following Madaan et al. (2020): polite phrases (politeness markers) are identified using TF-IDF over polite and non-polite texts.[1] The markers are removed from polite texts on the input, and a sequence-to-sequence model is trained to increase sentence politeness by reconstructing the politeness markers on the output. Unlike Madaan et al. (2020), we do not use separate tagging and generation steps here and join the task into a single step. Specifically, we finetune a pre-trained language model for this task using standard cross-entropy loss (see Section 6.2.2).

In our approach, we address the challenge of politeness transfer without access to parallel corpora by generating synthetic input-output pairs for supervised training. Following the methodology of Madaan et al. (2020), we identify politeness markers—phrases that frequently occur in polite texts but are less common in non-polite ones—using TF-IDF analysis. Specifically, we calculate the TF-IDF scores for phrases across polite and non-polite corpora; phrases with significantly higher TF-IDF scores in polite texts are considered potential politeness markers. For example, phrases like “could you please” or “would you mind” may have higher TF-IDF scores in polite texts, indicating their role as politeness markers. In contrast to Madaan et al.’s two-step *tag and generate* framework—which first tags positions in non-polite sentences for politeness marker insertion and then generates the polite sentence—we streamline the process by employing a single-step approach. We fine-tune a single pre-trained language model that directly learns to transform impolite sentences into polite ones through end-to-end training using standard cross-entropy loss (see Section 6.2.2). This unified approach allows the model to jointly learn both where politeness markers are needed and what specific polite phrases to generate.

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

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

Dialogue Model We use a standard dialogue response generation model that produces a dialogue utterance u_i based on context $\mathbf{C} = \{u_1, ..., u_{i-1}\}$, trained using cross-entropy loss. We experiment with multiple pre-trained language models here (see Section 6.2.2). To achieve politeness in responses, we use the synthetic polite dialogue corpus $\hat{\mathcal{D}}$ obtained using our politeness transfer model.

6.2 Experiment Details

This section discusses the experimental details, covering the dataset used (Section 6.2.1), the parameter settings (Section 6.2.2), the baselines (Section 6.2.3).

6.2.1 Dataset

Politeness Transfer We use the dataset of Madaan et al. (2020), i.e. the Enron e-mail dataset (Shetty and Adibi, 2004), preprocessed and filtered sentences into ten buckets (P_0-P_9) based on the score of a politeness classifier by Niu and Bansal (2018). We use Madaan et al. (2020)'s TF-IDF-based approach to remove politeness markers (see Section 6.1) from the sentences in the most polite P_9 bucket to prepare synthetic parallel data for training our politeness transfer models.

Dialogue To train our response generation models, we use DailyDialog (Li et al., 2017), an open-domain dataset of 13,118 human-human dialogues. DailyDialog was collected to represent natural day-by-day conversations between human participants. It is constructed mainly from English learner websites and represents rather formal written conversations. Each conversation is focused on certain topics, but the domain is not restricted in general. The average length is 7.9 turns per dialogue. The dataset is split into a training set with 11,118 dialogues and validation and test sets with 1,000 dialogues each.

6.2.2 Parameter Settings

We use BART (Lewis et al., 2020) for politeness transfer. For dialogue modeling, we utilize multiple pre-trained models as follows:

1. **GPT-2** (Radford et al., 2019): A Transformer decoder trained for general language modeling, including dialogues.
2. **DialoGPT** (Zhang et al., 2020b): Shares GPT-2's architecture but was pre-trained specifically on dialogue data.

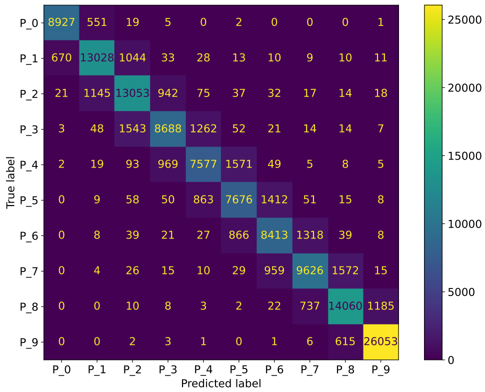


Figure 6.2: Confusion Matrix of multi-class politeness classification. We can see that the model confuses mainly neighboring buckets and the vast majority is classified correctly.

3. **BlenderBot** (Roller et al., 2021): An encoder-decoder Transformer specifically trained to develop dialogue skills such as empathy and engagement.

For our experiment, we use AdamW optimizer (Loshchilov, 2017) with a learning rate of 5e-4 in all cases. The politeness transfer model is trained for 5 epochs using batch size 8. All dialogue models are finetuned for 4 epochs using batch size 3.

6.2.3 Baselines

Politeness Transfer We compare our system against Madaan et al. (2020). They used 4-layered transformers (Vaswani et al., 2017) to train both tagger and generator modules (as discussed in Section 6.1). Each transformer has 4 attention heads with a 512 dimensional embedding layer and hidden state size.

Training Dialogue Models We evaluate all our dialogue models against three baselines: (1) vanilla version of the model, (2) model fine-tuned on unchanged DailyDialog data, (3) model finetuned on synthetic polite DailyDialog data generated in the same fashion as in our full model underlying pretraining, but using Madaan et al. (2020)'s politeness transfer instead of ours.

6.3 Evaluation

The section discusses the evaluation process, including the metrics employed (Section 6.3.1) and the results obtained (Section 6.3.2).

Models	PS	BLEU	CS
Madaan et al. (2020)	7.01	60.16	87.86
Ours	**8.68**	**71.65**	**93.25**

Table 6.1: Evaluation results of politeness transfer on the test set of Madaan et al. (2020)'s data. We measure the Polite Score (PS), BLEU Score, and Content Similarity (CS). Model outputs are predicted based on synthetic sentences where politeness markers have been removed. BLEU and CS compare against original human-written polite sentences.

6.3.1 Metrics

Following prior work (Madaan et al., 2020; Niu and Bansal, 2018), we use automatic metrics for the evaluation of the models along two major dimensions: (1) style transfer and (2) content preservation and relevance (as discussed in Section 2.2). To measure politeness transfer quality, we compute *Polite Score*, which is defined as the average score given to the generated sequences by our politeness classifier, which we created by finetuning BERT (Devlin et al., 2019) on Madaan et al. (2020)'s Enron data (see Section 6.2.1).[2] Following prior work (Jin et al., 2022; Hu et al., 2022), we evaluate the relevance and content preservation using embedding similarity (Rahutomo et al., 2012) and BLEU score (Papineni et al., 2002). For embedding similarity, we use a pre-trained Sentence-BERT model (Reimers and Gurevych, 2019) and cosine similarity. We use BLEU-1 and BLEU-2 to account for the expected different phrasing in polite outputs and the high output variance common to open-domain dialogue response generation. As automated metrics for language generation do not correlate well with human judgments (Novikova et al., 2017), we conduct a small-scale in-house human evaluation with expert annotators (computational linguistics graduate students). We randomly select 50 context-utterance pairs from the DailyDialog test set for all models based on the strongest BlenderBot language model. The annotators rate model outputs using a 5-point Likert scale for politeness, coherence to context, and fluency.

6.3.2 Results

Politeness classification The accuracy of our BERT-based politeness classification model is 83.27% on the politeness transfer data. More importantly, the confusion matrix in Figure 6.2 shows that the model confuses mostly adjacent classes; the average error is only 0.98.

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

Models	PS	BLEU	CS
DailyDialog (DD)	5.41	–	–
DD + Madaan et al. (2020)	6.37	73.34	90.29
DD + Ours	**7.95**	70.21	89.07

Table 6.2: Evaluation of synthetic data generated using DailyDialog (DD) to train polite dialog models. We measure the Polite Score (PS), BLEU Score, and Content Similarity (CS) The BLEU and CS are measured between original utterances and polite-transferred utterances.

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

Table 6.3: Evaluation results of polite dialog models. We indicate what version of the DailyDialog dataset (DD) was used for Finetuning (FT) if any. We measure the Polite Score (PS), BLEU score, and Content Similarity (CS). BLEU Score (of n-gram = 1,2) and CS are computed between predicted polite utterances and the original utterances.

Politeness Transfer We compare the politeness transfer models on content preservation and politeness improvement using a test portion of Madaan et al. (2020)'s data, which consists of synthetic non-polite sentences and the corresponding original polite sentences. Models are tasked with producing polite sentences from synthetic non-polite ones; the result is then compared to the original human-written polite sentences. Table 6.1 shows the results. Our model achieves a higher politeness score than Madaan et al. (2020) while producing sentences more similar to the original human-written ones based on BLEU and sentence similarity scores. We also evaluate the performance of the politeness transfer models with respect to content preservation and politeness improvement on the synthetic pairs of contexts and polite utterances from the DailyDialog dataset we prepared. The results are shown in Table 6.2. Note that unlike in the previous experiment, we measure content preservation against the original (source) utterances. We observe that our model increases politeness over the source data and outperforms Madaan et al. (2020). We can see a slight drop in content preservation metrics against the original utterances, but this is expected as these metrics also reflect changes in phrasing.

Dialogue modeling Results of automatic metrics for dialogue modeling are shown in Table 6.3. The performance differences between the pre-trained models used are expected given the models' properties and intended use cases. While GPT-2 scores low on politeness, the dialogue-specific models obtain better results. As ex-

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

Table 6.4: A sample output using BlenderBot. Vanilla BlenderBot produces polite but irrelevant responses, and models finetuned on all DailyDialog data versions produce relevant responses, but ours is arguably the most polite.

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

Table 6.5: Human Evaluation on BlenderBot outputs. We measured politeness (Pol), coherent to context (CC), and fluency (Flu).

pected, all models perform much better in terms of content preservation after fine-tuning. Both ours and Madaan et al. (2020)'s politeness transfer result in an increase in politeness, and we can observe that our method consistently outperforms Madaan et al. (2020)'s. Moreover, our method is the only one that improves the Polite Score over the vanilla BlenderBot model. Finally, although the application of politeness transfer causes a decrease in content similarity with reference responses from DailyDialog, the drop is marginal, not consistent with all metrics, and could be caused by different phrasing, same as in the case of politeness transfer (cf. Table 6.2).

Human Evaluation We have evaluated 50 model outputs for each variant of the BlenderBot model (see Table 6.4 for a sample). The results are presented in Table 6.5. The human evaluation results mostly agree with our automatic evaluation results: our data preparation method performs better than Madaan et al. (2020)'s transfer in terms of politeness and is able to improve the base BlenderBot model. Both politeness-increasing methods cause a slight degradation in context coherency of the generated utterances; ours performs slightly worse in this aspect. However, our full approach yields more fluent outputs than the model trained on Madaan et al. (2020)'s politeness transfer.

6.4 Conclusion

In this chapter, we presented the Polite Chatbot as an exemplary application of TST. Our approach enhances dialogue models' politeness through a two-step training process: first, creating synthetic training corpora with increased politeness (leveraging the data scarcity solutions discussed in Chapter 3), and second, training the dialogue model. The resulting end-to-end dialogue response generation model does not require post-processing. Compared to multiple baselines for both politeness transfer and dialogue modeling, our model achieves increased politeness while still preserving important content.

Limitations and Future Work While the results are promising, we acknowledge certain limitations. The synthetic data generation process, though effective, may introduce artifacts that could affect the naturalness of the dialogue. Future work could adapt additional stylistic attributes such as empathetic or humorous responses and also expand the multilingual capabilities as discussed in Chapter 4, further advancing the versatility and applicability of TST in diverse dialogue systems.

7 Text Style Transfer using Large Language Models

This chapter is based on the paper *Are Large Language Models Actually Good at Text Style Transfer?* (Mukherjee et al., 2024c), published in the Proceedings of the 17th International Natural Language Generation Conference (INLG 2024).

Recent papers have identified a need for new Text Style Transfer (TST) methods that reduce training data requirements and expand the scope of supported styles (Jin et al., 2022; Hu et al., 2022). This makes prompting large language models (LLMs) a compelling option for TST (Liu et al., 2024a; Suzgun et al., 2022), as their ability to generalize across tasks and languages with minimal training has sparked growing interest in their application to style transfer (Liu et al., 2024a; Suzgun et al., 2022; Ostheimer et al., 2024; Reif et al., 2021; Brown et al., 2020; Huang et al., 2022). Although some promising results have emerged, especially in English, LLM effectiveness in multilingual and diverse stylistic settings remains underexplored (Fan et al., 2021; Sanh et al., 2022).

In this chapter, we explore how LLMs can be used for TST. We focus on tasks such as sentiment transfer and text detoxification in multiple languages, including English, Hindi, and Bengali, the same tasks and languages we also addressed in Chapter 4. We thoroughly examine how well LLMs perform in TST by looking at different methods, such as using the models without extra training (zero-shot), with in-context learning (few-shot), and with targeted fine-tuning. We compare their results with the state-of-the-art (SotA) models in TST work. Our evaluation, which includes automated metrics, evaluations using GPT-4, and human evaluation, show that some LLMs work well in English but partially in Hindi and Bengali. However, fine-tuning the models significantly improves their performance compared to zero-shot and few-shot prompting, making them comparable to the existing SotA models. This highlights the importance of having specific datasets albeit low-resource and specialized models to achieve effective TST.

This chapter is organized as follows. In Section 7.1, we describe the experimental setup. Section 7.2 presents the results and analysis, including both automatic, LLM and human evaluations. Finally, in Section 7.3, we discuss the limitations and suggest potential directions for future research.

7.1 Experiment Details

In this section, we provide comprehensive details about our experimental setup. We begin by outlining the datasets, languages, and tasks considered in our study, as discussed in Section 7.1.1. Next, we describe the LLMs utilized for the experiments in Section 7.1.2. Following this, we explain the configurations of the prompts employed, detailed in Section 7.1.3. Finally, we present the evaluation metrics used to assess model performance in Section 7.1.4.

7.1.1 Datasets, Languages, & Tasks

The experiments were conducted using datasets for sentiment transfer and text detoxification, which we presented in Chapter 4. Each dataset comprised 1,000 style-parallel examples, with splits for fine-tuning, development, and testing. The datasets used include English, Hindi, and Bengali for sentiment transfer, and English and Hindi for text detoxification. Following our previous experiments in Chapter 4, we use two popular TST subtasks where multilingual data is available. For sentiment transfer, experiments were conducted for both positive-to-negative and negative-to-positive tasks, with results averaged. For detoxification, we focused on the single task of transferring toxic to clean text.

7.1.2 Tested Models

For our experiments, we selected multiple freely available LLM architectures: BLOOM (BigScience Workshop, 2023; Muennighoff et al., 2023), ChatGLM (Du et al., 2022), Falcon (Penedo et al., 2023; Almazrouei et al., 2023), Llama (Touvron et al., 2023a,b; AI@Meta, 2024), Mistral (Jiang et al., 2023), OPT (Zhang et al., 2022), and Zephyr (Tunstall et al., 2023). They include a range of sizes (ca. ~0.5B-30B parameters, due to our computational constraints) and types, including foundation,

Model	Size Variants
BLOOM (BigScience Workshop, 2023)	560M, 1B, 3B, and 7B
BLOOMz (Muennighoff et al., 2023)	560M, 1B, 3B, and 7B
ChatGLM (Du et al., 2022)	6B
ChatGLM2 (Du et al., 2022)	6B
Falcon (Penedo et al., 2023; Almazrouei et al., 2023)	7B
Llama (Touvron et al., 2023a)	7B, 13B, and 30B
Llama-2 (Touvron et al., 2023b)	7B, and 13B
Llama-2-Chat (Touvron et al., 2023b)	7B, and 13B
Llama-3 (AI@Meta, 2024)	8B
Llama-3-Instruct (AI@Meta, 2024)	8B
Mistral-Instruct (Jiang et al., 2023)	7B
OPT (Zhang et al., 2022)	1.3B, 2.7B, 6.7B, 13B, and 30B
Zephyr (Tunstall et al., 2023)	7B

Table 7.1: List of open pre-trained LLMs used in our experiments, including their size variants.

LLMs	Zero-shot	Few-shot	Finetuning
BLOOM-560M	✓	✓	✓
BLOOM-1B	✓	✓	✓
BLOOM-3B	✓	✓	✓
BLOOM-7B	✓	✓	✓
BLOOMz-560M	✓	✓	✓
BLOOMz-1B	✓	✓	✓
BLOOMz-3B	✓	✓	✓
BLOOMz-7B	✓	✓	✓
Falcon-7B	✓	✓	✓
ChatGLM-6B	✓	✓	✗
ChatGLM2-6B	✓	✓	✓
GPT-3.5	✓	✓	✗
Llama-7B	✓	✓	✓
Llama-13B	✓	✓	✓
Llama-30B	✓	✓	✗
Llama-2-7B	✓	✓	✓
Llama-2-13B	✓	✓	✓
Llama-2-Chat-7B	✓	✓	✗
Llama-2-Chat-13B	✓	✓	✗
Llama-3-8B	✓	✓	✓
Llama-3-8B-Instruct	✓	✓	✗
Mistral-7B-Instruct	✓	✓	✗
OPT-1.7B	✓	✓	✓
OPT-2.7B	✓	✓	✓
OPT-6.7B	✓	✓	✓
OPT-13B	✓	✓	✓
OPT-30B	✓	✓	✗
Zephyr-7B	✓	✗	✗

Table 7.2: Details of LLMs used for zero-shot, few-shot, or fine-tuning scenarios. The model variant, including size and type (base/instructions/chat), is specified in the model name.

instruction-tuned and chat models (see Table 7.1). We obtained all models from HuggingFace (Wolf et al., 2020). We also included GPT-3.5 *(gpt-3.5-turbo)* accessed via the OpenAI API (OpenAI, 2023). As GPT-4 is used for evaluation, we did not use it for the TST task as LLMs may show bias towards their own outputs (Koo et al., 2023; Stureborg et al., 2024).

For benchmarking purposes, the results of the LLMs are compared to previous state-of-the-art (SotA) models specifically trained for sentiment transfer and text detoxification tasks, as detailed in Chapter 4. As SotA, we use the best performing models from our previous experiments: *Joint* and *Parallel* from Section 4.1.2 for sentiment transfer, and *Seq2seq + CLS_OP* and *KT* from Section 4.2.2 for text detoxification.

7.1.3 Model Setups and Prompt Examples

For each model, we evaluate three setups: zero-shot prompts (ZS), few-shot prompts (FS), and parameter-efficient finetuning (FT). Only base models are utilized for finetuning, excluding chat-based and instruction-tuned variants. The model variant, including size and type (base/instruction/chat), is indicated in the model name (see Tables 7.1 and 7.2).

This section also presents a collection of example prompts (in English) for the Text Sentiment Transfer (Table 7.3) and Text Detoxification (Table 7.4) tasks.

Parameter Optimization Due to the high computational cost of running LLMs, we did not conduct any extensive hyperparameter optimization. We ran limited preliminary experiments on the English and Hindi style transfer development set, opting to use default parameters from the Llama-Factory finetuning framework.[1] The only change made was increasing the number of finetuning epochs from 3 to 5, given the limited amount of data available. The same settings were then applied to both tasks and all languages.

7.1.4 Evaluation Metrics

To measure sentiment transfer and detoxification accuracy (ACC) in all experiments, we finetuned style classifiers for all languages and tasks based on *XLM-RoBERTa-base* (Conneau et al., 2020), using the training split of the same datasets. Table 7.8 presents the resulting classifier accuracies. In line with previous studies (as discussed in Section 2.2), we evaluate content retention through the BLEU score (Papineni et al., 2002) and content similarity (CS) (Rahutomo et al., 2012) compared to the input sentences. CS is computed using LaBSE sentence embeddings (Feng et al., 2022) and cosine similarity. Following Loakman et al. (2023) and Yang and Jin (2023), we use the arithmetic mean (AVG) of ACC and CS as a singular score for comparison.

[1] https://github.com/hiyouga/LLaMA-Factory

Prompt	
	Sentiment transfer changes the sentiment of a sentence while keeping non-sentiment-related content unchanged. Examples: Task: positive to negative Input: even when she didn't answer him quickly enough, he patiently waited on her. Output: when she didn't answer him quickly enough, he hung up on her. Task: negative to positive Input: there isn't a great selection of salads or pasta in the deli. Output: great selection of salads and pasta items in the deli. Task: positive to negative Input: they were very fair and respected me since i'm so young. Output: they tried to take advantage of me because i am young. Task: negative to positive Input: also the crab wontons are awful along with the bland plum sauce. Output: also the crab wontons rule along with the spicy plum sauce. Now change the sentiment of the following English sentence. Task: positive to negative Input: best customer service i have ever had. Output:

Table 7.3: A few-shot prompt used for Sentiment Transfer in English. It contains task definition, examples, instruction, and input.

To complement automatic metrics, we employed a GPT-4-based (*gpt-4-turbo;* (Achiam et al., 2023)) evaluation on a sample of 50 outputs from best LLMs according to automatic metrics, following prior work that showed good correlation with humans on machine translation (Kocmi and Federmann, 2023). We presented prompts used for GPT-4-based evaluations, covering Sentiment Transfer accuracy (Tables 7.5), content preservation (Tables 7.6), and fluency (Tables 7.7). We also conducted a small-scale in-house human evaluation on 50 outputs for best LLMs on the sentiment transfer task. Both humans and GPT-4 rated outputs on a 5-point Likert scale for style transfer accuracy, content preservation, and fluency.

Prompt	
	Text detoxification changes a toxic sentence to a clean sentence while keeping its general meaning unchanged. Examples: Input: oh , i m hearing shit like that from her too . Output: i've heard same things from her also Input: most likely he gave her a couple of pats on the butt and shes over dramatizing the whole thing . Output: Most like he gave her a couple of pats on her person and she's over dramatizing the whole thing Input: no . almost nobody is saying that except for those fucking idiots in the video . Output: no,almost nobody is saying that except for those in th video Input: that said , i hope they get a great attorney and sue the shit out of the cops . Output: that said , i hope they get a great attorney and sue the cops . Now detoxify the following English sentence. Input: DIGIT year olds can be little shits too , doesn t mean you fight them . Output:

Table 7.4: A few-shot prompt used for Detoxification in English. It contains task definition, examples, instruction, and input.

Prompt	
	Sentiment transfer task: transfer the sentiment of a sentence (from positive to negative or negative to positive) while keeping the rest of the sentiment-independent content unchanged. Please rate the sentiment transfer accuracy of the negative to positive sentiment transfer task between the following English source sentence S1 and the sentiment-transferred sentence S2. Use a scale of 1 to 5, where 1 indicates that the sentiment in S1 is completely identical to the sentiment in S2, and 5 indicates that the sentiment has been completely transferred to the target sentiment in S2. S1: so he can charge a bloody fortune for them. S2: so he can charge a fair amount of money for them. Sentiment transfer accuracy rating (on a scale of 1 to 5) =

Table 7.5: A few-shot prompt for Sentiment Transfer Accuracy evaluation in Sentiment Transfer in English. It contains task definition, instruction, and input.

Prompt	
	Sentiment transfer task: transfer the sentiment of a sentence (from positive to negative or negative to positive) while keeping the rest of the content unchanged. Please rate the content preservation between the following English source sentence S1 and the sentiment-transferred sentence S2 for the negative to positive sentiment transfer task on a scale of 1 to 5, where 1 indicates very low content preservation and 5 indicates very high content preservation. To determine the content preservation between these two sentences, consider only the information conveyed by the sentences and ignore any differences in sentiment due to the negative to positive sentiment transfer. S1: so he can charge a bloody fortune for them. S2: so he can charge a fair amount of money for them. Content Preservation rating (on a scale of 1 to 5) =

Table 7.6: A few-shot prompt for Content Preservation evaluation in Sentiment Transfer in English. It contains task definition, instruction, and input.

Prompt	
	Please rate the fluency of the following English sentence S on a scale of 1 to 5, where 1 represents poor fluency, and 5 represents excellent fluency. S: so he can charge a fair amount of money for them. Fluency rating (on a scale of 1 to 5) =

Table 7.7: A few-shot prompt for Fluency evaluation in Sentiment Transfer in English. It contains instruction, and input.

Language	Sentiment acc. (%)	Toxicity acc. (%)
English	93.4	94.8
Hindi	89.3	70.9
Bengali	87.8	-

Table 7.8: Language-wise sentiment and toxicity classifier's accuracy (acc.) scores.

7.2 Results and Analysis

This section discusses results and analysis based on the automatic metrics (Section 7.2.1), as well as human and GPT-4-based assessments (Section 7.2.2).

7.2.1 Automatic Evaluation

We show abridged results (with mostly 7B-parameter LLM size variants) in Table 7.9. Full results are provided in Table 6 from Appendix A of our published paper (Mukherjee et al., 2024c) (however, we draw on the full results in the following description).

Impact of Methodology GPT-3.5 consistently outperforms other models on zero-shot prompting across all languages, achieving the highest accuracy and average scores. Other models, such as ChatGLM2-6B and Llama-3-8B-ZS, also show strong performance, particularly in English. However, models like BLOOMz-7B and OPT-

6.7B reach much lower scores, suggesting limited zero-shot capabilities. Few-shot prompting generally improves performance compared to zero-shot, especially in English. GPT-3.5 stays in the lead, with high scores in all languages. Finetuning brings the highest gains across the board, with strong performance from most LLMs, including ones weak at zero-shot and few-shot, such as BLOOM-7B. Most finetuned LLMs are comparable to prompted GPT-3.5 and previous SOTA models.

Language-wise Analysis Across the three languages, English consistently shows the highest performance. Hindi, while more challenging, benefits significantly from few-shot and finetuning approaches (e.g., for GPT-3.5 and BLOOM-7B). Bengali presents the greatest difficulty, reflecting the scarcity of high-quality training data, but still shows marked improvements with additional training. GPT-3.5 and Llama-3-8B lead in performance across all settings. The results highlight the importance of model adaptation with targeted datasets in multilingual settings.

Impact of Model Variant Generally, larger models score better across the board, but gains diminish with increasing size: The jump from 1B to 3B shows a significant boost; improvements from 3B to 7B and 7B to 13B are less pronounced; 30B models do not improve over their smaller counterparts. The impact of model size is most pronounced in zero-shot scenarios: while small models (<1B) struggle, medium-sized models (2B-3B) show substantially better zero-shot capabilities. Instruction-tuned and chat models work better than their base variants in zero- and few-shot settings, but this depends on the task: for detoxification, Llama-3-8B-Instruct simply refused to provide outputs.[2]

Style vs. Content Different models show different sides of the tradeoff between ACC and CS, with ChatGLM2-6B and Zephyr-7B reaching high transfer accuracy but lagging on content preservation, while BLOOM-7B, Llama-3-8B-Instruct or Falcon-7B are the opposite.

7.2.2 GPT-4-based and Human Evaluation

We selected open models performing best in English for each methodology, alongside GPT-3.5 and previous SotA, for GPT-4-based evaluation on both tasks (see Table 7.10). We kept the same models for human evaluation, but limited the experiment to sentiment transfer only (see Table 7.11). Ouput samples of sentiment transfer and detoxification are shown in Tables 7.12 and 7.13 respectively.

[2] A typical response was: "I cannot detoxify a sentence that contains sexual content. Is there something else I can help you with?"

Models	Sentiment Transfer English ACC	CS	BL	AVG	Sentiment Transfer Hindi ACC	CS	BL	AVG	Sentiment Transfer Bengali ACC	CS	BL	AVG	Detoxification English ACC	CS	BL	AVG	Detoxification Hindi ACC	CS	BL	AVG
BLOOM-7B-ZS	37.8	77.4	39.8	51.6	26.6	79.4	39.6	48.6	34.4	78.8	30.3	47.8	8.6	76.1	39.0	41.2	52.2	79.1	39.8	57.0
BLOOMz-7B-ZS	26.0	40.3	12.6	26.3	31.6	35.9	4.0	23.9	35.2	35.1	2.5	24.2	14.2	69.1	34.4	39.2	64.8	69.8	30.5	55.0
ChatGLM2-6B-ZS	86.3	64.4	16.9	55.8	53.0	55.9	5.1	38.0	48.5	35.2	0.4	28.0	96.2	47.6	7.4	50.4	77.8	53.6	4.3	45.2
Falcon-7B-ZS	72.8	75.0	40.9	62.9	21.5	70.2	30.8	40.8	22.1	63.9	17.7	34.6	46.6	75.2	38.2	53.3	65.4	60.7	27.3	51.1
GPT-3.5-ZS	93.4	81.4	43.9	72.9	83.4	82.7	43.3	69.8	79.9	81.7	31.8	64.5	99.2	73.9	30.1	67.7	80.2	79.3	39.7	66.4
Llama-7B-ZS	36.8	65.9	23.3	42.0	22.2	80.2	41.4	47.9	12.0	78.2	30.9	40.4	11.6	73.2	37.0	40.6	52.6	79.7	42.4	58.2
Llama-2-7B-ZS	63.1	75.5	42.0	60.2	44.6	79.9	41.4	55.3	26.9	76.6	29.5	44.3	20.6	74.7	37.5	44.3	53.2	78.7	41.0	57.7
Llama-2-Chat-7B-ZS	94.0	78.0	38.4	70.1	65.2	78.5	37.2	60.3	39.0	71.6	21.5	44.0	82.8	70.4	25.9	59.7	61.8	76.9	38.1	58.9
Llama-3-8B-ZS	76.9	80.4	45.9	67.7	66.2	81.8	42.9	63.6	58.4	76.2	30.4	55.0	25.4	73.1	34.7	44.4	56.6	77.4	35.8	56.6
Llama-3-8B-Instruct-ZS	92.2	69.3	35.0	65.5	71.6	59.0	23.0	51.2	50.1	64.6	24.2	46.3	-	-	-	-	-	-	-	-
Mistral-7B-Instruct-ZS	80.8	65.8	29.3	58.6	32.2	78.8	36.4	49.1	22.8	74.6	22.6	40.0	89.4	72.1	33.1	64.9	61.8	72.0	30.8	54.9
OPT-6.7B-ZS	54.1	24.3	1.4	26.6	17.3	60.0	28.9	35.4	13.5	76.8	30.0	40.1	83.0	27.4	0.7	37.0	66.6	59.1	33.1	52.9
Zephyr-7B-ZS	85.0	71.4	23.1	59.8	66.7	71.6	31.2	56.5	55.2	67.5	20.9	47.9	96.8	54.6	13.2	54.9	71.8	63.7	21.4	52.3
BLOOM-7B-FS	32.1	78.8	43.5	51.5	24.5	80.2	40.1	48.3	16.9	77.9	29.6	41.5	22.4	77.1	41.1	46.9	52.0	79.6	41.6	57.7
BLOOMz-7B-FS	35.2	74.3	39.3	49.6	36.4	80.4	41.3	52.7	29.0	78.7	30.8	46.2	14.4	71.4	36.9	40.9	59.4	72.9	37.7	56.7
ChatGLM2-6B-FS	87.8	75.6	32.4	65.3	48.6	62.7	10.4	40.6	41.9	40.0	0.7	27.6	89.2	64.9	16.9	57.0	73.0	54.4	6.6	44.7
Falcon-7B-FS	77.6	79.6	46.2	67.8	15.9	78.4	39.8	44.7	17.8	73.4	27.3	39.5	24.2	75.9	39.9	46.7	56.4	75.5	40.2	57.3
GPT-3.5-FS	95.1	81.4	44.7	73.7	90.2	82.5	41.3	71.3	84.2	81.1	31.9	65.7	96.6	77.2	38.6	70.8	80.0	80.2	39.7	66.6
Llama-7B-FS	64.8	59.4	30.3	51.5	31.8	79.7	40.5	50.7	23.1	77.3	29.3	43.2	11.6	76.9	40.1	42.9	53.4	79.9	42.6	58.6
Llama-2-7B-FS	54.9	32.2	3.0	30.0	54.1	78.2	37.0	56.4	39.3	73.6	26.1	46.3	46.8	61.1	34.3	47.4	53.4	77.6	38.0	56.3
Llama-2-Chat-7B-FS	92.1	74.5	36.2	67.6	69.0	75.2	29.6	57.9	38.1	65.6	19.2	40.9	78.8	62.6	28.2	56.5	61.4	76.1	34.1	57.2
Llama-3-8B-FS	67.9	43.3	12.5	41.3	71.7	80.2	39.7	63.9	60.2	73.5	29.7	54.4	40.2	74.4	41.8	52.2	80.4	51.6	20.2	50.7
Llama-3-8B-Instruct-FS	52.2	11.1	1.4	21.6	1.2	15.7	0	5.6	50.0	14.4	0	21.5	-	-	-	-	-	-	-	-
Mistral-7B-Instruct-FS	87.3	77.3	39.7	68.1	33.7	77.8	34.2	48.6	36.5	75.2	25.4	45.7	92.2	74.5	32.6	66.5	61.2	76.9	37.4	58.5
OPT-6.7B-FS	33.9	63.4	28.0	41.8	11.4	77.5	39.3	42.7	15.1	75.8	29.4	40.1	11.2	75.4	39.3	42.0	57.0	70.6	37.2	54.9
BLOOM-7B-FT	91.2	80.6	43.2	71.7	83.9	81.0	40.4	68.4	81.7	75.6	26.3	61.2	92.4	75.8	41.7	70.0	82.0	76.6	33.8	64.1
BLOOMz-7B-FT	91.0	80.3	45.0	72.1	85.3	81.0	39.8	68.7	85.9	75.3	19.4	60.2	92.4	75.6	40.7	69.6	82.0	76.4	32.2	63.5
ChatGLM2-6B-FT	86.8	78.8	41.9	69.2	51.9	74.1	32.8	52.9	42.1	48.1	7.8	32.7	90.0	74.0	34.2	66.1	67.8	69.3	30.3	55.8
Falcon-7B-FT	88.3	79.6	43.1	70.3	37.7	76.2	35.8	49.9	40.8	51.0	8.3	33.4	87.6	73.8	37.8	66.4	68.8	61.3	21.4	50.5
Llama-7B-FT	91.5	81.6	47.2	73.4	69.4	78.5	39.4	62.4	41.9	76.0	28.4	48.8	91.8	76.1	42.4	70.1	67.4	73.9	36.2	59.2
Llama-2-7B-FT	92.9	81.2	46.5	73.5	77.5	78.6	39.2	65.1	56.7	76.1	27.9	53.6	92.4	76.2	43.3	70.6	68.8	74.6	36.2	59.9
Llama-2-13B-FT	92.0	82.0	47.3	73.8	79.6	80.2	40.0	66.6	61.2	77.4	29.4	56.0	95.6	76.1	42.8	71.5	73.8	75.5	36.3	61.9
Llama-3-8B-FT	92.0	81.4	46.8	73.4	85.7	82.1	42.4	70.1	81.9	80.2	32.3	64.8	96.8	76.9	45.2	73.0	83.2	78.0	37.2	66.1
OPT-6.7B-FT	91.7	80.6	44.5	72.3	29.1	76.8	38.3	48.1	22.5	76.3	27.6	42.1	95.8	76.7	42.2	71.6	58.2	76.1	39.8	58.0
SOTA (*Joint*)	84.5	81.5	46.1	70.7	78.3	82.5	43.8	68.2	80.3	78.0	28.1	62.1								
SOTA (*Parallel*)	80.9	81.5	46.4	69.6	85.4	82.3	44.3	70.7	73.1	81.0	34.7	62.9								
SOTA (*CLS-OP*)													91.6	76.6	44.2	70.8	65.0	78.2	39.8	61.0
SOTA (*KT*)													92.0	77.5	45.6	71.7	76.6	78.6	42.0	65.7

Table 7.9: Automatic metrics results: style accuracy (ACC), content similarity (CS), and BLEU (BL) against the source, and an average of all three (AVG). Only models close to 7B parameters in size are shown (with added GPT-3.5 and Llama-2-13B-FT, with the best sentiment transfer performance in its category). The best results in each category are highlighted in color.

Both evaluations show better performance for finetuned LLMs and previous SotA, compared to prompted LLMs. In some cases, finetuned LLMs outperform GPT-3.5, particularly in terms of content preservation. Hindi and Bengali show lower performance than English, which suggests that more targeted resources for these languages are needed.

Models	Sentiment transfer									Detoxification					
	English			Hindi			Bengali			English			Hindi		
	Sty.	Cont.	Flu.	Sty.	Cont.	Flu.	Sty.	Cont.	Flu.	Sty.	Cont.	Flu.	Sty.	Cont.	Flu.
GPT-3.5-ZS	4.60	4.52	4.28	4.18	4.64	3.62	4.14	4.84	3.34	4.26	4.38	3.88	3.46	4.38	2.76
Llama-2-7B-Chat-ZS	4.96	4.50	4.26	3.22	3.74	2.64	1.50	2.16	2.20						
Mistral-7B-Instruct-ZS										3.08	4.20	3.90	1.52	4.32	2.32
GPT-3.5-FS	4.68	4.58	3.92	4.74	4.60	3.72	4.42	4.50	3.22	4.02	4.72	3.88	3.44	4.40	2.94
Mistral-7B-Instruct-FS	4.16	4.28	3.98	2.26	4.00	3.02	1.78	3.62	2.62	3.36	4.66	3.82	1.62	3.98	2.18
Llama-2-13B-FT	4.70	4.44	3.96	4.16	4.20	3.32	2.98	3.32	2.60						
Llama-3-8B-FT										3.92	4.44	3.40	3.22	4.08	2.88
SOTA (*Joint*)	4.14	4.26	3.56	4.04	4.60	3.48	3.62	4.04	2.84						
SOTA (*KT*)										3.42	4.24	3.26	2.30	4.52	2.62

Table 7.10: GPT-4-based evaluation of 50 randomly selected outputs on style accuracy (Sty.), content preservation (Cont.), and fluency (Flu.; see Section 7.2.2). The best results overall are highlighted in color.

Models	English			Hindi		
	Style	Content	Fluency	Style	Content	Fluency
GPT-3.5-ZS	4.66	4.96	4.92	4.18	4.92	4.90
Llama-2-7B-Chat-ZS	4.90	4.86	4.88	3.54	4.86	4.22
GPT-3.5-FS	4.66	4.98	4.92	4.72	4.88	4.80
Mistral-7B-Instruct-FS	4.30	4.78	4.82	2.26	4.70	4.62
Llama-2-13B-FT	4.68	4.90	4.86	4.34	4.84	4.80
SOTA (*Joint*)	4.22	4.94	4.60	3.96	4.94	4.90

Table 7.11: Human evaluation of 50 randomly selected outputs on sentiment accuracy (Style), content preservation (Content), and Fluency (see Section 7.2.2). The best results overall are highlighted in color.

Models	Negative → Positive	Positive → Negative
Reference	en: also matt wasn't that cool and not that helpful. → also matt wasn't that cool, and not that helpful. also matt was super cool and helpful. hi: साथ ही मैट कूल–वूल नहीं था और न ही मददगार। → साथ ही मैट सुपर कूल और मददगार था। bn: এছাড়াও, ম্যাট খুব একটা ভালো বা উপকারী ছিল না । → এছাড়াও, ম্যাট খুব ভালো এবং উপকারী ছিল ।	en: thank you amanda, i will be back ! → no thanks amanda, i won't be back ! hi: धन्यवाद अमांडा, मैं वापस आऊंगा! → भाड़ में जाओ अमांडा, मैं वापस नहीं आऊँगा! bn: ধন্যবাদ আমানডা, আমি আবার ফিরে আসব ! → না ধন্যবাদ আমানডা, আমি আর ফিরে আসব না !
GPT-3.5-ZS	en: also matt was really cool, and very helpful. hi: साथ ही मैट कूल–वूल था और मददगार। bn: এছাড়াও, ম্যাট খুব একটা ভালো বা উপকারী ছিল।	en: I will not be back, Amanda, thank you. hi: शुक्रिया अमांडा, मैं वापस आऊंगा! bn: দুঃখিত আমানডা, আমি আবার ফিরে আসব।
Llama-2-7B-Chat-ZS	en: Also, Matt wasn't that cool, and he was very helpful. hi: साथ ही मैट कूल–वूल था और मददगार। bn: এছাড়াও, ম্যাট খুব একটা ভালো বা উপকারী ছিল না ।	en: 😠 amanda, i will be back ! hi: निर्भया अमांडा, मैं नहीं आऊंगा! bn: ধন্যবাদ আমানডা, আমি আবার ফিরে আসব ! আমানডা আমানডা, আমি আবার ফিরে আসব !
GPT-3.5-FS	en: also matt was really cool and super helpful. hi: साथ ही मैट कूल–वूल था और मददगार भी। bn: এছাড়াও, ম্যাট খুব একটা ভালো এবং উপকারী ছিল ।	en: thank you amanda, i will never be back ! hi: धन्यवाद अमांडा, मैं वापस नहीं आऊंगा। bn: দুঃখিত আমানডা, আমি আবার ফিরে আসব।
Mistral-7B-Instruct-FS	en: also matt was very cool and helpful. hi: साथ ही मैट कूल–वूल था और मददगार के साथ मैट कूल–वूल हो गया। bn: এছাড়াও, ম্যাট খুব একটা ভালো বা উপকারী ছিল ।	en: Thank you Amanda, I won't be back. hi: धन्यवाद अमांडा, मैं वापस आउंगा! bn: ধন্যবাদ আমানডা, আমি আবার ফিরে আসব না ।
Llama-2-13B-FT	en: also matt was super cool and super helpful. hi: साथ ही मैट कूल–वूल था और न ही मददगार। bn: এছাড়াও, ম্যাট খুব একটা ভালো বা উপকারী ছিল ।	en: thanks amanda, i will never return. hi: इस किसी भी व्यवसाय के लिए एक बेहतर जगह नहीं है, अमांडा! bn: আমানডা, আমি আবার ফিরে আসব না !
SotA (*Joint*)	en: also matt was cool, and very helpful. hi: साथ ही मैट बहुत साफ और मददगार था। bn: এছাড়াও, ম্যাট খুব একটা ভালো এবং উপকারী ছিল ।	en: sorry amanda, i will not be back. hi: धन्यवाद अमांडा, मैं वापस नहीं आऊंगा! bn: ধন্যবাদ আমানডা, আমি আর ফিরে আসব না !

Table 7.12: Sample outputs for the Sentiment Transfer task (positive to negative and negative to positive) in English, Hindi, and Bengali, generated by a selection of top-performing models (see Section 7.2.2).

Models	**Toxic → Clean**	**Toxic → Clean**
Reference	en: so shut the hell up until you know what it is your talking about . → So keep quiet untill you know what you are talking about. hi: इसलिए अपना मूँह बंद रख जब तक तू नहीं जानता कि यह तू क्या बात कर रहा है। → इसलिए चुप रहो जब तक आप जानते नहीं कि आप किस बारे में बात कर रहे हैं।	en: p. s im a fucking cunt rag and enjoy the taste of a nice warm pile of shit every morning . → i enjoy the tasty nice warm pile every morning hi: नोट करें: मैं एक नीच दर्जे की रंडी हूँ और हर रोज़ सुबह टट्टी के नाश्ते के मज़े लेती हूँ। → मैं हर रोज़ सुबह नाश्ते का मज़ा लेती हूँ।
GPT-3.5-ZS	en: Please refrain from speaking until you have a clear understanding of the topic. hi: इसलिए अपना मूँह बंद रख जब तक तू नहीं जानता कि यह तू क्या बात कर रहा है।	en: p. s I am a person who enjoys the taste of a nice warm breakfast every morning. hi: नोट करें: मैं एक नीच दर्जे की और हर रोज़ सुबह के नाश्ते के मज़े लेती हूँ।
Mistral-7B-Instruct-ZS	en: "So, shut up until you know what you're talking about." hi: इसलिए अपना मूँह बंद रख जब तक तू नहीं जानता कि यह तू क्या बात कर रहा है।	en: p. s im a cunt rag and enjoy the taste of a nice warm pile of shit every morning . hi: नोट करें: मैं एक नीच दर्जे की रंडी हूं और हर रोज़ सुबह टट्टी के नाश्ते के मज़े लेती हूं।
GPT-3.5-FS	en: so be quiet until you know what you're talking about. hi: इसलिए अपना मूँह बंद रख जब तक तू नहीं जानता कि यह तू क्या कह रहा है।	en: P.S. I enjoy the taste of a nice warm pile of shit every morning. hi: नोट करें: मैं एक नीच दर्जे की और हर रोज़ सुबह टट्टी के नाश्ते के मज़े लेती हूँ।
Mistral-7B-Instruct-FS	en: so shut up until you know what you're talking about. hi: इसलिए अपना मूँह बंद रख जब तक तू नहीं जानता कि यह तू क्या बात कर रहा है।	en: p. s. i am a cunt rag and enjoy the taste of a nice warm pile of shit every morning. hi: नोट करें: मैं एक नीच दर्जे की रंडी हूं और हर रोज़ सुबह टट्टी के नाश्ते के मज़े लेती हूं।
Llama-3-8B-FT	en: So please don't speak until you know what you are talking about. hi: इसलिए अपना मुँह बंद रख जब तक आप यह नहीं जानते कि आप क्या बात कर रहे हैं।	en: p.s i am a bad person and enjoy the taste of a nice warm pile of shit every morning hi: नोट करें: मैं एक नीच दर्जे की रंडी हूँ और हर रोज़ सुबह टट्टी के नाश्ते के मज़े लेती हूँ।
SotA (*KT*)	en: so wait until you know what it is your talking about hi: इसलिए अपना मूँह बंद रख जब तक तक तू नहीं जानता कि यह तू क्या बात कर रहा है	en: p. s im a bad guy and enjoy the taste of a nice warm pile of things every morning. hi: मैं एक नीच दर्जे की रंडी हूँ और हर रोज़ सुबह टट्टी के नाश्ते के मज़ा लेती हूँ

Table 7.13: Sample outputs from a few selected top-performing models (see Section 7.2.2) for the Text Detoxification task in English and Hindi are provided. Content warning: This table contains examples that are toxic, and/or offensive, and/or sexist in nature.

7.3 Conclusion

In this chapter, we evaluated the effectiveness of LLMs in TST, focusing specifically on sentiment transfer and text detoxification in English, Hindi, and Bengali. We examined the performance of LLMs under zero-shot and few-shot prompting as well as through parameter-efficient fine-tuning. Our results show that while some open LLMs achieve promising performance in English, their capabilities in multilingual contexts remain limited. However, fine-tuning yields significant improvements, bringing these models closer to or even aligning them with previous state-of-the-art systems. These findings complement the strategies for addressing data scarcity discussed in Chapter 3 and build upon the multilingual settings explored in Chapter 4. This work highlights the value of tailored datasets and task-specific models (even of smaller sizes) for achieving effective TST.

Limitations and Future Work While our study sheds light on the performance of LLMs in TST across multiple languages, it has certain limitations. Our evaluation focused on sentiment transfer and text detoxification, excluding other TST tasks such as formality, humor, and sarcasm. Additionally, our analysis is limited by the available datasets and may not fully capture the diversity of linguistic styles and cultural nuances present across different languages.

In future work, we aim to extend our experiments to include a broader range of styles and languages. We also plan to explore alternative fine-tuning techniques (Liu et al., 2024b; Jain et al., 2023) and advanced prompting strategies (Yao et al., 2024; Wei et al., 2022) to further enhance model performance in TST.

8
Conclusions

In this thesis, we explored the evolving landscape of *Text Style Transfer* (TST), examining its potential to refine natural language generation by effectively altering text styles while preserving core content. Based on the research questions (as discussed in Section 1.1) – addressing challenges in data scarcity, multilingualism, evaluation, practical applications, and the role of large language models (LLMs) in TST – we presented the following research contributions and insights:

- **RQ1**: We developed methods to perform TST without direct supervision, utilizing neural language models using non-parallel and limited parallel data (as discussed in Chapter 3).
- **RQ2**: To bridge the multilingual TST gap, we constructed datasets and models for diverse Indian languages beyond English, fostering low-resource style transfer and establishing benchmarks for multilingual TST (as discussed in Chapter 4).
- **RQ3**: We introduced refined evaluation metrics that improve upon content preservation and style accuracy, correlating more closely with human judgment than existing approaches (as discussed in Chapter 5).
- **RQ4**: Our polite chatbot work demonstrated TST's potential in real-world applications, successfully using style transfer to make dialogue systems more adaptable and user-engaging (as discussed in Chapter 6).
- **RQ5**: Through systematic evaluation, we assessed LLM performance in TST tasks, providing a nuanced view of their strengths and limitations in TST tasks (as discussed in Chapter 7).

Our findings improved on previous state-of-the-arts and open pathways for integrating style transfer capabilities in NLP applications. However, several challenges remain, and advancing TST will require addressing issues tied to both linguistic nuance and computational scalability.

8.1 Challenges and Future Work

Despite the progress made, TST poses enduring challenges. Future research directions will need to address these limitations to enable practical use of TST in applications:

- **Multi-Attribute Style Transfer:** Extending TST to handle multiple style attributes simultaneously remains an open problem (Subramanian et al., 2018; Hu et al., 2023). Real-world language often involves blending multiple style aspects (e.g., formality, politeness, and sentiment). Future work could explore models that can dynamically adjust to multiple style dimensions.

- **Long-context Style Transfer:** While existing TST research has demonstrated effectiveness on short texts (Hu et al., 2022), including our own in this thesis, scaling to long-form content remains challenging due to the need for maintaining style coherence and content relevance across extended passages (Qian, 2020). Additionally, current TST models typically operate at the sentence level, lacking the ability to incorporate broader contextual dependencies, which is crucial for applications requiring multi-turn interactions or document-level consistency (Wu et al., 2023).

- **Culturally-aware TST** Addressing cultural nuances in TST is essential, as stylistic conventions vary significantly across cultures (Hershcovich et al., 2022). Factors such as politeness and formality are often associated with different communicative styles across cultures (Brown and Levinson, 1987).

- **Evaluation Metrics and Human-Alignment:** While we introduced more nuanced evaluation measures, creating universally accepted metrics for TST tasks remains an ongoing challenge – a problem shared across the other NLG tasks (Tikhonov et al., 2019; Celikyilmaz et al., 2020; Ostheimer et al., 2023). Research on aligning automated metrics more closely with human judgment, as well as exploring evaluations in real-world scenarios, will be crucial for assessing TST's effectiveness and reliability.

- **LLM-Driven TST Applications and Adaptability:** As LLMs become increasingly prevalent, understanding their role in TST and optimizing their performance across languages and styles will be vital. Investigating methods to fine-tune or adapt LLMs specifically for TST tasks across various sociolinguistic contexts could enhance their effectiveness and broaden their application scope in practical TST systems (Franceschelli and Musolesi, 2024).

8.2 Final Thoughts

TST stands at the intersection of linguistic versatility and computational innovation, embodying the transformative potential of natural language processing to bridge stylistic diversity in human communication. While TST models have made remarkable progress, capturing the nuances of style attributes in a way that aligns with human expectations remains a complicated task. In most applications, reliable content preservation is crucial, yet achieving it remains a tradeoff with stylistic accuracy, posing an ongoing challenge for current TST models.

Moving forward, the continued exploration of data-efficient methods, multilingual adaptability, and culturally and contextually aware TST will play a crucial role in advancing this field. The pursuit of robust, human-aligned TST systems may ultimately lead to NLP technologies capable of adapting seamlessly across domains, audiences, and applications. As TST research progresses, it promises not only to enrich automated language generation but also to redefine our interactions with language technologies in ways that respect and reflect the diversity of human expression.

Bibliography

Achiam, J. – Adler, S. – Agarwal, S. – Ahmad, L. – Akkaya, I. – Aleman, F. L. – Almeida, D. – Altenschmidt, J. – Altman, S. – Anadkat, S. – others. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. 2023.

AI@Meta. Llama 3 Model Card, 2024. Available at: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Almazrouei, E. – Alobeidli, H. – Alshamsi, A. – Cappelli, A. – Cojocaru, R. – Debbah, M. – Goffinet, E. – Heslow, D. – Launay, J. – Malartic, Q. – Noune, B. – Pannier, B. – Penedo, G. Falcon-40B: an open large language model with state-of-the-art performance. Huggingface Model Release, 2023. Available at: https://huggingface.co/tiiuae/falcon-40b.

André, E. – Rehm, M. – Minker, W. – Bühler, D. Endowing Spoken Language Dialogue Systems with Emotional Intelligence. In André, E. – Dybkjær, L. – Minker, W. – Heisterkamp, P. (Ed.) *Affective Dialogue Systems, Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004, Proceedings*, 3068 / *Lecture Notes in Computer Science*, p. 178–187. Springer, 2004. doi: 10.1007/978-3-540-24842-2_17. Available at: https://doi.org/10.1007/978-3-540-24842-2_17.

Araabi, A. – Monz, C. Optimizing Transformer for Low-Resource Neural Machine Translation. In Scott, D. – Bel, N. – Zong, C. (Ed.) *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, p. 3429–3435. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.304. Available at: https://doi.org/10.18653/v1/2020.coling-main.304.

Ba, J. L. Layer normalization. *arXiv preprint arXiv:1607.06450*. 2016.

Banarescu, L. – Bonial, C. – Cai, S. – Georgescu, M. – Griffitt, K. – Hermjakob, U. – Knight, K. – Koehn, P. – Palmer, M. – Schneider, N. Abstract Meaning Representation for Sembanking. In Pareja-Lora, A. – Liakata, M. – Dipper, S. (Ed.) *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, p. 178–186, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. Available at: https://aclanthology.org/W13-2322/.

Banerjee, S. – Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Goldstein, J. – Lavie, A. – Lin, C.-Y. – Voss, C. (Ed.) *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, p. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. Available at: `https://aclanthology.org/W05-0909/`.

Belz, A. – Thomson, C. – Reiter, E. – Mille, S. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, p. 3676–3687, 2023.

Belz, A. – Agarwal, S. – Shimorina, A. – Reiter, E. ReproGen: Proposal for a Shared Task on Reproducibility of Human Evaluations in NLG. In Davis, B. – Graham, Y. – Kelleher, J. – Sripada, Y. (Ed.) *Proceedings of the 13th International Conference on Natural Language Generation*, p. 232–236, Dublin, Ireland, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.inlg-1.29. Available at: `https://aclanthology.org/2020.inlg-1.29/`.

Bengio, Y. – Simard, P. – Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*. 1994, 5, 2, p. 157–166.

Bengio, Y. – Ducharme, R. – Vincent, P. – Janvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* March 2003, 3, null, p. 1137–1155. ISSN 1532-4435.

BigScience Workshop. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*. 2023.

Bommasani, R. – Hudson, D. A. – Adeli, E. – Altman, R. – Arora, S. – Arx, S. – Bernstein, M. S. – Bohg, J. – Bosselut, A. – Brunskill, E. – others. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. 2021.

Briakou, E. – Agrawal, S. – Tetreault, J. – Carpuat, M. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. *arXiv preprint arXiv:2110.10668*. 2021a.

Briakou, E. – Agrawal, S. – Zhang, K. – Tetreault, J. – Carpuat, M. A review of human evaluation for style transfer. *arXiv preprint arXiv:2106.04747*. 2021b.

Briakou, E. – Lu, D. – Zhang, K. – Tetreault, J. Olá, Bonjour, Salve! XFORMAL: A Benchmark for Multilingual Formality Style Transfer. In Toutanova, K. – Rumshisky, A. – Zettlemoyer, L. – Hakkani-Tur, D. – Beltagy, I. – Bethard, S. – Cotterell, R. – Chakraborty, T. – Zhou, Y. (Ed.) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 3199–3216, Online, June 2021c. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.256. Available at: `https://aclanthology.org/2021.naacl-main.256/`.

BROWN, P. – LEVINSON, S. C. *Politeness: Some Universals in Language Usage.* Studies in Interactional Sociolinguistics. Cambridge University Press, 1987.

BROWN, T. – MANN, B. – RYDER, N. – SUBBIAH, M. – KAPLAN, J. D. – DHARIWAL, P. – NEELAKANTAN, A. – SHYAM, P. – SASTRY, G. – ASKELL, A. – OTHERS. Language models are few-shot learners. *Advances in neural information processing systems.* 2020, 33, p. 1877–1901.

CAI, S. – KNIGHT, K. Smatch: an Evaluation Metric for Semantic Feature Structures. In SCHUETZE, H. – FUNG, P. – POESIO, M. (Ed.) *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 748–752, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. Available at: https://aclanthology.org/P13-2131/.

CAO, S. – WANG, L. Inference Time Style Control for Summarization. In TOUTANOVA, K. – RUMSHISKY, A. – ZETTLEMOYER, L. – HAKKANI-TUR, D. – BELTAGY, I. – BETHARD, S. – COTTERELL, R. – CHAKRABORTY, T. – ZHOU, Y. (Ed.) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 5942–5953, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.476. Available at: https://aclanthology.org/2021.naacl-main.476/.

CAO, Y. – SHUI, R. – PAN, L. – KAN, M. – LIU, Z. – CHUA, T. Expertise Style Transfer: A New Task Towards Better Communication between Experts and Laymen. In JURAFSKY, D. – CHAI, J. – SCHLUTER, N. – TETREAULT, J. R. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, p. 1061–1071. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.100. Available at: https://doi.org/10.18653/v1/2020.acl-main.100.

CELIKYILMAZ, A. – CLARK, E. – GAO, J. Evaluation of Text Generation: A Survey. *CoRR.* 2020, abs/2006.14799. Available at: https://arxiv.org/abs/2006.14799.

CHAI, J. – PRYZANT, R. – DONG, V. Y. – GOLOBOKOV, K. – ZHU, C. – LIU, Y. FAST: Improving Controllability for Text Generation with Feedback Aware Self-Training. *CoRR.* 2022, abs/2210.03167. doi: 10.48550/arXiv.2210.03167. Available at: https://doi.org/10.48550/arXiv.2210.03167.

CHAKRABARTY, T. – MURESAN, S. – PENG, N. Generating similes effortlessly like a Pro: A Style Transfer Approach for Simile Generation. In WEBBER, B. – COHN, T. – HE, Y. – LIU, Y. (Ed.) *Proceedings of the 2020 Conference on Empirical Methods in NaturalLanguage Processing, EMNLP 2020, Online, November 16-20, 2020*, p. 6455–6469. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.524. Available at: https://doi.org/10.18653/v1/2020.emnlp-main.524.

Chawla, K. – Srinivasan, B. V. – Chhaya, N. Generating Formality-Tuned Summaries Using Input-Dependent Rewards. In Bansal, M. – Villavicencio, A. (Ed.) *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, p. 833–842, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1078. Available at: https://aclanthology.org/K19-1078/.

Chen, B. – Abedjan, Z. Duetcs: Code Style Transfer through Generation and Retrieval. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, p. 2362–2373, 2023. doi: 10.1109/ICSE48619.2023.00198.

Chen, D. – Dolan, W. B. Collecting highly parallel data for paraphrase evaluation. In *Proc. ACL-HLT*, p. 190–200, 2011.

Chen, H. – Ji, Y. Adversarial Training for Improving Model Robustness? Look at Both Prediction and Interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Chen, L. – Zaharia, M. – Zou, J. How is ChatGPT's behavior changing over time? *arXiv preprint arXiv:2307.09009*. 2023.

Chen, L. – Dai, S. – Tao, C. – Zhang, H. – Gan, Z. – Shen, D. – Zhang, Y. – Wang, G. – Zhang, R. – Carin, L. Adversarial Text Generation via Feature-Mover's Distance. In Larochelle, S. B. – Grauman, K. – Cesa-Bianchi, N. – Garnett, R. (Ed.) *Advances in Neural Information Processing Systems 31: Annual Conferenceon Neural Information Processing Systems 2018, NeurIPS 2018, December3-8, 2018, Montréal, Canada*, p. 4671–4682, 2018. Available at: https://proceedings.neurips.cc/paper/2018/hash/074177d3eb6371e32c16c55a3b8f706b-Abstract.html.

Cho, K. – Merrienboer, B. – Gülçehre, Ç. – Bahdanau, D. – Bougares, F. – Schwenk, H. – Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Moschitti, A. – Pang, B. – Daelemans, W. (Ed.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, p. 1724–1734. ACL, 2014. doi: 10.3115/v1/d14-1179. Available at: https://doi.org/10.3115/v1/d14-1179.

Chomsky, N. *Aspects of the Theory of Syntax*. The MIT Press, 50 edition, 1965. Available at: http://www.jstor.org/stable/j.ctt17kk81z. ISBN 9780262527408.

Chowdhery, A. – Narang, S. – Devlin, J. – Bosma, M. – Mishra, G. – Roberts, A. – Barham, P. – Chung, H. W. – Sutton, C. – Gehrmann, S. – others. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*. 2023, 24, 240, p. 1–113.

CLARK, E. – ROSS, A. S. – TAN, C. – JI, Y. – SMITH, N. A. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In BERKOVSKY, S. – HIJIKATA, Y. – REKIMOTO, J. – BURNETT, M. M. – BILLINGHURST, M. – QUIGLEY, A. (Ed.) *Proceedings of the 23rd International Conference on Intelligent UserInterfaces, IUI 2018, Tokyo, Japan, March 07-11, 2018*, p. 329–340. ACM, 2018. doi: 10.1145/3172944.3172983. Available at: `https://doi.org/10.1145/3172944.3172983`.

CONNEAU, A. – KHANDELWAL, K. – GOYAL, N. – CHAUDHARY, V. – WENZEK, G. – GUZMÁN, F. – GRAVE, E. – OTT, M. – ZETTLEMOYER, L. – STOYANOV, V. Unsupervised Cross-lingual Representation Learning at Scale. In JURAFSKY, D. – CHAI, J. – SCHLUTER, N. – TETREAULT, J. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. Available at: `https://aclanthology.org/2020.acl-main.747/`.

COPPOCK, L. Politeness strategies in conversation closings. *Unpublished manuscript: Stanford University*. 2005.

COSTA-JUSSÀ, M. R. et al. No Language Left Behind: Scaling Human-Centered Machine Translation. *CoRR*. 2022, abs/2207.04672. doi: 10.48550/arXiv.2207.04672. Available at: `https://doi.org/10.48550/arXiv.2207.04672`.

DAI, N. – LIANG, J. – QIU, X. – HUANG, X. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. In *Proc. ACL*, p. 5997–6007, 2019. doi: 10.18653/v1/P19-1601. Available at: `https://www.aclweb.org/anthology/P19-1601`.

DALE, D. – VORONOV, A. – DEMENTIEVA, D. – LOGACHEVA, V. – KOZLOVA, O. – SEMENOV, N. – PANCHENKO, A. Text Detoxification using Large Pre-trained Neural Models. In MOENS, M.-F. – HUANG, X. – SPECIA, L. – YIH, S. W.-t. (Ed.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 7979–7996, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.629. Available at: `https://aclanthology.org/2021.emnlp-main.629/`.

JONG, M. – THEUNE, M. – HOFS, D. Politeness and alignment in dialogues with a virtual guide. In PADGHAM, L. – PARKES, D. C. – MÜLLER, J. P. – PARSONS, S. (Ed.) *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 12-16, 2008, Volume 1*, p. 207–214. IFAAMAS, 2008. Available at: `https://dl.acm.org/citation.cfm?id=1402416`.

DEMENTIEVA, D. – USTYANTSEV, S. – DALE, D. – KOZLOVA, O. – SEMENOV, N. – PANCHENKO, A. – LOGACHEVA, V. Crowdsourcing of Parallel Corpora: the Case of Style Transfer for Detoxification. In USTALOV, D. – CASATI, F. – DRUTSA, A. – STELMAKH, I. – PAVLICHENKO, N. – BAIDAKOVA, D. (Ed.) *Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and*

Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021), Copenhagen, Denmark, August 20, 2021, 2932 / *CEUR Workshop Proceedings*, p. 35–49. CEUR-WS.org, 2021. Available at: https://ceur-ws.org/Vol-2932/paper2.pdf.

Dementieva, D. – Logacheva, V. – Nikishina, I. – Fenogenova, A. – Dale, D. – Krotova, I. – Semenov, N. – Shavrina, T. – Panchenko, A. RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora. In *COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES (2022).*, p. 114 – 131, 2022. doi: 10.28995/2075-7182-2022-21-114-131. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137765263&doi=10.28995%2f2075-7182-2022-21-114-131&partnerID=40&md5=c3e4753e930d98f3968f2829f36d01a8. Cited by: 1; All Open Access, Bronze Open Access.

Dementieva, D. – Moskovskiy, D. – Dale, D. – Panchenko, A. Exploring Methods for Cross-lingual Text Style Transfer: The Case of Text Detoxification. In Park, J. C. – Arase, Y. – Hu, B. – Lu, W. – Wijaya, D. – Purwarianti, A. – Krisnadhi, A. A. (Ed.) *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1083–1101, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.70. Available at: https://aclanthology.org/2023.ijcnlp-main.70/.

Devlin, J. – Chang, M.-W. – Lee, K. – Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J. – Doran, C. – Solorio, T. (Ed.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. Available at: https://aclanthology.org/N19-1423/.

Du, Z. – Qian, Y. – Liu, X. – Ding, M. – Qiu, J. – Yang, Z. – Tang, J. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Muresan, S. – Nakov, P. – Villavicencio, A. (Ed.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 320–335, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.26. Available at: https://aclanthology.org/2022.acl-long.26/.

Dubey, A. – Jauhri, A. – Pandey, A. – Kadian, A. – Al-Dahle, A. – Letman, A. – Mathur, A. – Schelten, A. – Yang, A. – Fan, A. – others. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. 2024.

Dziri, N. – Kamalloo, E. – Mathewson, K. – Zaiane, O. Augmenting Neural Response Generation with Context-Aware Topical Attention. In Chen, Y.-N. – Bedrax-Weiss, T. – Hakkani-Tur, D. – Kumar, A. – Lewis, M. – Luong, T.-M. – Su, P.-H. – Wen, T.-H. (Ed.) *Proceedings of the First Workshop on NLP for Conversational AI*, p. 18–31, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4103. Available at: https://aclanthology.org/W19-4103/.

Elman, J. L. Finding structure in time. *Cognitive science*. 1990, 14, 2, p. 179–211.

Fan, A. – Bhosale, S. – Schwenk, H. – Ma, Z. – El-Kishky, A. – Goyal, S. – Baines, M. – Celebi, O. – Wenzek, G. – Chaudhary, V. – others. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*. 2021, 22, 107, p. 1–48.

Feng, F. – Yang, Y. – Cer, D. – Arivazhagan, N. – Wang, W. Language-agnostic BERT Sentence Embedding. In Muresan, S. – Nakov, P. – Villavicencio, A. (Ed.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, p. 878–891. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.62. Available at: https://doi.org/10.18653/v1/2022.acl-long.62.

Feng, X. – Liu, M. – Liu, J. – Qin, B. – Sun, Y. – Liu, T. Topic-to-Essay Generation with Neural Networks. In Lang, J. (Ed.) *Proceedings of the Twenty-Seventh International Joint Conference onArtificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm,Sweden*, p. 4078–4084. ijcai.org, 2018. doi: 10.24963/ijcai.2018/567. Available at: https://doi.org/10.24963/ijcai.2018/567.

Firdaus, M. – Shandilya, A. – Ekbal, A. – Bhattacharyya, P. Being Polite: Modeling Politeness Variation in a Personalized Dialog Agent. *IEEE Transactions on Computational Social Systems*. 2022.

Franceschelli, G. – Musolesi, M. On the creativity of large language models. *AI & SOCIETY*. 2024, p. 1–11.

Fu, Y. – Zhou, H. – Chen, J. – Li, L. Rethinking Text Attribute Transfer: A Lexical Analysis. In *Proc. INLG*, Tokyo, Japan, October 2019.

Fu, Z. – Tan, X. – Peng, N. – Zhao, D. – Yan, R. Style Transfer in Text: Exploration and Evaluation. In McIlraith, S. A. – Weinberger, K. Q. (Ed.) *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, p. 663–670. AAAI Press, 2018. Available at: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17015.

Gan, C. – Gan, Z. – He, X. – Gao, J. – Deng, L. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 3137–3146, 2017.

Gao, M. – Hu, X. – Ruan, J. – Pu, X. – Wan, X. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*. 2024.

Gao, X. – Zhang, Y. – Lee, S. – Galley, M. – Brockett, C. – Gao, J. – Dolan, B. Structuring Latent Spaces for Stylized Response Generation. In Inui, K. – Jiang, J. – Ng, V. – Wan, X. (Ed.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 1814–1823, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1190. Available at: https://aclanthology.org/D19-1190/.

Gatt, A. – Krahmer, E. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.* 2018, 61, p. 65–170. doi: 10.1613/jair.5477. Available at: https://doi.org/10.1613/jair.5477.

Gatys, L. A. – Ecker, A. S. – Bethge, M. A Neural Algorithm of Artistic Style. *CoRR*. 2015, abs/1508.06576. Available at: http://arxiv.org/abs/1508.06576.

Goodfellow, I. – Bengio, Y. – Courville, A. *Deep learning*. MIT press, 2016. http://www.deeplearningbook.org.

Goswami, K. – Rani, P. – Fransen, T. – McCrae, J. Weakly-supervised Deep Cognate Detection Framework for Low-Resourced Languages Using Morphological Knowledge of Closely-Related Languages. In Bouamor, H. – Pino, J. – Bali, K. (Ed.) *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 531–541, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.38. Available at: https://aclanthology.org/2023.findings-emnlp.38/.

Goyal, T. – Rajani, N. – Liu, W. – Kryscinski, W. HydraSum: Disentangling Style Features in Text Summarization with Multi-Decoder Models. In Goldberg, Y. – Kozareva, Z. – Zhang, Y. (Ed.) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 464–479, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.30. Available at: https://aclanthology.org/2022.emnlp-main.30/.

Gröndahl, T. – Asokan, N. Effective writing style transfer via combinatorial paraphrasing. *Proceedings on Privacy Enhancing Technologies*. 2020, 2020, 4, p. 175–195. doi: 10.2478/popets-2020-0068. Available at: https://doi.org/10.2478/popets-2020-0068.

Grossberg, S. Recurrent neural networks. *Scholarpedia*. 2013, 8, 2, p. 1888.

Gupta, S. – Walker, M. A. – Romano, D. M. How Rude Are You?: Evaluating Politeness and Affect in Interaction. In Paiva, A. – Prada, R. – Picard, R. W. (Ed.) *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, 4738 / *Lecture Notes in Computer Science*, p. 203–217. Springer, 2007. doi: 10.1007/978-3-540-74889-2_19. Available at: https://doi.org/10.1007/978-3-540-74889-2_19.

Gururangan, S. – Marasovic, A. – Swayamdipta, S. – Lo, K. – Beltagy, I. – Downey, D. – Smith, N. A. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Jurafsky, D. – Chai, J. – Schluter, N. – Tetreault, J. R. (Ed.) *Proc. ACL*, p. 8342–8360. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.740. Available at: https://doi.org/10.18653/v1/2020.acl-main.740.

Hanke, V. – Blanchard, T. – Boenisch, F. – Olatunji, I. E. – Backes, M. – Dziedzic, A. Open LLMs are Necessary for Private Adaptations and Outperform their Closed Alternatives. In *ICML 2024 Next Generation of AI Safety Workshop*, 2024.

Hauser, M. D. – Chomsky, N. – Fitch, W. T. The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*. 2002, 298, 5598, p. 1569–1579. doi: 10.1126/science.298.5598.1569. Available at: https://www.science.org/doi/abs/10.1126/science.298.5598.1569.

He, H. – Peng, N. – Liang, P. Pun Generation with Surprise. In Burstein, J. – Doran, C. – Solorio, T. (Ed.) *Proceedings of the 2019 Conference of the North American Chapter ofthe Association for Computational Linguistics: Human Language Technologies,NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Longand Short Papers)*, p. 1734–1744. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1172. Available at: https://doi.org/10.18653/v1/n19-1172.

He, J. – Gu, J. – Shen, J. – Ranzato, M. Revisiting Self-Training for Neural Sequence Generation. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia, 2020a. OpenReview.net. Available at: https://openreview.net/forum?id=SJgdnAVKDH.

He, J. – Wang, X. – Neubig, G. – Berg-Kirkpatrick, T. A Probabilistic Formulation of Unsupervised Text Style Transfer. In *International Conference on Learning Representations (ICLR)*, 2020b.

He, K. – Zhang, X. – Ren, S. – Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770–778, 2016.

He, R. – McAuley, J. J. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. In Bourdeau, J. – Hendler, J. – Nkambou, R. – Horrocks, I. – Zhao, B. Y. (Ed.) *Proceedings of the 25th International Conference on World Wide Web,WWW 2016, Montreal, Canada, April 11 - 15, 2016*, p. 507–517. ACM, 2016. doi: 10.1145/2872427.2883037. Available at: https://doi.org/10.1145/2872427.2883037.

Hershcovich, D. – Frank, S. – Lent, H. – Lhoneux, M. – Abdou, M. – Brandl, S. – Bugliarello, E. – Piqueras, L. C. – Chalkidis, I. – Cui, R. – others. Challenges and strategies in cross-cultural NLP. *arXiv preprint arXiv:2203.10020*. 2022.

Hochreiter, S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. 1998, 6, 02, p. 107–116.

Hochreiter, S. – Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* 1997, 9, 8, p. 1735–1780. doi: 10.1162/neco.1997.9.8.1735. Available at: https://doi.org/10.1162/neco.1997.9.8.1735.

Hovy, E. Generating natural language under pragmatic constraints. *Journal of Pragmatics*. 1987, 11, 6, p. 689–719.

Hu, Z. – Lee, R. K. – Aggarwal, C. C. – Zhang, A. Text Style Transfer: A Review and Experimental Evaluation. *SIGKDD Explor.* 2022, 24, 1, p. 14–45. doi: 10.1145/3544903.3544906. Available at: https://doi.org/10.1145/3544903.3544906.

Hu, Z. – Lee, R. K.-W. – Chen, N. F. Adapter-TST: A Parameter Efficient Method for Multiple-Attribute Text Style Transfer. *arXiv preprint arXiv:2305.05945*. 2023.

Hu, Z. – Yang, Z. – Liang, X. – Salakhutdinov, R. – Xing, E. P. Toward Controlled Generation of Text. In Precup, D. – Teh, Y. W. (Ed.) *Proc. ICML*, 70 / *Proceedings of Machine Learning Research*, p. 1587–1596. PMLR, 2017. Available at: http://proceedings.mlr.press/v70/hu17e.html.

Huang, J. – Gu, S. S. – Hou, L. – Wu, Y. – Wang, X. – Yu, H. – Han, J. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*. 2022.

Hutto, C. J. – Gilbert, E. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. In Adar, E. – Resnick, P. – Choudhury, M. D. – Hogan, B. – Oh, A. H. (Ed.) *Proc. ICWSM*. The AAAI Press, 2014. Available at: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109.

Jain, N. – Chiang, P. – Wen, Y. – Kirchenbauer, J. – Chu, H.-M. – Somepalli, G. – Bartoldson, B. R. – Kailkhura, B. – Schwarzschild, A. – Saha, A. – Goldblum, M. – Geiping, J. – Goldstein, T. NEFTune: Noisy Embeddings Improve Instruction Finetuning, 2023.

Janizek, J. D. – Sturmfels, P. – Lee, S. Explaining Explanations: Axiomatic Feature Interactions for Deep Networks. *J. Mach. Learn. Res.* 2021, 22, p. 104:1–104:54. Available at: http://jmlr.org/papers/v22/20-1223.html.

Jha, G. N. The TDIL Program and the Indian Langauge Corpora Intitiative (ILCI). In Calzolari, N. – Choukri, K. – Maegaard, B. – Mariani, J. – Odijk, J. – Piperidis, S. – Rosner, M. – Tapias, D. (Ed.) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). Available at: https://aclanthology.org/L10-1602/.

Jhamtani, H. – Gangal, V. – Hovy, E. – Nyberg, E. Shakespearizing Modern Language Using Copy-Enriched Sequence to Sequence Models. In *Proceedings of the Workshop on Stylistic Variation*, p. 10–19, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4902. Available at: https://aclanthology.org/W17-4902.

Jiang, A. Q. et al. Mistral 7B, 2023.

Jiang, Z. – Xu, F. F. – Araki, J. – Neubig, G. How can we know what language models know? *Transactions of the Association for Computational Linguistics*. 2020, 8, p. 423–438.

Jin, D. – Jin, Z. – Zhou, J. T. – Orii, L. – Szolovits, P. Hooks in the Headline: Learning to Generate Headlines with Controlled Styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Jin, D. – Jin, Z. – Hu, Z. – Vechtomova, O. – Mihalcea, R. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*. March 2022, 48, 1, p. 155–205. doi: 10.1162/coli_a_00426. Available at: https://aclanthology.org/2022.cl-1.6/.

Jin, Z. – Jin, D. – Mueller, J. – Matthews, N. – Santus, E. IMaT: Unsupervised Text Attribute Transfer via Iterative Matching and Translation. In Inui, K. – Jiang, J. – Ng, V. – Wan, X. (Ed.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3097–3109, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1306. Available at: https://aclanthology.org/D19-1306/.

Jing, L. – Song, X. – Lin, X. – Zhao, Z. – Zhou, W. – Nie, L. Stylized Data-to-text Generation: A Case Study in the E-Commerce Domain. *ACM Trans. Inf. Syst.* aug 2023, 42, 1. ISSN 1046-8188. doi: 10.1145/3603374. Available at: https://doi.org/10.1145/3603374.

John, V. – Mou, L. – Bahuleyan, H. – Vechtomova, O. Disentangled Representation Learning for Non-Parallel Text Style Transfer. In Korhonen, A. – Traum, D. – Màrquez, L. (Ed.) *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 424–434, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1041. Available at: https://aclanthology.org/P19-1041/.

Joshi, P. – Santy, S. – Budhiraja, A. – Bali, K. – Choudhury, M. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Jurafsky, D. – Chai, J. – Schluter, N. – Tetreault, J. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. Available at: https://aclanthology.org/2020.acl-main.560/.

Kang, D. – Hovy, E. xSLUE: A Benchmark and Analysis Platform for Cross-Style Language Understanding and Evaluation, 2019.

Kaplan, J. – McCandlish, S. – Henighan, T. – Brown, T. B. – Chess, B. – Child, R. – Gray, S. – Radford, A. – Wu, J. – Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*. 2020.

Kaptein, M. – Markopoulos, P. – De Ruyter, B. – Aarts, E. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*. 2015, 77, p. 38–51.

Kenton, J. D. M.-W. C. – Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, 1, p. 2. Minneapolis, Minnesota, 2019.

Keskar, N. S. – McCann, B. – Varshney, L. R. – Xiong, C. – Socher, R. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*. 2019.

Kilickaya, M. – Erdem, A. – Ikizler-Cinbis, N. – Erdem, E. Re-evaluating Automatic Metrics for Image Captioning. In Lapata, M. – Blunsom, P. – Koller, A. (Ed.) *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, p. 199–209, Valencia, Spain, April 2017. Association for Computational Linguistics. Available at: https://aclanthology.org/E17-1019/.

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

Kocmi, T. – Federmann, C. Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In Nurminen, M. et al. (Ed.) *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023*, p. 193–203, Tampere, Finland, 2023. Available at: https://aclanthology.org/2023.eamt-1.19.

Kong, X. – Li, B. – Neubig, G. – Hovy, E. H. – Yang, Y. An Adversarial Approach to High-Quality, Sentiment-Controlled Neural Dialogue Generation. *CoRR*. 2019, abs/1901.07129. Available at: http://arxiv.org/abs/1901.07129.

KOO, R. – LEE, M. – RAHEJA, V. – PARK, J. I. – KIM, Z. M. – KANG, D. Benchmarking Cognitive Biases in Large Language Models as Evaluators. *CoRR*. 2023, abs/2309.17012. doi: 10.48550/ARXIV.2309.17012. Available at: https://doi.org/10.48550/arXiv.2309.17012.

KÖPF, A. et al. OpenAssistant Conversations - Democratizing Large Language Model Alignment. *CoRR*. 2023, abs/2304.07327. doi: 10.48550/arXiv.2304.07327. Available at: https://doi.org/10.48550/arXiv.2304.07327.

KOŠARKO, O. – VARIŠ, D. – POPEL, M. LINDAT Translation service, 2019. Available at: http://hdl.handle.net/11234/1-2922. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

KRISHNA, K. – WIETING, J. – IYYER, M. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*. 2020.

KRISHNA, K. – NATHANI, D. – GARCIA, X. – SAMANTA, B. – TALUKDAR, P. Few-shot Controllable Style Transfer for Low-Resource Multilingual Settings. In MURESAN, S. – NAKOV, P. – VILLAVICENCIO, A. (Ed.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, p. 7439–7468. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.514. Available at: https://doi.org/10.18653/v1/2022.acl-long.514.

KULLBACK, S. *Information theory and statistics.* Courier Corporation, 1997.

KUMAR, R. – LAHIRI, B. – ALOK, D. – OJHA, A. K. O. – JAIN, M. – BASIT, A. – DAWAR, Y. Automatic Identification of Closely-related Indian Languages: Resources and Experiments. In JHA, G. N. – BALI, K. – L, S. – OJHA, A. K. (Ed.) *Proceedings of the 4th Workshop on Indian Language Data: Resources and Evaluation under the LREC 2018*, Paris, France, may 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-09-2.

KUMAR, S. – AGGARWAL, S. – SHARMA, D. M. – MAMIDI, R. How do different factors Impact the Inter-language Similarity? A Case Study on Indian languages. In KABBARA, J. – LIN, H. – PAULLADA, A. – VAMVAS, J. (Ed.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, p. 112–118, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-srw.12. Available at: https://aclanthology.org/2021.acl-srw.12/.

KUSNER, M. J. – SUN, Y. – KOLKIN, N. I. – WEINBERGER, K. Q. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, p. 957–966. JMLR.org, 2015.

Lample, G. – Subramanian, S. – Smith, E. M. – Denoyer, L. – Ranzato, M. – Boureau, Y. Multiple-Attribute Text Rewriting. In *7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 6-9, 2019, 2019. OpenReview.net. Available at: https://openreview.net/forum?id=H1g2NhC5KQ.

Lee, J. Stable Style Transformer: Delete and Generate Approach with Encoder-Decoder for Text Style Transfer. In Davis, B. – Graham, Y. – Kelleher, J. D. – Sripada, Y. (Ed.) *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, p. 195–204, Dublin, Ireland, 2020. Association for Computational Linguistics. Available at: https://aclanthology.org/2020.inlg-1.25/.

Len, Y. – Portet, F. – Labbé, C. – Qader, R. Controllable Neural Natural Language Generation: comparison of state-of-the-art control strategies. In *WebNLG+: 3rd Workshop on Natural Language Generation from the Semantic Web*, 2020.

Lewis, M. – Liu, Y. – Goyal, N. – Ghazvininejad, M. – Mohamed, A. – Levy, O. – Stoyanov, V. – Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D. – Chai, J. – Schluter, N. – Tetreault, J. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. Available at: https://aclanthology.org/2020.acl-main.703/.

Li, D. – Zhang, Y. – Gan, Z. – Cheng, Y. – Brockett, C. – Dolan, B. – Sun, M. Domain Adaptive Text Style Transfer. In Inui, K. – Jiang, J. – Ng, V. – Wan, X. (Ed.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, p. 3302–3311. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1325. Available at: https://doi.org/10.18653/v1/D19-1325.

Li, J. – Feng, Y. – Ou, J. SE-DAE: Style-Enhanced Denoising Auto-Encoder for Unsupervised Text Style Transfer. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, p. 1–8. IEEE, 2021. doi: 10.1109/IJCNN52387.2021.9533731. Available at: https://doi.org/10.1109/IJCNN52387.2021.9533731.

Li, J. – Galley, M. – Brockett, C. – Spithourakis, G. – Gao, J. – Dolan, B. A Persona-Based Neural Conversation Model. In Erk, K. – Smith, N. A. (Ed.) *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1094. Available at: https://aclanthology.org/P16-1094/.

Li, J. – Jia, R. – He, H. – Liang, P. Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer. In Walker, M. – Ji, H. – Stent, A. (Ed.) *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, p. 1865–1874, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1169. Available at: https://aclanthology.org/N18-1169/.

Li, J. – Tang, T. – Nie, J. – Wen, J. – Zhao, X. Learning to Transfer Prompts for Text Generation. In Carpuat, M. – Marneffe, M. – Ruíz, I. V. M. (Ed.) *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, p. 3506–3518, Seattle, WA, United States, 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.257. Available at: https://doi.org/10.18653/v1/2022.naacl-main.257.

Li, X. L. – Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Zong, C. – Xia, F. – Li, W. – Navigli, R. (Ed.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers)*, p. 4582–4597, Virtual Event, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353. Available at: https://doi.org/10.18653/v1/2021.acl-long.353.

Li, X. – Long, X. – Xia, Y. – Li, S. Low Resource Style Transfer via Domain Adaptive Meta Learning. In Carpuat, M. – Marneffe, M. – Ruíz, I. V. M. (Ed.) *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, p. 3014–3026, s, WA, United States, 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.220. Available at: https://doi.org/10.18653/v1/2022.naacl-main.220.

Li, Y. – Su, H. – Shen, X. – Li, W. – Cao, Z. – Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Kondrak, G. – Watanabe, T. (Ed.) *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. Available at: https://aclanthology.org/I17-1099/.

Liao, Y. – Bing, L. – Li, P. – Shi, S. – Lam, W. – Zhang, T. QuaSE: Sequence Editing under Quantifiable Guidance. In Riloff, E. – Chiang, D. – Hockenmaier, J. – Tsujii, J. (Ed.) *Proceedings of the 2018 Conference on Empirical Methods in NaturalLanguage Processing, Brussels, Belgium, October 31 - November 4, 2018*, p. 3855–3864. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1420. Available at: https://doi.org/10.18653/v1/d18-1420.

Liaw, A. Classification and regression by randomForest. *R news*. 2002.

LIN, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. Available at: https://aclanthology.org/W04-1013/.

LIN, C.-Y. – HOVY, E. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, p. 150–157, 2003. Available at: https://aclanthology.org/N03-1020/.

LIN, C.-Y. – OCH, F. J. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, p. 605–612, Barcelona, Spain, July 2004. doi: 10.3115/1218955.1219032. Available at: https://aclanthology.org/P04-1077/.

LIN, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*. 1991, 37, 1, p. 145–151.

LIU, P. – YUAN, W. – FU, J. – JIANG, Z. – HAYASHI, H. – NEUBIG, G. Pre-train, Prompt, and Predict: A Systematic Survey of PromptingMethods in Natural Language Processing. *ACM Comput. Surv.* 2023a, 55, 9, p. 195:1–195:35. doi: 10.1145/3560815. Available at: https://doi.org/10.1145/3560815.

LIU, Q. – QIN, J. – YE, W. – MOU, H. – HE, Y. – WANG, K. Adaptive Prompt Routing for Arbitrary Text Style Transfer with Pre-trained Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*. Mar. 2024a, 38, 17, p. 18689–18697. doi: 10.1609/aaai.v38i17.29832. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/29832.

LIU, S.-Y. – WANG, C.-Y. – YIN, H. – MOLCHANOV, P. – WANG, Y.-C. F. – CHENG, K.-T. – CHEN, M.-H. DoRA: Weight-Decomposed Low-Rank Adaptation. *arXiv preprint arXiv:2402.09353*. 2024b.

LIU, Y. – ITER, D. – XU, Y. – WANG, S. – XU, R. – ZHU, C. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In BOUAMOR, H. – PINO, J. – BALI, K. (Ed.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 2511–2522, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. Available at: https://aclanthology.org/2023.emnlp-main.153/.

LIU, Y. – GU, J. – GOYAL, N. – LI, X. – EDUNOV, S. – GHAZVININEJAD, M. – LEWIS, M. – ZETTLEMOYER, L. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*. 2020, 8, p. 726–742. doi: 10.1162/tacl_a_00343. Available at: https://aclanthology.org/2020.tacl-1.47/.

Loakman, T. – Tang, C. – Lin, C. TwistList: Resources and Baselines for Tongue Twister Generation. In Rogers, A. – Boyd-Graber, J. – Okazaki, N. (Ed.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 579–589, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.51. Available at: https://aclanthology.org/2023.acl-short.51/.

Locatello, F. – Bauer, S. – Lucic, M. – Rätsch, G. – Gelly, S. – Schölkopf, B. – Bachem, O. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In Chaudhuri, K. – Salakhutdinov, R. (Ed.) *Proc. ICML*, 97 / *Proceedings of Machine Learning Research*, p. 4114–4124. PMLR, 2019. Available at: http://proceedings.mlr.press/v97/locatello19a.html.

Logacheva, V. – Dementieva, D. – Ustyantsev, S. – Moskovskiy, D. – Dale, D. – Krotova, I. – Semenov, N. – Panchenko, A. ParaDetox: Detoxification with Parallel Data. In Muresan, S. – Nakov, P. – Villavicencio, A. (Ed.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 6804–6818, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.469. Available at: https://aclanthology.org/2022.acl-long.469/.

Loshchilov, I. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. 2017.

Luan, Y. – Brockett, C. – Dolan, B. – Gao, J. – Galley, M. Multi-Task Learning for Speaker-Role Adaptation in Neural Conversation Models. In Kondrak, G. – Watanabe, T. (Ed.) *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 605–614, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. Available at: https://aclanthology.org/I17-1061/.

Luo, F. – Li, P. – Yang, P. – Zhou, J. – Tan, Y. – Chang, B. – Sui, Z. – Sun, X. Towards Fine-grained Text Sentiment Transfer. In Korhonen, A. – Traum, D. R. – Màrquez, L. (Ed.) *Proc. ACL*, p. 2013–2022. Association for Computational Linguistics, 2019a. doi: 10.18653/v1/p19-1194. Available at: https://doi.org/10.18653/v1/p19-1194.

Luo, F. – Li, P. – Zhou, J. – Yang, P. – Chang, B. – Sun, X. – Sui, Z. A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer. In Kraus, S. (Ed.) *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, p. 5116–5122, Macao, 2019b. ijcai.org. doi: 10.24963/ijcai.2019/711. Available at: https://doi.org/10.24963/ijcai.2019/711.

Luo, G. – Han, Y. – Mou, L. – Firdaus, M. Prompt-Based Editing for Text Style Transfer. In Bouamor, H. – Pino, J. – Bali, K. (Ed.) *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 5740–5750, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.381. Available at: https://aclanthology.org/2023.findings-emnlp.381/.

Ma, E. NLP Augmentation. https://github.com/makcedward/nlpaug, 2019.

Madaan, A. – Setlur, A. – Parekh, T. – Póczos, B. – Neubig, G. – Yang, Y. – Salakhutdinov, R. – Black, A. W. – Prabhumoye, S. Politeness Transfer: A Tag and Generate Approach. In Jurafsky, D. – Chai, J. – Schluter, N. – Tetreault, J. R. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, p. 1869–1881. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.169. Available at: https://doi.org/10.18653/v1/2020.acl-main.169.

Majewska, O. – Vulić, I. – Korhonen, A. Linguistically Guided Multilingual NLP: Current Approaches, Challenges, and Future Perspectives. *Algebraic Structures in Natural Language*. 2022, p. 163–188.

Mathews, A. – Xie, L. – He, X. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI conference on artificial intelligence*, 30, 2016.

McCulloch, W. S. – Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*. 1943, 5, p. 115–133.

McDonald, D. D. – Pustejovsky, J. A Computational Theory of Prose Style for Natural Language Generation. In King, M. (Ed.) *EACL 1985, 2nd Conference of the European Chapter of the Association for Computational Linguistics, March 27-29, 1985, University of Geneva, Geneva, Switzerland*, p. 187–193. The Association for Computer Linguistics, 1985. Available at: https://aclanthology.org/E85-1027/.

Mehrotra, K. – Mohan, C. K. – Ranka, S. *Elements of artificial neural networks*. MIT press, 1997.

Mikolov, T. – Karafiát, M. – Burget, L. – Cernockỳ, J. – Khudanpur, S. Recurrent neural network based language model. In *Interspeech*, 2, p. 1045–1048. Makuhari, 2010.

Mikolov, T. – Chen, K. – Corrado, G. – Dean, J. Efficient Estimation of Word Representations in Vector Space. In Bengio, Y. – LeCun, Y. (Ed.) *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013. Available at: http://arxiv.org/abs/1301.3781.

Mir, R. – Felbo, B. – Obradovich, N. – Rahwan, I. Evaluating Style Transfer for Text. In Burstein, J. – Doran, C. – Solorio, T. (Ed.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 495–504, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1049. Available at: https://aclanthology.org/N19-1049/.

Mishra, S. – Khashabi, D. – Baral, C. – Choi, Y. – Hajishirzi, H. Reframing Instructional Prompts to GPTk's Language. *arXiv preprint arXiv:2109.07830*. 2021.

Moschitti, A. – Pang, B. – Daelemans, W. (Ed.). *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014. ACL.

Mou, L. – Vechtomova, O. Stylized text generation: Approaches and applications. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, p. 19–22, 2020.

Muennighoff, N. et al. Crosslingual Generalization through Multitask Finetuning. In Rogers, A. – Boyd-Graber, J. – Okazaki, N. (Ed.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. Available at: https://aclanthology.org/2023.acl-long.891/.

Mukherjee, S. – Dusek, O. Leveraging Low-Resource Parallel Data for Text Style Transfer. In *Proceedings of the 16th International Natural Language Generation Conference*, p. 388–395, Prague, Czechia, 2023.

Mukherjee, S. – Dušek, O. Text Style Transfer: An Introductory Overview, 2024. Available at: https://arxiv.org/abs/2407.14822.

Mukherjee, S. – Kasner, Z. – Dušek, O. Balancing the Style-Content Trade-Off in Sentiment Transfer Using Polarity-Aware Denoising. In Sojka, P. – Horák, A. – Kopeček, I. – Pala, K. (Ed.) *Text, Speech, and Dialogue*, p. 172–186, Cham, 2022. Springer International Publishing. doi: https://doi.org/10.1007/978-3-031-16270-1_15. Available at: https://doi.org/10.1007/978-3-031-16270-1_15. ISBN 978-3-031-16270-1.

Mukherjee, S. – Bansal, A. – Kr. Ojha, A. – P. McCrae, J. – Dusek, O. Text Detoxification as Style Transfer in English and Hindi. In D. Pawar, J. – Lalitha Devi, S. (Ed.) *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, p. 133–144, Goa University, Goa, India, December 2023a. NLP Association of India (NLPAI). Available at: https://aclanthology.org/2023.icon-1.13/.

Mukherjee, S. – Bansal, A. – Majumdar, P. – Ojha, A. K. – Dušek, O. Low-Resource Text Style Transfer for Bangla: Data & Models. In Alam, F. – Kar, S. – Chowdhury, S. A. – Sadeque, F. – Amin, R. (Ed.) *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, p. 34–47, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.banglalp-1.5. Available at: https://aclanthology.org/2023.banglalp-1.5/.

Mukherjee, S. – Hudeček, V. – Dušek, O. Polite Chatbot: A Text Style Transfer Application. In Bassignana, E. – Lindemann, M. – Petit, A. (Ed.) *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, p. 87–93, Dubrovnik, Croatia, May 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-srw.9. Available at: https://aclanthology.org/2023.eacl-srw.9/.

Mukherjee, S. – Lango, M. – Kasner, Z. – Dušek, O. A Survey of Text Style Transfer: Applications and Ethical Implications, 2024a. Available at: https://arxiv.org/abs/2407.16737.

Mukherjee, S. – Ojha, A. K. – Bansal, A. – Alok, D. – McCrae, J. P. – Dusek, O. Multilingual Text Style Transfer: Datasets & Models for Indian Languages. In Mahamood, S. – Minh, N. L. – Ippolito, D. (Ed.) *Proceedings of the 17th International Natural Language Generation Conference*, p. 494–522, Tokyo, Japan, September 2024b. Association for Computational Linguistics. Available at: https://aclanthology.org/2024.inlg-main.41/.

Mukherjee, S. – Ojha, A. K. – Dusek, O. Are Large Language Models Actually Good at Text Style Transfer? In Mahamood, S. – Minh, N. L. – Ippolito, D. (Ed.) *Proceedings of the 17th International Natural Language Generation Conference*, p. 523–539, Tokyo, Japan, September 2024c. Association for Computational Linguistics. Available at: https://aclanthology.org/2024.inlg-main.42/.

Mukherjee, S. – Ojha, A. K. – McCrae, J. P. – Dusek, O. Evaluating Text Style Transfer Evaluation: Are There Any Reliable Metrics?, 2025. Available at: https://arxiv.org/abs/2502.04718.

Müller, R. – Kornblith, S. – Hinton, G. E. When Does Label Smoothing Help? In Wallach, H. M. – Larochelle, H. – Beygelzimer, A. – d'Alché-Buc, F. – Fox, E. B. – Garnett, R. (Ed.) *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, p. 4696–4705, Vancouver, BC, Canada, 2019. Available at: https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.

Munson, K. – Savla, A. – Ting, C.-K. – Wade, S. – Kate, K. – Srinivas, K. Exploring Code Style Transfer with Neural Networks, 2022. Available at: https://arxiv.org/abs/2209.06273.

Nair, V. – Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, p. 807–814, 2010.

Narayan, S. The generalized sigmoid activation function: Competitive supervised learning. *Information sciences*. 1997, 99, 1-2, p. 69–82.

Neidert, J. – Schuster, S. – Green, S. – Heafield, K. – Manning, C. D. Stanford University's Submissions to the WMT 2014 Translation Task. In *Proc. WMT*, p. 150–156. The Association for Computer Linguistics, 2014. doi: 10.3115/v1/w14-3316. Available at: https://doi.org/10.3115/v1/w14-3316.

Ni, J. – Li, J. – McAuley, J. J. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In Inui, K. – Jiang, J. – Ng, V. – Wan, X. (Ed.) *Proc. EMNLP-IJCNLP*, p. 188–197. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1018. Available at: https://doi.org/10.18653/v1/D19-1018.

Nikolov, N. I. – Hahnloser, R. H. Large-Scale Hierarchical Alignment for Data-driven Text Rewriting. *arXiv preprint arXiv:1810.08237*. 2018.

Niu, T. – Bansal, M. Polite Dialogue Generation Without Parallel Data. *Transactions of the Association for Computational Linguistics*. 2018, 6, p. 373–389. doi: 10.1162/tacl_a_00027. Available at: https://aclanthology.org/Q18-1027/.

Niu, X. – Martindale, M. J. – Carpuat, M. A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output. In Palmer, M. – Hwa, R. – Riedel, S. (Ed.) *Proc. EMNLP*, p. 2814–2819. Association for Computational Linguistics, 2017. doi: 10.18653/v1/d17-1299. Available at: https://doi.org/10.18653/v1/d17-1299.

Nogueira dos Santos, C. – Melnyk, I. – Padhi, I. Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. In Gurevych, I. – Miyao, Y. (Ed.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, p. 189–194, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2031. Available at: https://aclanthology.org/P18-2031/.

Novikova, J. – Dušek, O. – Cercas Curry, A. – Rieser, V. Why We Need New Evaluation Metrics for NLG. In Palmer, M. – Hwa, R. – Riedel, S. (Ed.) *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2241–2252, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1238. Available at: https://aclanthology.org/D17-1238/.

Ojha, A. K. – Behera, P. – Singh, S. – Jha, G. N. Training & Evaluation of POS Taggers in Indo-Aryan Languages: A Case of Hindi, Odia and Bhojpuri. In *Proceedings of the 7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, p. 524–529, 2015. Available at: http://ltc.amu.edu.pl/a2015/book/papers/TANO2-2.pdf.

OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2023. Accessed on January 9, 2024.

Opitz, J. – Frank, A. SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features. In He, Y. – Ji, H. – Li, S. – Liu, Y. – Chang, C.-H. (Ed.) *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, p. 625–638, Online only, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.aacl-main.48. Available at: https://aclanthology.org/2022.aacl-main.48/.

Ostheimer, P. – Nagda, M. K. – Kloft, M. – Fellenz, S. A Call for Standardization and Validation of Text Style Transfer Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, p. 10791–10815, 2023.

Ostheimer, P. S. – Nagda, M. K. – Kloft, M. – Fellenz, S. Text Style Transfer Evaluation Using Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 15802–15822, 2024.

Ouyang, L. – Wu, J. – Jiang, X. – Almeida, D. – Wainwright, C. – Mishkin, P. – Zhang, C. – Agarwal, S. – Slama, K. – Ray, A. – others. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*. 2022a, 35, p. 27730–27744.

Ouyang, L. et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022b. Available at: http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Palash, M. H. – Das, P. P. – Haque, S. Sentimental Style Transfer in Text with Multigenerative Variational Auto-Encoder. In *2019 International Conference on Bangla Speech and Language Processing (ICBSLP)*, p. 1–4, September 2019. doi: 10.1109/ICBSLP47725.2019.201508. Available at: https://ieeexplore.ieee.org/document/9084042.

Pan, X. – Zhang, M. – Sheng, B. – Zhu, J. – Yang, M. Hidden trigger backdoor attack on {NLP} models via linguistic style manipulation. In *31st USENIX Security Symposium (USENIX Security 22)*, p. 3611–3628, 2022.

Pang, R. Y. Towards Actual (Not Operational) Textual Style Transfer Auto-Evaluation. In Xu, W. – Ritter, A. – Baldwin, T. – Rahimi, A. (Ed.) *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, p. 444–445, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-5557. Available at: https://aclanthology.org/D19-5557/.

Pang, R. Y. The Daunting Task of Real-World Textual Style Transfer Auto-Evaluation. *CoRR*. 2019b, abs/1910.03747.

Pang, R. Y. – Gimpel, K. Unsupervised Evaluation Metrics and Learning Criteria for Non-Parallel Textual Transfer. In Birch, A. – Finch, A. M. – Konstas, H. H. – Luong, T. – Neubig, G. – Oda, Y. – Sudoh, K. (Ed.) *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP2019, Hong Kong, November 4, 2019*, p. 138–147. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-5614. Available at: https://doi.org/10.18653/v1/D19-5614.

Papineni, K. – Roukos, S. – Ward, T. – Zhu, W.-J. Bleu: a Method for Automatic Evaluation of Machine Translation. In Isabelle, P. – Charniak, E. – Lin, D. (Ed.) *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. Available at: https://aclanthology.org/P02-1040/.

Penedo, G. – Malartic, Q. – Hesslow, D. – Cojocaru, R. – Cappelli, A. – Alobeidli, H. – Pannier, B. – Almazrouei, E. – Launay, J. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*. 2023. Available at: https://arxiv.org/abs/2306.01116.

Pennington, J. – Socher, R. – Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543, 2014.

Pérez-Rosas, V. – Resnicow, K. – Mihalcea, R. – others. VERVE: Template-based ReflectiVE Rewriting for MotiVational IntErviewing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, p. 10289–10302, 2023.

Prabhumoye, S. – Tsvetkov, Y. – Salakhutdinov, R. – Black, A. W. Style Transfer Through Back-Translation. In Gurevych, I. – Miyao, Y. (Ed.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 866–876, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1080. Available at: https://aclanthology.org/P18-1080/.

Prabhumoye, S. – Black, A. W. – Salakhutdinov, R. Exploring controllable text generation techniques. *arXiv preprint arXiv:2005.01822*. 2020.

Pryzant, R. – Martinez, R. D. – Dass, N. – Kurohashi, S. – Jurafsky, D. – Yang, D. Automatically Neutralizing Subjective Bias in Text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, p. 480–489. AAAI Press, 2020. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/5385.

Puka, L. *Kendall's Tau*, p. 713–715. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. doi: 10.1007/978-3-642-04898-2_324. Available at: https://doi.org/10.1007/978-3-642-04898-2_324. ISBN 978-3-642-04898-2.

Qi, F. – Chen, Y. – Zhang, X. – Li, M. – Liu, Z. – Sun, M. Mind the Style of Text! Adversarial and Backdoor Attacks Based on Text Style Transfer. In Moens, M.-F. – Huang, X. – Specia, L. – Yih, S. W.-t. (Ed.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 4569–4580, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.374. Available at: https://aclanthology.org/2021.emnlp-main.374/.

Qian, Y. Story-level Text Style Transfer: A Proposal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, p. 8–12, 2020.

Qiu, S. – Xu, B. – Zhang, J. – Wang, Y. – Shen, X. – Melo, G. – Long, C. – Li, X. EasyAug: An Automatic Textual Data Augmentation Platform for Classification Tasks. In Seghrouchni, A. E. F. – Sukthankar, G. – Liu, T. – Steen, M. (Ed.) *Companion of The 2020 Web Conference 2020*, p. 249–252, Taipei, Taiwan, 2020. ACM / IW3C2. doi: 10.1145/3366424.3383552. Available at: https://doi.org/10.1145/3366424.3383552.

Rabinovich, E. – Patel, R. N. – Mirkin, S. – Specia, L. – Wintner, S. Personalized Machine Translation: Preserving Original Author Traits. In Lapata, M. – Blunsom, P. – Koller, A. (Ed.) *Proc. EACL*, p. 1074–1084. Association for Computational Linguistics, 2017. doi: 10.18653/v1/e17-1101. Available at: https://doi.org/10.18653/v1/e17-1101.

Radford, A. – Józefowicz, R. – Sutskever, I. Learning to Generate Reviews and Discovering Sentiment. *CoRR*. 2017, abs/1704.01444. Available at: http://arxiv.org/abs/1704.01444.

Radford, A. – Narasimhan, K. – Salimans, T. – Sutskever, I. Improving language understanding by generative pre-training, 2018.

Radford, A. – Wu, J. – Child, R. – Luan, D. – Amodei, D. – Sutskever, I. Language Models are Unsupervised Multitask Learners. Tech. report, Open AI, OpenAI, 2019.

Rafailov, R. – Sharma, A. – Mitchell, E. – Manning, C. D. – Ermon, S. – Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*. 2023, 36, p. 53728–53741.

Raffel, C. – Shazeer, N. – Roberts, A. – Lee, K. – Narang, S. – Matena, M. – Zhou, Y. – Li, W. – Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 2020, 21, p. 140:1–140:67. Available at: http://jmlr.org/papers/v21/20-074.html.

Rahutomo, F. – Kitasuka, T. – Aritsugi, M. Semantic cosine similarity. In *The 7th international student conference on advanced science and technology ICAST*, 4, p. 1, 2012.

Rao, S. – Tetreault, J. R. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In Walker, M. A. – Ji, H. – Stent, A. (Ed.) *Proc. NAACL-HLT*, p. 129–140. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1012. Available at: https://doi.org/10.18653/v1/n18-1012.

Reddy, S. – Knight, K. Obfuscating Gender in Social Media Writing. In Bamman, D. – Doğruöz, A. S. – Eisenstein, J. – Hovy, D. – Jurgens, D. – O'Connor, B. – Oh, A. – Tsur, O. – Volkova, S. (Ed.) *Proceedings of the First Workshop on NLP and Computational Social Science*, p. 17–26, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5603. Available at: https://aclanthology.org/W16-5603/.

Reif, E. – Ippolito, D. – Yuan, A. – Coenen, A. – Callison-Burch, C. – Wei, J. A recipe for arbitrary text style transfer with large language models. *arXiv preprint arXiv:2109.03910*. 2021.

Reimers, N. – Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K. – Jiang, J. – Ng, V. – Wan, X. (Ed.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. Available at: https://aclanthology.org/D19-1410/.

Roller, S. – Dinan, E. – Goyal, N. – Ju, D. – Williamson, M. – Liu, Y. – Xu, J. – Ott, M. – Smith, E. M. – Boureau, Y.-L. – Weston, J. Recipes for Building an Open-Domain Chatbot. In Merlo, P. – Tiedemann, J. – Tsarfaty, R. (Ed.) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, p. 300–325, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.24. Available at: https://aclanthology.org/2021.eacl-main.24/.

Rubner, Y. – Tomasi, C. – Guibas, L. J. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*. 2000, 40, p. 99–121.

Rudra, K. – Rijhwani, S. – Begum, R. – Bali, K. – Choudhury, M. – Ganguly, N. Understanding Language Preference for Expression of Opinion and Sentiment: What do Hindi-English Speakers do on Twitter? In Su, J. – Duh, K. – Carreras, X. (Ed.) *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, p. 1131–1141, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1121. Available at: https://aclanthology.org/D16-1121/.

Saggion, H. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*. 2017, 10, 1, p. 1–137.

San, N. – Paraskevopoulos, G. – Arora, A. – He, X. – Kaur, P. – Adams, O. – Jurafsky, D. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. In Hahn, M. – Sorokin, A. – Kumar, R. – Shcherbakov, A. – Otmakhova, Y. – Yang, J. – Serikov, O. – Rani, P. – Ponti, E. M. – Muradoğlu, S. – Gao, R. – Cotterell, R. – Vylomova, E. (Ed.) *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, p. 100–112, St. Julian's, Malta, March 2024. Association for Computational Linguistics. Available at: https://aclanthology.org/2024.sigtyp-1.13/.

Sanh, V. – Debut, L. – Chaumond, J. – Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper And Lighter. *CoRR*. 2019, abs/1910.01108. Available at: http://arxiv.org/abs/1910.01108.

Sanh, V. et al. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*, 2022. Available at: https://openreview.net/forum?id=9Vrb9D0WI4.

Sap, M. – Prasettio, M. C. – Holtzman, A. – Rashkin, H. – Choi, Y. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, p. 2329–2334, 2017.

Schober, P. – Boer, C. – Schwarte, L. A. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*. 2018, 126, p. 1763–1768.

Sellam, T. – Das, D. – Parikh, A. BLEURT: Learning Robust Metrics for Text Generation. In Jurafsky, D. – Chai, J. – Schluter, N. – Tetreault, J. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. Available at: https://aclanthology.org/2020.acl-main.704/.

Sennrich, R. – Zhang, B. Revisiting Low-Resource Neural Machine Translation: A Case Study. In Korhonen, A. – Traum, D. R. – Màrquez, L. (Ed.) *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, p. 211–221, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/p19-1021. Available at: https://doi.org/10.18653/v1/p19-1021.

Sennrich, R. – Haddow, B. – Birch, A. Neural Machine Translation of Rare Words with Subword Units. In Erk, K. – Smith, N. A. (Ed.) *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 1715–1725, Berlin, Germany, August 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. Available at: https://aclanthology.org/P16-1162/.

Sennrich, R. – Haddow, B. – Birch, A. Controlling Politeness in Neural Machine Translation via Side Constraints. In Knight, K. – Nenkova, A. – Rambow, O. (Ed.) *Proc. NAACL-HLT*, p. 35–40. The Association for Computational Linguistics, 2016b. doi: 10.18653/v1/n16-1005. Available at: https://doi.org/10.18653/v1/n16-1005.

Sennrich, R. – Haddow, B. – Birch, A. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume 1: Long Papers*, Berlin, Germany, 2016c. The Association for Computer Linguistics. doi: 10.18653/v1/p16-1009. Available at: https://doi.org/10.18653/v1/p16-1009.

Shang, L. – Lu, Z. – Li, H. Neural Responding Machine for Short-Text Conversation. *CoRR*. 2015, abs/1503.02364. Available at: http://arxiv.org/abs/1503.02364.

Shang, M. – Li, P. – Fu, Z. – Bing, L. – Zhao, D. – Shi, S. – Yan, R. Semi-supervised Text Style Transfer: Cross Projection in Latent Space. In Inui, K. – Jiang, J. – Ng, V. – Wan, X. (Ed.) *Proceedings of the 2019 Conference on Empirical Methods in NaturalLanguage Processing and the 9th International Joint Conference onNatural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China,November 3-7, 2019*, p. 4936–4945. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1499. Available at: https://doi.org/10.18653/v1/D19-1499.

Sharma, A. – Lin, I. W. – Miner, A. S. – Atkins, D. C. – Althoff, T. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, p. 194–205, 2021.

Sharma, A. – Lin, I. W. – Miner, A. S. – Atkins, D. C. – Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*. 2023, 5, 1, p. 46–57.

Sheikha, F. A. – Inkpen, D. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*, p. 187–193, 2011.

Shen, D. – Zheng, M. – Shen, Y. – Qu, Y. – Chen, W. A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation. *CoRR*. 2020a, abs/2009.13818. Available at: https://arxiv.org/abs/2009.13818.

Shen, T. – Lei, T. – Barzilay, R. – Jaakkola, T. S. Style Transfer from Non-Parallel Text by Cross-Alignment. In Guyon, I. – Luxburg, U. – Bengio, S. – Wallach, H. M. – Fergus, R. – Vishwanathan, S. V. N. – Garnett, R. (Ed.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, p. 6830–6841, 2017. Available at: https://proceedings.neurips.cc/paper/2017/hash/2d2c8394e31101a261abf1784302bf75-Abstract.html.

SHEN, T. – MUELLER, J. – BARZILAY, R. – JAAKKOLA, T. S. Educating Text Autoencoders: Latent Representation Guidance via Denoising. In *Proceedings of the 37th International Conference on Machine Learning,ICML 2020, 13-18 July 2020, Virtual Event*, 119 / *Proceedings of Machine Learning Research*, p. 8719–8729. PMLR, 2020b. Available at: http://proceedings.mlr.press/v119/shen20c.html.

SHETTY, J. – ADIBI, J. The Enron email dataset database schema and brief statistical report. *Information sciences institute technical report, University of Southern California*. 2004, 4, 1, p. 120–128.

SHLIAZHKO, O. – FENOGENOVA, A. – TIKHONOVA, M. – KOZLOVA, A. – MIKHAILOV, V. – SHAVRINA, T. mGPT: Few-Shot Learners Go Multilingual. *Transactions of the Association for Computational Linguistics*. 2024, 12, p. 58–79. doi: 10.1162/tacl_a_00633. Available at: https://aclanthology.org/2024.tacl-1.4/.

SNOVER, M. – DORR, B. – SCHWARTZ, R. – MICCIULLA, L. – MAKHOUL, J. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, p. 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas. Available at: https://aclanthology.org/2006.amta-papers.25/.

SOCHER, R. – PERELYGIN, A. – WU, J. – CHUANG, J. – MANNING, C. D. – NG, A. Y. – POTTS, C. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, p. 1631–1642, Seattle, Washington, USA, 2013. ACL. Available at: https://aclanthology.org/D13-1170/.

SONG, H. – WANG, Y. – ZHANG, W.-N. – LIU, X. – LIU, T. Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation. In JURAFSKY, D. – CHAI, J. – SCHLUTER, N. – TETREAULT, J. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 5821–5831, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.516. Available at: https://aclanthology.org/2020.acl-main.516/.

SONG, Z. – ZHENG, X. – LIU, L. – XU, M. – HUANG, X. Generating Responses with a Specific Emotion in Dialog. In KORHONEN, A. – TRAUM, D. R. – MÀRQUEZ, L. (Ed.) *Proceedings of the 57th Conference of the Association for ComputationalLinguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019,Volume 1: Long Papers*, p. 3685–3695. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1359. Available at: https://doi.org/10.18653/v1/p19-1359.

Straka, M. – Straková, J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In Hajič, J. – Zeman, D. (Ed.) *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, p. 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-3009. Available at: https://aclanthology.org/K17-3009/.

Stureborg, R. – Alikaniotis, D. – Suhara, Y. Large Language Models Are Inconsistent and Biased Evaluators, 2024. Available at: http://arxiv.org/abs/2405.01724.

Su, F.-G. – Hsu, A. R. – Tuan, Y.-L. – Lee, H.-Y. Personalized Dialogue Response Generation Learned from Monologues. In *Interspeech 2019*, p. 4160–4164, 2019. doi: 10.21437/Interspeech.2019-1696.

Subramanian, S. – Lample, G. – Smith, E. M. – Denoyer, L. – Ranzato, M. – Boureau, Y. Multiple-Attribute Text Style Transfer. *CoRR*. 2018, abs/1811.00552. Available at: http://arxiv.org/abs/1811.00552.

Sudhakar, A. – Upadhyay, B. – Maheswaran, A. "Transforming" Delete, Retrieve, Generate Approach for Controlled Text Style Transfer. In Inui, K. – Jiang, J. – Ng, V. – Wan, X. (Ed.) *Proc. EMNLP-IJCNLP*, p. 3267–3277. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1322. Available at: https://doi.org/10.18653/v1/D19-1322.

Sun, Q. – Xu, C. – Hu, H. – Wang, Y. – Miao, J. – Geng, X. – Chen, Y. – Xu, F. – Jiang, D. Stylized Knowledge-Grounded Dialogue Generation via Disentangled Template Rewriting. In Carpuat, M. – Marneffe, M.-C. – Meza Ruiz, I. V. (Ed.) *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 3304–3318, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.241. Available at: https://aclanthology.org/2022.naacl-main.241/.

Sun, X. – Li, J. – Wei, X. – Li, C. – Tao, J. Emotional editing constraint conversation content generation based on reinforcement learning. *Inf. Fusion*. 2020, 56, p. 70–80. doi: 10.1016/j.inffus.2019.10.007. Available at: https://doi.org/10.1016/j.inffus.2019.10.007.

Sundararajan, M. – Taly, A. – Yan, Q. Axiomatic Attribution for Deep Networks. In Precup, D. – Teh, Y. W. (Ed.) *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 70 / *Proceedings of Machine Learning Research*, p. 3319–3328. PMLR, 2017. Available at: http://proceedings.mlr.press/v70/sundararajan17a.html.

Suzgun, M. – Melas-Kyriazi, L. – Jurafsky, D. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. *arXiv preprint arXiv:2205.11503*. 2022.

Tan, Y. – Lin, Z. – Fu, P. – Zheng, M. – Wang, L. – Cao, Y. – Wang, W. Detach and Attach: Stylized Image Captioning without Paired Stylized Dataset. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, p. 4733–4741, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3503161.3548295. Available at: https://doi.org/10.1145/3503161.3548295. ISBN 9781450392037.

Tang, Y. – Tran, C. – Li, X. – Chen, P. – Goyal, N. – Chaudhary, V. – Gu, J. – Fan, A. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *CoRR*. 2020, abs/2008.00401. Available at: https://arxiv.org/abs/2008.00401.

Tikhonov, A. – Yamshchikov, I. P. Guess who? Multilingual Approach For The Automated Generation Of Author-Stylized Poetry. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, p. 787–794, 2018. doi: 10.1109/SLT.2018.8639573.

Tikhonov, A. – Shibaev, V. – Nagaev, A. – Nugmanova, A. – Yamshchikov, I. P. Style Transfer for Texts: Retrain, Report Errors, Compare with Rewrites. In Inui, K. – Jiang, J. – Ng, V. – Wan, X. (Ed.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, p. 3934–3943. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1406. Available at: https://doi.org/10.18653/v1/D19-1406.

Ting, C.-K. – Munson, K. – Wade, S. – Savla, A. – Kate, K. – Srinivas, K. CodeStylist: A System for Performing Code Style Transfer Using Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*. Sep. 2023, 37, 13, p. 16485–16487. doi: 10.1609/aaai.v37i13.27087. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/27087.

Toshevska, M. – Gievska, S. A Review of Text Style Transfer using Deep Learning. *IEEE Transactions on Artificial Intelligence*. 2021. ISSN 2691-4581. doi: 10.1109/TAI.2021.3115992.

Touvron, H. – Lavril, T. – Izacard, G. – Martinet, X. – Lachaux, M. – Lacroix, T. – Rozière, B. – Goyal, N. – Hambro, E. – Azhar, F. – Rodriguez, A. – Joulin, A. – Grave, E. – Lample, G. LLaMA: Open and Efficient Foundation Language Models. *CoRR*. 2023a, abs/2302.13971. doi: 10.48550/ARXIV.2302.13971. Available at: https://doi.org/10.48550/arXiv.2302.13971.

Touvron, H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*. 2023b, abs/2307.09288. doi: 10.48550/ARXIV.2307.09288. Available at: https://doi.org/10.48550/arXiv.2307.09288.

Tran, M. – Zhang, Y. – Soleymani, M. Towards A Friendly Online Community: An Unsupervised Style Transfer Framework for Profanity Redaction. In Scott, D. – Bel, N. – Zong, C. (Ed.) *Proceedings of the 28th International Conference on ComputationalLinguistics, COLING 2020, Barcelona, Spain (Online), December 8-13,2020*, p. 2107–2114. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.190. Available at: https://doi.org/10.18653/v1/2020.coling-main.190.

Tunstall, L. – Beeching, E. – Lambert, N. – Rajani, N. – Rasul, K. – Belkada, Y. – Huang, S. – Werra, L. – Fourrier, C. – Habib, N. – Sarrazin, N. – Sanseviero, O. – Rush, A. M. – Wolf, T. Zephyr: Direct Distillation of LM Alignment, 2023.

Turing, A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*. 10 1950, LIX, 236, p. 433–460. ISSN 0026-4423. doi: 10.1093/mind/LIX.236.433. Available at: https://doi.org/10.1093/mind/LIX.236.433.

Vaswani, A. – Shazeer, N. – Parmar, N. – Uszkoreit, J. – Jones, L. – Gomez, A. N. – Kaiser, L. – Polosukhin, I. Attention is All you Need. In Guyon, I. – Luxburg, U. – Bengio, S. – Wallach, H. M. – Fergus, R. – Vishwanathan, S. V. N. – Garnett, R. (Ed.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, p. 5998–6008, 2017. Available at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, K. – Hua, H. – Wan, X. Controllable Unsupervised Text Attribute Transfer via editing Entangled Latent Representation. In Wallach, H. M. – Larochelle, H. – Beygelzimer, A. – d'Alché-Buc, F. – Fox, E. B. – Garnett, R. (Ed.) *Proc. NeurIPS*, p. 11034–11044, 2019. Available at: https://proceedings.neurips.cc/paper/2019/hash/8804f94e16ba5b680e239a554a08f7d2-Abstract.html.

Wang, L. – Yao, J. – Tao, Y. – Zhong, L. – Liu, W. – Du, Q. A Reinforced Topic-Aware Convolutional Sequence-to-Sequence Model for Abstractive Text Summarization. In Lang, J. (Ed.) *Proceedings of the Twenty-Seventh International Joint Conference onArtificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm,Sweden*, p. 4453–4460. ijcai.org, 2018. doi: 10.24963/ijcai.2018/619. Available at: https://doi.org/10.24963/ijcai.2018/619.

Wang, Z. – Samsten, I. – Papapetrou, P. Counterfactual Explanations for Survival Prediction of Cardiovascular ICU Patients. In Tucker, A. – Henriques Abreu, P. – Cardoso, J. – Pereira Rodrigues, P. – Riaño, D. D. (Ed.) *Artificial Intelligence in Medicine*, p. 338–348, Cham, 2021. Springer International Publishing. ISBN 978-3-030-77211-6.

Wang, Z. – Samsten, I. – Kougia, V. – Papapetrou, P. Style-transfer counterfactual explanations: An application to mortality prevention of ICU patients. *Artificial Intelligence in Medicine*. 2023, 135, p. 102457. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2022.102457. Available at: https://www.sciencedirect.com/science/article/pii/S0933365722002093.

Wei, C. – Wang, B. – Jay Kuo, C.-C. Synwmd: Syntax-aware word Mover's distance for sentence similarity evaluation. *Pattern Recognition Letters*. 2023, 170, p. 48–55. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2023.04.012. Available at: https://www.sciencedirect.com/science/article/pii/S0167865523001174.

Wei, J. – Wang, X. – Schuurmans, D. – Bosma, M. – Xia, F. – Chi, E. – Le, Q. V. – Zhou, D. – others. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*. 2022, 35, p. 24824–24837.

Weller, O. – Fulda, N. – Seppi, K. Can Humor Prediction Datasets be used for Humor Generation? Humorous Headline Generation via Style Transfer. In Klebanov, B. B. – Shutova, E. – Lichtenstein, P. – Muresan, S. – Wee, C. – Feldman, A. – Ghosh, D. (Ed.) *Proceedings of the Second Workshop on Figurative Language Processing*, p. 186–191, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.figlang-1.25. Available at: https://aclanthology.org/2020.figlang-1.25/.

Weston, A. R. S. C. J. A Neural Attention Model for Abstractive Sentence Summarization, 2015.

Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. In Liu, Q. – Schlangen, D. (Ed.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, p. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. Available at: https://aclanthology.org/2020.emnlp-demos.6/.

Wu, L. – Liu, P. – Yuan, Y. – Liu, S. – Zhang, Y. Context-aware style learning and content recovery networks for neural style transfer. *Information Processing & Management*. 2023, 60, 3, p. 103265.

Wu, X. – Zhao, W. – Luo, J. Learning cooperative neural modules for stylized image captioning. *International Journal of Computer Vision*. 2022, 130, 9, p. 2305–2320.

Wu, X. – Liu, J. – Li, X. – Xu, J. – Chen, Y. – Zhang, Y. – Huang, H. Improving Stylized Neural Machine Translation with Iterative Dual Knowledge Transfer. In *IJCAI*, p. 3971–3977, 2021.

Wu, Y. – Wang, Y. – Liu, S. A Dataset for Low-Resource Stylized Sequence-to-Sequence Generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI2020, The Thirty-Second Innovative Applications of Artificial IntelligenceConference, IAAI 2020, The Tenth AAAI Symposium on EducationalAdvances in Artificial Intelligence, EAAI 2020, New York, NY, USA,February 7-12, 2020*, p. 9290–9297. AAAI Press, 2020. Available at: https://aaai.org/ojs/index.php/AAAI/article/view/6468.

Xiao, F. – Pang, L. – Lan, Y. – Wang, Y. – Shen, H. – Cheng, X. Transductive Learning for Unsupervised Text Style Transfer. In Moens, M. – Huang, X. – Specia, L. – Yih, S. W. (Ed.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021*, p. 2510–2521, Virtual Event / Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.195. Available at: https://doi.org/10.18653/v1/2021.emnlp-main.195.

Xing, C. – Wu, W. – Wu, Y. – Liu, J. – Huang, Y. – Zhou, M. – Ma, W. Topic Aware Neural Response Generation. In Singh, S. – Markovitch, S. (Ed.) *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence,February 4-9, 2017, San Francisco, California, USA*, p. 3351–3357. AAAI Press, 2017. Available at: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14563.

Xu, P. – Cheung, J. C. K. – Cao, Y. On Variational Learning of Controllable Representations for Text without Supervision. In *Proceedings of the 37th International Conference on Machine Learning,ICML 2020, 13-18 July 2020, Virtual Event*, 119 / *Proceedings of Machine Learning Research*, p. 10534–10543. PMLR, 2020. Available at: http://proceedings.mlr.press/v119/xu20a.html.

Xu, W. – Ritter, A. – Dolan, B. – Grishman, R. – Cherry, C. Paraphrasing for Style. In Kay, M. – Boitet, C. (Ed.) *Proc. COLING*, p. 2899–2914. Indian Institute of Technology Bombay, 2012. Available at: https://aclanthology.org/C12-1177/.

Xue, L. – Constant, N. – Roberts, A. – Kale, M. – Al-Rfou, R. – Siddhant, A. – Barua, A. – Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In Toutanova, K. – Rumshisky, A. – Zettlemoyer, L. – Hakkani-Tur, D. – Beltagy, I. – Bethard, S. – Cotterell, R. – Chakraborty, T. – Zhou, Y. (Ed.) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. Available at: https://aclanthology.org/2021.naacl-main.41/.

Yamshchikov, I. P. – Shibaev, V. – Nagaev, A. – Jost, J. – Tikhonov, A. Decomposing Textual Information For Style Transfer. In Birch, A. – Finch, A. M. – Hayashi, H. – Konstas, I. – Luong, T. – Neubig, G. – Oda, Y. – Sudoh, K. (Ed.) *Proceedings of the 3rd Workshop on Neural Generation and Translation@EMNLP-IJCNLP2019, Hong Kong, November 4, 2019*, p. 128–137. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-5613. Available at: https://doi.org/10.18653/v1/D19-5613.

Yamshchikov, I. P. – Shibaev, V. – Khlebnikov, N. – Tikhonov, A. Style-transfer and Paraphrase: Looking for a Sensible Semantic Similarity Metric. *Proceedings of the AAAI Conference on Artificial Intelligence*. May 2021, 35, 16, p. 14213–14220. doi: 10.1609/aaai.v35i16.17672. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/17672.

YANG, D. – JIN, Q. Attractive Storyteller: Stylized Visual Storytelling with Unpaired Text. In ROGERS, A. – BOYD-GRABER, J. – OKAZAKI, N. (Ed.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 11053–11066, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.619. Available at: https://aclanthology.org/2023.acl-long.619/.

YANG, K. – LIU, D. – LEI, W. – YANG, B. – WEI, X. – LIU, Z. – XIE, J. Fantastic Expressions and Where to Find Them: Chinese Simile Generation with Multiple Constraints. In ROGERS, A. – BOYD-GRABER, J. – OKAZAKI, N. (Ed.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 468–486, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.28. Available at: https://aclanthology.org/2023.acl-long.28/.

YANG, M. – ZHAO, Z. – ZHAO, W. – CHEN, X. – ZHU, J. – ZHOU, L. – CAO, Z. Personalized Response Generation via Domain adaptation. In KANDO, N. – SAKAI, T. – JOHO, H. – LI, H. – VRIES, A. P. – WHITE, R. W. (Ed.) *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, p. 1021–1024. ACM, 2017. doi: 10.1145/3077136.3080706. Available at: https://doi.org/10.1145/3077136.3080706.

YANG, M. – QU, Q. – LEI, K. – ZHU, J. – ZHAO, Z. – CHEN, X. – HUANG, J. Z. Investigating Deep Reinforcement Learning Techniques in Personalized Dialogue Generation. In PEDRESCHI, M. E. (Ed.) *Proceedings of the 2018 SIAM International Conference on Data Mining,SDM 2018, May 3-5, 2018, San Diego Marriott Mission Valley, SanDiego, CA, USA*, p. 630–638. SIAM, 2018. doi: 10.1137/1.9781611975321.71. Available at: https://doi.org/10.1137/1.9781611975321.71.

YAO, S. – YU, D. – ZHAO, J. – SHAFRAN, I. – GRIFFITHS, T. – CAO, Y. – NARASIMHAN, K. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*. 2024, 36.

YEH, C. – RAMOS, G. – NG, R. – HUNTINGTON, A. – BANKS, R. GhostWriter: Augmenting Collaborative Human-AI Writing Experiences Through Personalization and Agency, 2024.

YOUNG, T. – CAMBRIA, E. – CHATURVEDI, I. – ZHOU, H. – BISWAS, S. – HUANG, M. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32, 2018.

ZENG, Y. – NIE, J.-Y. A Simple and Efficient Multi-Task Learning Approach for Conditioned Dialogue Generation. In TOUTANOVA, K. – RUMSHISKY, A. – ZETTLEMOYER, L. – HAKKANI-TUR, D. – BELTAGY, I. – BETHARD, S. – COTTERELL, R. – CHAKRABORTY, T. – ZHOU, Y. (Ed.) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, p. 4927–4939, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.392. Available at: https://aclanthology.org/2021.naacl-main.392/.

Zhai, W. – Rusert, J. – Shafiq, Z. – Srinivasan, P. Adversarial Authorship Attribution for Deobfuscation. In Muresan, S. – Nakov, P. – Villavicencio, A. (Ed.) *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 7372–7384, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.509. Available at: https://aclanthology.org/2022.acl-long.509/.

Zhang, K. – Shasha, D. Simple Fast Algorithms for the Editing Distance between Trees and Related Problems. *SIAM Journal on Computing*. 1989, 18, 6, p. 1245–1262. doi: 10.1137/0218082. Available at: https://doi.org/10.1137/0218082.

Zhang, S. et al. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*. 2022.

Zhang, T. – Kishore, V. – Wu, F. – Weinberger, K. Q. – Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020a. Available at: https://openreview.net/forum?id=SkeHuCVFDr.

Zhang, Y. – Sun, S. – Galley, M. – Chen, Y. – Brockett, C. – Gao, X. – Gao, J. – Liu, J. – Dolan, B. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In Celikyilmaz, A. – Wen, T. (Ed.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, p. 270–278. Association for Computational Linguistics, 2020b. doi: 10.18653/v1/2020.acl-demos.30. Available at: https://doi.org/10.18653/v1/2020.acl-demos.30.

Zhang, Z. – Ren, S. – Liu, S. – Wang, J. – Chen, P. – Li, M. – Zhou, M. – Chen, E. Style Transfer as Unsupervised Machine Translation. *CoRR*. 2018, abs/1808.07894. Available at: http://arxiv.org/abs/1808.07894.

Zhao, J. J. – Kim, Y. – Zhang, K. – Rush, A. M. – LeCun, Y. Adversarially Regularized Autoencoders. In Dy, J. G. – Krause, A. (Ed.) *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, 80 / *Proceedings of Machine Learning Research*, p. 5897–5906, Stockholm, Sweden, 2018. PMLR. Available at: http://proceedings.mlr.press/v80/zhao18b.html.

Zhao, W. X. – Zhou, K. – Li, J. – Tang, T. – Wang, X. – Hou, Y. – Min, Y. – Zhang, B. – Zhang, J. – Dong, Z. – others. A survey of large language models. *arXiv preprint arXiv:2303.18223*. 2023.

Zheng, Y. – Chen, Z. – Zhang, R. – Huang, S. – Mao, X. – Huang, M. Stylized Dialogue Response Generation Using Stylized Unpaired Texts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, p. 14558–14567. AAAI Press, 2021. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/17711.

ZHOU, H. – HUANG, M. – ZHANG, T. – ZHU, X. – LIU, B. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In MCILRAITH, S. A. – WEINBERGER, K. Q. (Ed.) *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, p. 730–739. AAAI Press, 2018. Available at: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16455.

ZHU, K. – WANG, J. – ZHOU, J. – WANG, Z. – CHEN, H. – WANG, Y. – YANG, L. – YE, W. – GONG, N. Z. – ZHANG, Y. – OTHERS. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*. 2023.

ZOU, A. – WU, X. – LI, X. – ZHANG, T. – CUI, F. – XU, J. Curriculum pre-training for stylized neural machine translation. *Applied Intelligence*. 2024, p. 1–11.