Evaluating Style Transfer in Natural Language

by

Nicholas Matthews

S.B., C.S. M.I.T., 2017

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Author .....
Department of Electrical Engineering and Computer Science
May 28, 2018

Certified by .....
Professor Regina Barzilay, Thesis Supervisor
May 28, 2018

Accepted by .....
Katrina LaCurts, Chair, Master of Engineering Thesis Committee
May 28, 2018

Evaluating Style Transfer in Natural Language

by

Nicholas Matthews

Submitted to the Department of Electrical Engineering and Computer Science
on May 28, 2018, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

Style transfer is an active area of research growing in popularity in the Natural Language setting. The goal of this thesis is present a comprehensive review of style transfer tasks used to date, analyze these tasks, and delineate important properties and candidate tasks for future methods researchers. Several challenges still exist, including the difficulty of distinguishing between content and style in a sentence. While some state of the art models attempt to overcome this problem, even tasks as simple as sentiment transfer are still non-trivial. Problems of granularity, transferability, and distinguishability have yet to be solved. I provide a comprehensive analysis of the popular sentiment transfer task along with a number of metrics that highlight its shortcomings. Finally, I introduce possible new tasks for consideration, news outlet style transfer and non-parallel error correction, and provide similar analysis for the feasibility of using these tasks as style transfer baselines.

Acknowledgments

Special thanks to my supervisor, Professor Regina Barzilay for her consistent guidance and support, and clear vision in this area of research. Thanks to Tianxiao Shen and Zhi-Jing Jin whose work was incredibly instrumental in writing this thesis. Additional thanks to Casey Crownhart, Andrew Mullen, and Suri Bandler for their helpful discussions and feedback; and my family for their love and support.

Contents

List of Tables

Chapter 1

Introduction

Style Transfer is an increasingly important and relevant text-to-text generation task in the field of Natural Language Processing (NLP) that aims to transform text in an intelligent way that preserves semantic content while modulating stylistic features. While state-of-the-art models continue to improve across a number of automatic and human evaluation metrics, the task itself remains difficult to define and evaluate [17, 7, 21, 14, 11, 1]. The goal of this thesis is present a comprehensive review of style transfer tasks used to date, analyze these tasks, and delineate important properties and candidate tasks for future methods researchers.

One of the fundamental challenges in style transfer is making the distinction between style and content in natural language, whether implicitly or explicitly. I present both important abstract properties and simple, pragmatic techniques for analyzing this distinction between style and content according to a given task definition and dataset. In light of this analysis I compare three style-transfer tasks by training and evaluating contemporary models. Finally, I analyze these results and present informed counsel on directions for future research that is both rigorous and meaningful.

Precedents in computer vision and machine translation inform style transfer in NLP, but key differences highlight the need for a unique research discipline. Unlike in most machine translation work, style transfer does not assume a parallel corpus of sentence

pairs [17]. And unlike in computer vision, linguistic representation complicates the distinction between style and content and generally increases the difficulty of generative tasks. Typically words are represented as vectors [13] but do not have the same convenient properties facilitating stylistic analysis.

Despite these challenges, researchers have built impressive and sometimes complicated models to learn how to translate between styles even in a non-parallel setting, while making some distinction between style and content. Most work either attempts to model the distinction between style and content or to construct a dataset of pseudopairs using various similarity metrics [1]. For considerations relating to modeling difficult and available training data, contemporary research and this thesis deal explicitly with the task of translating between two opposed style classes ("Binary style transfer").

I approach the question of evaluating style transfer models from two motivating questions: First, how can we evaluate our task definition and dataset to ensure they reflect what we hope to achieve at a conceptual and intuitional level? Second, what metrics can we use to determine what success looks like, both quickly and automatically for rapid model development, and methodically and rigorously at test time to compare different approaches more deeply? I report the findings of applying my methods to three candidate style transfer tasks:

- sentiment transfer on Yelp business reviews [1]
- general style transfer between news outlets with disparate audiences
- flexible, customized error correction (in the non-parallel, low-resource setting) on Yelp business reviews.

Results of the sentiment transfer demonstrate that the task is acceptable but major flaws result in low human performance, inconsistent measurements across experiments, and unclear evaluation.

Chapter 2

Background and Related Work

Style transfer has a long history in the context of machine learning, beginning in computer vision. In NLP, the task of style transfer is analogous to machine translation, although crucial details differ. After introducing the task setting, I present this background in computer vision and NLP, before moving on to state-of-the-art style transfer methods and tasks used to evaluate these methods.

General Task Setting Generally, the task of Style Transfer is to learn some transformation function f between styles:

$$f(x_i) = x_j \tag{2.1}$$

where x_i and x_j indicate the sentence x has some stylistic attributes i and j respectively. Further, the task posits that there is some content latent in x_i called z that should be maintained throughout the transformation. In other words, x_i and x_j share semantic content z, but differ in their stylistic attributes. In style transfer, contemporary research concerns itself with the binary case: there are just two styles s_1 and s_2, and the goal is to learn both transformations

$$f_{12}(x_1) = x_2 \tag{2.2}$$

$$f_{21}(x_2) = x_1 \tag{2.3}$$

where the subscript on f denotes the direction of the transformation between the two styles. This definition approximates everything from machine translation of natural language to stylistic image transformation; however, while those settings often involve a parallel training corpus of (x_1, x_2) pairs, this is often not the case for style transfer in NLP. Style transfer often operates in the non-parallel setting, so there are not direct (x_1, x_2) pairs in the training corpus. Of course, one *could* learn on direct training pairs, but conceptually one of the primary goals of style transfer in NLP is to conquer the goal of expressing the same ideas in different forms that almost never occur side by side in the world. With the exception of limited resources like paired translations of text (e.g. the King James vs New American Bible) and error correction datasets like Berzak et al., Yannakoudakis et al. consider, one would not expect to find a corpus that pairs stylistically diverse data (e.g. Ernest Hemingway sentences with James Joyce sentences, academic language with pop science writing) with exact matches in content. The non-parallel setting will introduce a number of challenges to the general task of style transfer. Namely, one needs to tease out the difference between content and style directly, or construct training pairs (x_1, x_2) in an unsupervised way [1].

Motivation Generating text that maintains or modulates various properties of some source text would be immensely useful. For NLP research, the style transfer task setting encourages novel work in representation for natural language and improvement in text-to-text generation models in the non-parallel setting. Successful style transfer models can be used to generate new training data for resource poor domains and alleviate class imbalance by transforming text with common stylistic attributes to text with uncommon stylistic attributes. In the general text generation setting, it will often be the case in practice that the output language should have certain properties e.g. adjusted tone and word usage to accommodate certain audiences and objectives.

2.1 Related Work in Vision

In computer vision, style transfer's feasibility stems from the fundamental representation: images are tensors of real numbers with locality relations that constitute forms, shapes and objects that humans can interpret. Further, these patterns in vision are resilient to small perturbations, and even some global transformations (e.g. modulating color properties with RGB, HSV channels, or applying other filters). Naturally then, some patterns can be taken as the content of the image while rigorous definitions of style can (theoretically) be constructed to describe the specific way in which these patterns are instantiated.

While early computer vision work in style transfer did precisely this, using classical signal processing techniques, recent deep learning methods have proved even more powerful and generally applicable. Many of these methods employ some form of Generative Adversarial Network (GAN) [5] to learn how to generate a synthetic image with some target style that is indistinguishable from true data in the target style.
The distinction between style and content is difficult to identify in language, but it is relatively robust in images: one can take an image in one style e.g. photo-realism and express it in different style e.g. painted, Monet-like brushstrokes, like in Zhu et al.'s work, while still representing the same scene. However, the filters and brush-strokes of language are difficult to identify so clearly.

2.2 Related work in NLP

In NLP several common problems arise in generation tasks, beginning with finding the right representation for language. Unlike in computer vision, changing a single value (word or token) can dramatically alter the meaning of the sentence. Further, words are discrete tokens that need to be transformed into the appropriate numerical representations if machine learning methods are to be used, and intricate annotations are needed for more traditional linguistic methods. In the context of style transfer

this means that small errors can result in completely different content, nonsensical grammar, or implausible stylistic detail. Still, effective models in machine translation and style transfer have begun to address these challenges.

Machine Translation Most precedents in task formulation and methodology for style transfer come from machine translation. Usually machine translation utilizes parallel training corpora, and the most competitive state of the art models use deep sequence to sequence models that first encode some text in the source language before transforming the latent content into the target language [2]. Recent improvements to these methods have produced translations that rival human ability according to Hassan et al. but to do so the models utilize parallel corpora on the order of 20 million sentences or more in production systems [6]. Recently though, some progress has been made in the non-parallel setting, where Lample et al. construct a common latent space between source and target languages. They train the model according to two principles: the model should be able to (1) reconstruct a sentence after noise is added and (2) translate the domain invariant encoding into the target language. They train the latter by generating target sentences with the model that are then translated into the source language and compared against original, natural sentences. This approach leads nicely into some recent work that is discussed in the next section (2.3); the main difference is that this recent work [1] always translates between two sentences that come from the different style sub-corpora (noisy pseudopairs), whereas Lample et al. use model outputs as noisy inputs.

Evaluation in Machine Translation Evaluating the quality of machine translation systems is difficult and time consuming, because it involves human judges evaluating every sentence. For this reason, automatic measures such as BLEU were invented as proxies or "understudies" to human evaluation [12]. With quality human references, high BLEU scores tend to correlate with favorable human judgments. However, BLEU scores have some known problems: a correct translation that diverges from the human references may have an inappropriately low score. While the

authors claim that over an entire test set the BLEU score is more accurate and pathological results average with other closely matching sentences, having a single reference for a task as subjective and nuanced as translation is problematic. Further, because of this rigidity to the reference, coupled with a lack of emphasis on grammaticality, BLEU scores typically can only distinguish between poor models and potentially acceptable models and are not applicable in distinguishing between human quality and or near human quality translations [6].

With these problems in mind, it is apparent that human evaluations are still necessary and cannot be completely replaced with automatic evaluation; however, it is not always apparent how to define success. Different researchers tend to produce their own definitions of success. For example, Hassan et al. use the metric of human parity, which they define as:

Definition 2.1. *If a bilingual human judges the quality of a candidate translation produced by a human to be equivalent to one produced by a machine, then the machine has achieved* HUMAN PARITY

They generalize this definition to an entire test set as:

Definition 2.2. *If there is no statistically significant difference between human quality scores for a test set of candidate translations from a machine translation system and the scores for the corresponding human translations then the machine has achieved* HUMAN PARITY

While human parity might seem like a noble goal, for style transfer it is important to keep in mind that human parity is not always indicative of general success, because the task definition itself may make it difficult for even humans to produce good results, or for humans to make fair and consistent judgments, as will be discussed in later sections. While the task definition of language translation is more concrete, there are still a number of extraneous variables that affect human evaluations which Hassan et al. point out:

- **Variability of quality measure** How sensitive is our quality measure to different subsets of data?
- **Test set selection** Would we likely obtain the same result on slightly different test data?
- **Annotator errors** What if some human annotators become inattentive, unfairly improving or damaging the score of one system over the other?
- **Annotator consistency** What if the human annotators produce different scores given the same data? Would using different human annotators still lead to the same conclusion?

In this thesis, as has been done in recent work [1], I account for the effects of test set selection and annotator consistency specifically as discussed in Chapter 4.

2.3 State of the Art

2.3.1 Methods

Many new approaches to style transfer in NLP have emerged to respond to the difficult non-parallel task setting. In "Iterative Matching and Translation for Non-Parallel Style Transfer," the authors succinctly report on the failure of adversarial methods (see [17, 4] for more on adversarial methods) to separate content and style in the latent space as indicated by consistently poor content preservation and fluency evaluations by human judges [1]. This trend changed with the introduction of simpler methods: Li et al.'s DeleteAndRetrieve model deletes salient (see equation 3.7) words from a sentence, retrieves related salient words from the opposite style, and then uses a recurrent neural editor to smoothly insert the relevant words. Their success was based on the observation that in simple cases that occur frequently in tasks like sentiment transfer, the solution primarily involves recognizing the sentiment bearing phrases and choosing appropriate antonymous phrases, which I denote as the "weak" case of style transfer in Section 3.2.1. To better generate grammatical sentences and

handle the "strong" case where more intricate rewriting is required, the authors of "Iterative Matching and Translation for Non-Parallel Style Transfer" introduce an iterative method of constructing pseudo-pairs to learn direct sequence-to-sequence translation.

2.3.2 SotA Evaluation

For every new style transfer method developed, a different evaluation procedure has been proposed, making it difficult to compare different methods and discern how well state-of-the-art performs. I will briefly recap these datasets and metrics here before proposing my own evaluation methods and results in Chapters 3 and 4.

Toy Tasks in Style Transfer

Shen et al. utilize two notable baseline tasks that are both simple and clearly defined. The first is word substitution deciphering. In this task, a 1-to-1 word substitution mapping is chosen and then applied to the non-parallel training corpus to generate an alternative cipher style. This task highlights the difficulty of the non-parallel setting as parallel translation models solve the task handily. On the other hand, this toy task does little to test a model's ability to generate novel, grammatical sentences. It also does little to directly judge more subtle aspects of the style transfer task, which include picking up on word usage and learning the contextual distinction between style and content. Shen et al. also propose word order recovery on scrambled sentences, which suffers from the same lack of testing coverage. Additionally, Shen et al. note that some generated re-orderings will be valid (well-formed English sentences) but nevertheless penalized, unlike in other style transfer tasks where a diverse set of solutions are both acceptable and sought after.

Linguistic Feature Transfer

When asked to think of what attributes constitute "style" in natural language, sentence structure, verb mood, tense, word choice, and voice all come to mind as candi-

dates. Importantly, the distinction between content and style is always contextual; in many situations tense or time setting would be considered content, but modulating the tense of verbs fluently while maintaining all other content is also a reasonable and meaningful goal. In this case then, verb tense would be style, not content. Hu et al. propose using verb tense labels of {"past", "present", "future"} as stylistic categories and learning to translate between them. They do so by compiling from the TimeBank (timeml.org) dataset and obtain a lexicon of 5250 words and phrases labeled accordingly. They demonstrate promising results as reproduced in Table 2.2. Linguistic transformations could in general could be effective baseline tasks for judging style transfer success, but often they only apply to limited subsets of data. Namely, many sentences have several verb phrases; transforming all of them will not always result in a sensible sentence, and refining the task objective for all possible input sentences is impossible. These problems arise because the stylistic property's granularity is phrase or constituent level, rather than sentence level (See Section 3.2.2). The same problems plague alternatives like translating between active and passive voice. Less control over possible output features means less applicability.

Varying verb tense in simple cases
this was one of the outstanding thrillers of the last decade
this is one of the outstanding thrillers of the all time
this will be one of the outstanding thrillers of the all time

Table 2.1: Example model output from [7] where the stylistic property of verb tense is modulated.

Varying verb tense in more complex cases
I **walked** in hoping for a thriller but little **did I know** I **would** be sorely disappointed
I **walked** in hoping for a thriller but little **do I know** I **will** be sorely disappointed
I **walk** in hoping for a thriller but little **do I know** I **will** be sorely disappointed
I **will walk** in hoping for a thriller but little **do I know** I **will** be sorely disappointed
I **will walk** in hoping for a thriller but little **will I know** I **will** be sorely disappointed

Table 2.2: Selected example of the case where the verb-tense modification task breaks down. There are multiple verbs, modifying the tense of each changes the meaning, and ultimately if all solutions are accepted, the task becomes less focused and meaningful.

Topic and Perspective Transfer

Several authors explore the possibility of re-expressing sentences as they pertain to different topics and perspectives. Zhao et al. report their results translating between three different Yahoo Answers topics, music, science and politics. For example, "republicans : would you vote for a cheney / satan ticket in 2008 ?"[Politics] becomes "guys : how would you solve this question ?"[Science] when translated by their model. There are several obvious problems, including (1) how does one judge success with such a loose definition of content? and (2) to what end would this transformation be useful?

Prabhumoye et al. look at translating between the styles of different political slants, namely democratic and republican. They cite different ways users thank their representatives: "thank u james, praying for all the work u do ."[Republican] and "on behalf of the hard-working nh public school teachers- thank you !"[Democrat]. Because meaning preservation is difficult to define in style transfer, the authors relax the constraint of literal meaning preservation to *affect* or *intent* preservation in the context of the discourse. Relaxing the constraints is a tradeoff between difficulty (in both translation and evaluation) and utility. There may be scenarios where their objective makes sense, but treating content so loosely may make the task almost trivial.

I propose one goal of style transfer should be to define task properties and evaluation methods more precisely so constraint relaxation is not necessary. In the context of Prabhumoye et al.'s political slant transfer task, it is possible that democrats and republicans have different stylistic tendencies apart from content words like author identity ("teachers") such that the constraint need not be relaxed. They report that human evaluators prefer neither candidate model output 45.9% of the time, reflecting the difficulty of ascertaining the political slant and judging intent preservation. They do not report human performance on the task.

Sentiment Transfer

The most common benchmark task for judging style-transfer capability is sentiment transfer: given a sentence expressing positive or negative sentiment, transform the sentence into a similar one expressing the same content but the opposite sentiment [17, 7, 21, 14, 11]. One of the main reasons for the popularity of this task is the abundance of data in the form of Yelp, Amazon, and IMDB reviews rather than the conceptual properties of sentiment transfer. While the task is clear in simple cases (e.g. translating "delicious food" into "gross food") sentences can be dramatically more complex and ill-formed. Often the line between content and style/sentiment blur like in example 1 in Table 4.6. Granularity of sentiment is an issue (Section 3.2.2), as well as the noisy "gold" labels in training data: Shen et al., Zhu et al., Li et al. all take review level labels in place of non-existent sentence level labels. Further, not all input sentences have a clear overall sentiment (Section 3.2.2) and not all sentences' sentiment can be transferred (Section 3.2.3). The quality of model evaluations and human performance documented in Section 4 illustrates the effects of these concerns.

Chapter 3

Evaluation Methods

3.1 Formal Definition

First, we define the style function as a mapping from some text to a style representation.

$$S(x) : x \longrightarrow s \tag{3.1}$$

x is typically a sentence (sequence of tokens); s is typically a binary variable or discrete class. In contemporary research this definition is implicit and instead the definition of style is instantiated implicitly by the choice of corpora. Defining style explicitly would be cumbersome (see section 3.2). x is then labeled x_s when it belongs to some corpus that represents style s. If $S(x)$ is considered at all, it has to be learned by some model using the training corpus or data that comes from the same distribution.

Ideally style transfer would be the task of learning the function to translate between explicit styles s_1 and s_2

$$f(x_1) = x_2 | S(x_1) = 1, S(x_2) = s_2 \tag{3.2}$$

but without an explicit definition of style this usually becomes

$$f(x_1) = x_2 | x_1 \in X_1, S'(x_2) = s_2 \tag{3.3}$$

where $S'(x)$ is a style judgment made by either a classifier or human judge. Crucially, parallel training pairs (x_1, x_2) are not available. Finally, the additional constraint of content preservation can be represented as

$$z_1 = z_2 \tag{3.4}$$

where z represents the content of each x, although the distinction between content z and style s need not be modeled explicitly to complete the task. Human judges are typically asked to judge this constraint, although BLEU scores calculated with respect to human references correlate at least at current performance levels ([1], Section 3.4.1).

3.2 Defining a Style

While the definitions of content and style are instantiated implicitly by corpus selection and division, some notion of what they should be is necessary to propose evaluation metrics. Because the distinction between content and style need not be fixed, I instead propose a number of general properties that would favor useful, meaningful tasks with measurable results.

3.2.1 Strong versus Weak Transfer

The transfer task can be roughly divided into two categories: *weak* and *strong*. The weak case involves replacing or changing select words in a sentence, leaving the overall structure intact, and most words unchanged.

The weak case arises when either (1) the definition of the styles or attributes are sufficiently related, so that any given string of text having one of the two attributes can be converted with a few word deletions and/or insertions or (2) the specific string of text happens to be solvable in a weak way. For example, two authors may have very

different styles that normally require sophisticated rewriting of sentences to transfer between the two voices, but an especially short sentence with uncomplicated subject matter might be transferable in a simple way. The strong case arises when many words and the syntactic structure of the sentence change dramatically.

Note: Edits are loosely defined on purpose: a transformation from active to passive voice on a single clause sentence could keep almost all words the same and wed define this task as weak transfer; however, if you look at a fixed metric like the number of insertions and deletions, it might imply that simply reversing the order of clauses in a sentence and modulating verb voice is an example of strong transfer. This example highlights the need to be wary of the distance metrics used in style transfer, and further points to the fact that strong and weak are guideposts that inform the complexity of method appropriate for the task, rather than perfect measures at the case by case level.

Mathematically, strong and weak can be distinguished by either

$$\text{Distance}(x_1, x_2) > \gamma \tag{3.5}$$

or

$$\text{Distance}(x_1, f(x_1)) > \gamma \tag{3.6}$$

for some choice of distance function, such as token Jaccard Distance [CITE or note], and reasonable threshold γ, depending on whether or not access to exact pairs is available. If the distance is above the threshold it would be the strong case; γ is not necessary to define in general as it depends on the task and distance metric used, but this formulation is useful for comparing definitions of style across an entire corpus. In practice, because of the nonparallel constraint, one would likely use a test set translated by humans to estimate the corpus value.

3.2.2 Classifiability

Any given definition of a style or attribute may not be defined on any given string of text. In the case of sentiment transfer, if you define the domain of attribute values as ("positive", "negative") most sentences may have undefined style. e.g. "I walked into the restaurant around noon." is neutral.

Granularity and Distinguishability For some attributes or styles, the granularity of the values in relation to what they represent should be considered. Even if you extend the aforementioned attribute domain to "positive", "negative", "neutral" many sentences may have unclear values — e.g. a sentence of the form "I liked x but I didnt like y" could be classified as any of the three values. Individual phrases like "I liked x" have clear sentiment, but the typical granulariy used in contemporary research is sentence-level style. On the other hand, it would at least be an improvement to consider the set of values "positive", "negative", and "neither" because you can simply place ambiguous sentences in the third category. Of course, generally adding the complement of the union of all attributes as your final, default class, will result in a relatively large class that may not have any meaning.

To generalize the earlier "I liked x but I didnt like y" case, consider any body of text that is formed by splicing text together from all possible styles. When defining a framework for the task, method, and evaluation, we must consider to what extent such a pathological case is relevant, and if the style or attribute should be defined in this case. It might seem like such a sentence would be nonsense, but you can easily construct a sentence that looks like this in general, "I liked x but I didn't like y" being a prime example, so it cannot be ignored. If one wants a rigorous and robust definition of style or stylistic-attributes, the only solution (besides "catch-all" classes) is to consider every piece of text (word, sentence, or sequence depending on how your model works) as having some combination of all possible styles or attributes. Crafting such a representation by hand or with neural models is difficult or impossible without the right data, so I will continue to focus on binary class based style and

use granularity and distinguishability as guiding principles for metrics I introduce in chapter 4.

3.2.3 Transferability

Not all sequences of tokens can be transferred. Even if you restrict the set of sentences to (1) grammatical and (2) classifiable sentences, it may still be the case that a sentence with a different style attribute and related content cannot be generated. Transferability depends on the definition of content and more broadly the relationship between the various attributes and content.

Entanglement Some combinations of content and attributes (or content and style) will inherently result in a strong entanglement between content and style. The exact nature of the relation depends on the definitions and metrics adopted, so the vagueness of the word entanglement is appropriate. One possible solution to entanglement would be to consider something like the covariance between content and style, provided the underlying representations accept this measurement. For an attribute definition like sentiment, you can analyze the number and frequency of words that bear content and attribute: in many cases, content words would be details like what food was ordered, and attribute words would might be words like delicious, or cold. In Section 3.3.1 I propose a practical way of doing this, namely using Li et al.'s saliency metric. Other times, the same words convey content and style, as in sentences like They raised my bill without telling me! Unfortunately, cases like these are difficult or impossible to translate without relaxing the content preservation constraint. This is a trade-off: depending on your goals it could be appropriate to relax the content preservation constraint, but in general style transfer becomes less meaningful the less content preservation is considered: Zhao et al.'s Yahoo Answers topic transfer and to a lesser degree in Prabhumoye et al.'s political slant transfer show the potential pitfalls of pushing aside content preservation. For example, in the Yahoo topic transfer task, there is practically no constraint on content and transfer becomes trivial; it can be solved by simply randomly sampling from the target style corpus.

3.2.4 Degeneracy

There may be degenerate identifiers (for classification) and degenerate transformations (for transfer). For example, for sentiment transfer, you could often add but I liked it anyway to transform a negative review into a positive one. One major difference between strong, general style transfer and simpler tasks like sentiment transfer is the presence of degenerate cases. In the former case, where you might be doing something like rewriting text as if it were written by a different author, its unlikely there are simple degenerate transformations or template-based solutions.

Further, the very existence of degenerate cases can be mitigated by tighter definitions of style and content. For example, if typical sentiment transfer were reformulated to specifically reverse the sentiment or perspective towards each item (rather than sentence level reversal) the aforementioned degenerate addition "but I liked it anyway" no longer represents a correct solution. Unfortunately this formulation requires detailed, expensive annotations so it is not considered further here.

3.2.5 Grammaticality or Fluency

Important in any generative natural language task is grammaticality and/or fluency. The terms are typically used interchangeably in literature to denote the quality of a sentence holistically, which is a product of correct grammatical usage, interpretability and overall "naturalness" or "fluency" best judged by native English speakers.

3.3 Analyzing Corpora

All of the properties outlined above are important to consider when defining style or picking a criteria for dividing a corpus into sub-corpora X_1 and X_2 for translation; however, most of them are difficult to measure in practice, so I propose a number of metrics to use to validate the features of a corpus before finalizing and beginning training. I will present the results of evaluating these metrics on my candidate tasks

in chapter 4.

3.3.1 Automatic Corpus Analysis

Because the definition of style is implicitly instantiated in corpus selection and division criteria it is crucial to validate the division against desired properties.

Term Frequencies Perhaps the simplest and most straightforward step in analyzing a style transfer corpus is looking at word counts. Despite immense effort to train models at the task, vocabulary analysis on sentiment transfer corpora is ill-reported. It may be the case that unexpected content-bearing words are particularly frequent causing translation models to suffer. Recently Li et al. suggested a specific term-frequency based metric for analyzing the "saliency" of n-grams which they define as:

$$\text{saliency}(u, v) = \frac{\text{count}(u, \mathcal{D}_v) + \lambda}{\sum_{v' \in \mathcal{V}, v' \neq v} \text{count}(u, \mathcal{D}_{v'}) + \lambda} \tag{3.7}$$

where $\text{count}(u, \mathcal{D}_v)$ denotes the number of times that n-gram u appears in $\mathcal{D}_v$ and λ is a smoothing parameter. Although they exploit a difference in keywords and phrases to learn a simple translation model for sentiment and topic transfer, they do little to show that this definition is valid. If most of the most salient words were completely unrelated to a desired definition of style, the task is likely not meaningful and doomed to fail. For example, if "Italian" is one of the most salient n-grams in a positive-sentiment Yelp corpus, the model may have trouble learning the distinction between style and content. Simply ranking your sub-corpora by saliency can be illuminating. See Section 4.1.1 for results on the Yelp dataset. After extracting sentences with these salient words, calculating the nearest neighbor in the other corpus reveals an ad hoc impression of the sentences that are likely to be compared, especially in recent methods like [1]. Nearest neighbors can be calculated by comparing the cosine distance of TF-IDF vector representations of sentences.

Classifiers Normally distinguishability is a goal for any corpora; to obtain a proxy measure of the ability to distinguish between style classes I train classifiers on all candidate corpora. In sentiment analysis, state of the art methods ranging from CNN text classification [8] to using intensively trained character-based representations like in [15] which reports upwards of 93% accuracy on a large Amazon dataset. It is crucial to note that for some arbitrary corpus this metric may be misleading: e.g. if one constructs a corpus to translate between the writing styles of two authors, but the two corpora talk about very different subjects, the classifier can use the disjoint vocabulary without relying on stylistic markers important to the spirit of the task. Thus it is important to consider content overlap alongside classifier accuracy, for example by using the saliency metric defined above and inspecting the most salient words of each corpus. Using a model that can supply rationales for its predictions can also alleviate this problem. For example Lei et al.'s text classifier is split into separate rationale extraction and prediction models such that a limited and mostly contiguous set of tokens used to make some prediction is always available. By inspecting the rationales used to make stylistic judgments one can get a rough but contextual sense for the difference between two styles.

3.3.2 Human Corpus Analysis

Automatic methods are still crude measurements and state-of-the-art stylistic analysis still requires human judgment. Tasking humans with distinguishing between two candidate styles should be a part of every experimental setup, and can help reveal potential issues with classifiability and transferability. Even in sentiment transfer on Yelp data, widely utilized as a style transfer task, is imperfect. Independent judgments by two humans on each sentence over 3,000 sentences indicates that [NUM] percent of the Yelp corpus has ambiguous sentiment, as shown in [TABLE].

I give humans the style transfer task itself, both as a means of justifying the task feasibility and as reference data for automatic measures like BLEU [12].

3.4 Evaluation Metrics

3.4.1 Automatic Evaluation

Automatic style classifiers like the aforementioned sentiment classifiers can effectively indicate style attribute success. In [17] the automatic classifier prediction inflates the true success rate by over 10% absolute, as shown in [TABLE 1+2 from tianxiao]. This is likely because of the entanglement between style and content: classifiers can learn degenerate content-bearing supports that just happen to be more present in one corpus or entirely absent in the other, as discussed in [11].

BLEU scores are widely used in Machine Translation and can be used to approximate content preservation scores in style transfer. Zhao et al. calculate BLEU scores with respect to the original input sentences positing that most of the words and phrases in the target style sentence will overlap with words and phrases in the input sentence. This assumption is only true in weak style transfer cases and even in those cases it would be difficult to compare models that perform close to human levels using BLEU scores with respect to inputs. BLEU scores with respect to human references completing the transfer task are more relevant in general, but do not necessarily capture fluency judgments well.

3.4.2 Human Evaluation

Because of the issues with classifiers and metrics like BLEU, human evaluations are still the gold standard. In style transfer, they typically come on three dimensions: (1) content preservation (2) fluency and (3) style correctness; all three are measured on Likert scales, usually from 1-5 [11, 1]. In judging content preservation, humans are asked to rate how well the output sentence preserves the sentence content apart from the stylistic attribute, compared with the original input sentence. (In [14] they are instead asked how well it preserves the affect or intent but as discussed in section

2.3 I do not relax to that constraint). Fluency requires humans to rate how natural-sounding and grammatical a sentence is. Finally attribute transfer success measures how well the output sentence embodies the target style attribute. See Table 4.6 for examples.

Input	Blue cheese dressing wasn't the best by any means.
Good Con.	Blue cheese dressing was delightful.
Poor Con.	The burger was yummy.
Input	Suzanne and her staff were excellent!
Good Gra.	The staff were rude and unhelpful.
Poor Gra.	I staff unhelpful.
Input	This is the best seafood joint in town.
Good Cor.	The seafood was not good.
Poor Cor.	The food here was great!

Table 3.1: Exemplary and bad transfer sentences in each of the evaluation categories: content preservation (Con.), fluency & grammaticality (Gra.) and style correctness (Cor.)

Chapter 4

Results and Discussion

4.1 Yelp Sentiment Transfer

In this task, the models are given a corpus of Yelp business reviews split into positive and negative sentiment sentences. The sentences coming from reviews with rating 4 or 5 are taken as positive (270K reviews), and rating 1 or 2 as negative (180K reviews). Long reviews are discarded because they are more likely to include background narration sentences with sentiment neither positive nor negative [17]. Mapping the review ratings to sentence-level sentiment is noisy, but acceptable as reported in Shen et al..

4.1.1 Yelp Sentiment Task Analysis

In Section 2.3 I discuss sentiment transfer as a task, and use the same formulation for these tests. My main objective is to demonstrate the pros and cons of sentiment transfer through the concepts and metrics delineated in this thesis.

Saliency One risk in using the Yelp dataset is that certain topics or content will occur more often in the positive reviews than in the negative reviews or vice versa. While this is not necessarily a blow to the task definition of sentiment transfer, it is a practical concern for certain model architectures that may be susceptible to these kinds of biases. To understand whether or not that bias is present I use saliency to rank the vocabulary (Equation 3.7). The top-10 words for each sub-corpora are

Negative	Positive
('unacceptable', 2151.0)	('must-try', 377.0)
('rudely', 1888.0)	('addicting', 277.3)
('**refund**', 1374.0)	('cutest', 243.0)
('rudest', 938.0)	('unmatched', 238.0)
('**cockroach**', 825.0)	('amazeballs', 170.0)
('roach', 744.0)	('delicioso', 163.0)
('slowest', 720.0)	('yumm', 150.5)
('unprofessional', 580.8)	('unassuming', 141.3)
('disgusted', 485.0)	('splendid', 136.0)
('shrugged', 461.0)	('delectable', 128.8)

Table 4.1: The most salient words and their saliency values (Equation 3.7) from the negative and positive sub-corpora of the Yelp dataset respectively. Most words are appropriately salient, although words like "refund" and "cockroach" are problematic.

reported in Table 4.1. After inspecting this list (and the next few hundred words to follow), it is apparent that on the basis of the lexicon, the sentiment transfer task with Yelp data is fairly sound. Most words on the list are words you would expect to be salient for the respective class e.g. "must-try", "addicting", and "yumm" for positive reviews, and "unacceptable", "rudely", and "unprofessional" for negative; on the other hand words like "refund" and "cockroach" denote content, indicating a model that over-utilizes saliency may suffer in content preservation.

Rationale In order to better understand what features are important for classifying a sentence as positive or negative, besides simple statistics like saliency, I use Lei et al.'s rationale model [1]. The F1 score of the model reaches 95% reliably, but has some trouble producing useful rationales for corpus analysis. Early in training (circa epoch 1), the model extracts short single word rationales that can actually allow the classifier to reach an 85% F1 score, and similar overall accuracy; this result makes sense given the importance of salient sentiment bearing words. On the other hand, later in training the model tends to extract long rationales and often the entire review. It is clear that extensive hyper parameter tuning is necessary to mitigate the effects

[1] I use Adam Yala's PyTorch implementation [19]

Negative Likelihood	Positive Likelihood	Rationale
0.69	0.32	hungry
0.94	0.06	rude
0.53	0.49	busy
0.28	0.76	tasty
0.85	0.16	skip
0.41	0.63	jonesing

Table 4.2: Early Rationale model predictions for likelihood of belonging to the Negative sentiment and Positive sentiment sub-corpora of the Yelp dataset. After only 1-2 epochs the model uses small rationales which often, but of course fail in non-trivial cases.

Negative Likelihood	Positive Likelihood	Rationale
0.91614914	0.09816217	few hours later and we were so hungry again
0.9730762	0.03818279	when you are new to a game they are soo rude...
0.05112737	0.9340056	can be really busy at times
0.00359428	0.994449	loved the tasty , crispy , cheese toast
0.98269063	0.02461735	if you are hungry skip this place
0.54868597	0.4402116	i just had a big lunch at nami down the street...

Table 4.3: Late Rationale model predictions for likelihood of belonging to the Negative sentiment and Positive sentiment sub-corpora of the Yelp dataset. After only 10+ epochs the model uses large rationales and even the entire sentence.

of noise in the dataset.

Human Distinguishability and Transfer Sentiment classification is a popular task in NLP but human distinguishability between classes should not be taken for granted. There are problems with unclassifiability and granularity in sentiment analysis in general, and noisy sentiment labels for this dataset in particular, because review level ratings are used in place of non-existent sentiment ratings. The annotator agreement and sentiment classification results reported in Table 4.4 corroborate these challenges.

Total Sentences	Annotators Agreed	Positive or Negative
3000	2369	1933
100%	78.97%	64.43%

Table 4.4: Human workers from Amazon Mechanical Turk classify randomly selected sentences from the Yelp 2018 open dataset. Two humans label each sentence totaling 6000 tasks. Only 64.43% of sentences are classified as positive or negative by both annotators.

	Human			Automatic	
	Con	Gra	Sen	BLEU	Sen(%)
CrossAlignment	2.72	2.81	3.00	9.50	70.56
DeleteAndRetrieve	3.13	3.16	3.25	18.13	78.64
IterativeAlignment	**3.17**	**3.97**	**3.50**	**21.63**	**86.10**
Human	3.70	4.09	3.81	52.63	80.14

Table 4.5: Human evaluation and automatic evaluation results. Human ratings on content preservation (Con), grammaticality (Gra), and Sentiment correction (Sen) are on a 1 to 5 scale. Results are averaged across three human annotators and again across 400 test sentences, 200 positive and 200 negative. [1]

4.1.2 Yelp Sentiment Evaluation

Evaluation work for this task was performed in conjunction with the authors of "Iterative Matching and Translation for Non-Parallel Style Transfer." We hire workers on Amazon Mechanical Turk to evaluate and compare model outputs in terms of content preservation, grammaticality and sentiment correctness. Each task consists of the outputs of all three models and one random human reference to be rated together. Model outputs are combined into one evaluation task because judgments are not precise and giving judges more context facilitates their understanding of how the ratings generalize to new sentences. The three judges' scores on a 1-5 Likert scale are then averaged, and the summary stats average across all sentences in the test set and reported in Table 4.5.

There are several important trends to note. First, our automatic measures *do* correlate with human evaluations, although with some caveats. This suggests the feasibility

of using BLEU and Classifier correctness judgments as tools for rapidly evaluating and iterating on model architectures, as is done in machine translation, at least in the **sentiment transfer** task setting. Caveats include the over-inflated classifier judgments discussed in Section 3.4.1 and susceptibility of BLEU scores to misjudging novel but correct transfer outputs.

Human performance on all metrics is relatively low, considering that successful results are almost always rated 4 or 5, empirically speaking. A low rating on human content preservation corroborates the problematic distinction between content and style in the sentiment transfer task. It also justifies the possibility of many inputs $x \in$ Yelp being nearly unclassifiable or untransferrable resulting in low quality outputs with poor content preservation.

Somewhat low sentiment transfer success can also be explained by the problematic task definition, specifically issues with granularity and distinguishability. If humans cannot distinguish well between positive and negative sentences at the sentence level they may not know how to transfer the sentence, or what target sentiment to transfer to. This lack of distinguishability often occurs because sentiment is most clearly expressed at the word and phrase level as discussed in section 3.2.2.

Table 4.6 presents two sentences from the test set and outputs from all models considered. DeleteAndRetrieve has trouble likely because of the overly-simplifying constraints that delete and substitute paradigm places on the model [1]. Further we notice that the human output for the first example changes "simple" to "complicated." These words have the correct connotations of "positive" and "negative" in most contexts, but they arguably represent content. While the same entree can be reviewed positively or negatively by a reviewer, a simple menu may be "plain" or "uninteresting" or "very limited" but unlikely "complex" in the eyes of another reviewer. Indeed the Iterative Alginment model's success here highlights the possibility that state-of-the-art models are not entirely handicapped by a noisy corpus, but

Input	While **the menu is *simple***, what it does offer is ***truly first-rate***.
CrossAlignment	While the owner is, when i can't expect to speak.
DeleteAndRetrieve	While **the menu**, what there is no hand has would truly ethnic.
IterativeAlignment	**The menu is *very limited*** and **the food is *o.k.***
Human	The menu is ***complicated***, ***very low quality*** items.
Input	The ***surly*** **older waitress** was ***a huge bummer***.
CrossAlignment	The burgers of they was a huge.
DeleteAndRetrieve	The spanish was ***very nice*** and **the older waitress** was ***a huge bummer***.
IterativeAlignment	**The older lady is *nice* too.**
Human	The older waitress was *a real sweetheart*.

Table 4.6: Example outputs of different systems: *CrossAlignment*, *DeleteAndRetrieve*, *IterativeAlignment*, and human reference. Whether or not descriptions like "complicated" fall into content or style is subjective. [1]

human performance and proper evaluation is nevertheless hindering advancement.

4.2 News Style Transfer

In the hopes of constructing a task and dataset with a more intuitive notion of style in natural language, I test and analyze a news corpus with data queried from Webhoseio archives [18]. I collected all posts from The Wall Street Journal (WSJ) and Breitbart from December 2016 to March 2018 that contain at least one designated keyword. The keywords are chosen in the hopes of limiting the difference in content between the two sources. Overall, the much longer sentence lengths and large disjoint vocabulary were too challenging for state of the art models, despite trying to control for both.

4.2.1 News Task Analysis

As with Yelp, I begin by analyzing the lexicon ranked according to saliency. This exercise is especially revealing for the news dataset: most of the salient words in both datasets have to do with content rather than style. WSJ talks more about businesses and products; Breitbart apparently talks more about 2nd amendment rights and

WSJ	Breitbart
('deloitte', 287.0)	('jong-un', 225.0)
('cfos', 231.0)	('idf', 138.0)
('corp.', 166.0)	('amador', 135.0)
('cio', 128.0)	('aliens', 120.0)
('pts', 128.0)	('amnesty', 112.7)
('iphone', 100.0)	('jarrar', 106.0)
('wpp', 89.0)	('2nd', 102.0)
('xerox', 84.0)	('pre-viral', 96.0)
('cfo', 82.8)	('riots', 87.0)
('subscriber', 82.0)	('boko', 84.0)

Table 4.7: The most salient words from the Wall Street Journal (WSJ) and Breitbart corpora. Most or all words reflect differences in content across the two corpora, as salient words in one corpus do not have appropriate synonyms or antonyms in the other corpus.

uses words like "riots" and "aliens". If the WSJ most-salient words included like "protests" and "immigrants" that would indicate the two corpora *do* talk about the same content, but use different words—exactly what makes a good style transfer dataset. The fact that we *do not* see many related words across the saliency-sorted vocabulary indicates that models will have to be very good at overcoming this bias, something state-of-the-art has not necessarily achieved yet..

Rationales Because of the disjoint set of content-bearing words, the rationale extractor and classifier would be susceptible to relying on everything but the hypothesized stylistic markers. For this reason, I test rationale extraction on a smaller news dataset that includes TF-IDF matches between the news sources. Specifically, all of the sentences in the smaller sub-corpora are chosen, and only sentences from the other corpora with TF-IDF vectors within a specified cosine distance of a chosen sentence are kept [2].

Unfortunately results on classifier prediction and rationale extraction were poor. One explanation for these results is the difficulty of the task, which humans failed to do

[2]Zhi-Jing Jin helped construct this filtered dataset using code developed for [1]

CNN	Breitbart
('–', 27103)	('articleslike', 1590)
('unfolds', 3089)	('___', 168)
('caption', 1526)	('martel', 128)
('\xc2 \xa9', 1223)	('coahuila', 68)
('hkt', 933)	('p.s.', 68)
('¨', 907)	('regnery', 67)
('tech30', 741)	('isd', 63)
('indices', 645)	('arce', 61)
("ś", 628)	('pmf', 60)
('slams', 599)	('cheap-labor', 59)

Table 4.8: Because of the disjoint vocabulary between WSJ and Breitbart, CNN and Breitbart were also compared. There are some improvements, such as lower raw saliency values for Breitbart suggesting a less unique lexicon; CNN on the other hand still has noisy salient vocabulary items.

Total Sentences	Agreed (0)	Agreed (1)	Agreed Total	Matched Truth
1000	156	121	277	169
100%	15.6%	12.1%	27.7%	16.9%

Table 4.9: Human workers from Amazon Mechanical Turk classify 500 randomly selected sentences as published by CNN (Source 0) or Breitbart (Source 1). Two humans label each sentence totaling 1100 tasks. Only 27.7% of sentences are classified the same by both annotators suggesting the task is difficult and likely not a good candidate for style transfer.

well as reported in the next section.

Human Distinguishability Human judges was asked to predict which news source a given sentence came from.

4.2.2 News Evaluation

In line with expectations, given the task difficulty and large set of disjoint vocabulary, the most performant transfer model [1] did not produce any meaningful results. The data was substantially filtered by title keywords, and further filtered by only including sentences from one source that had a TF-IDF match in the other source. Even with

this additional filtering, the task did not involve a cohesive style transfer task that the model could learn to perform. Perhaps a bigger blow to the possible news outlet style transfer task is the human performance on distinguishability. Without reliable human evaluation, it would not be possible to judge the success of any models regardless of how well the model might be performing.

4.3 Error Correction in Yelp

Because of the problems with sentiment transfer and news source transfer, the field still needs a more reliable baseline task to test style transfer models; ideally this task will have a clear definition of success that does not muddle content and style, and will stylistically distinguishable to humans. With all of these criteria in mind I propose a task for future exploration and provide some initial analysis here: Non-parallel error correction. Traditionally in error correction, supervised data pairs are assumed. Contemporary work like Rei et al. uses artificial error generation to augment the limited training data. Even with such techniques, the field primarily works with relatively small highly annotated training pairs like the English Learner's Corpus, e.g. UD English-ESL / Treebank of Learner English (TLE) [3, 20]. In practice, one might have access to a large amount of data filled with typos, or similar defects, and another separate exemplary dataset, but no one-to-one matching nor expensive annotations of the error-prone data.

Properties of Non-Parallel Error Correction With that motivation in mind, the non-parallel error correction task is attractive, but is it well-formed as a style transfer task? Here, I consider this task in the same way Yelp and news sources were considered, but to some degree this analysis is not as essential. First, it is clear that every sentence will fit into either the correct or not-correct "styles," with the added benefit that one can decide on flexible rules for what constitutes an error in this context, so long as the dataset is constructed well. Every sentence that has some error can be corrected (with the exception of some small minority of cases where

the error-laden text is nonsensical), and every sentence that is well-formed can have errors introduced. Thus, the task has no issues with classifiability nor transferability. Finally, granularity is not an issue so long as the sentence is considered a part of the "Error" style as long as there is an error anywhere in the sentence. Although errors often occur in relation to a single word or phrase, the presence of well-formed phrases does not compete with the presence of an error elsewhere; in sentiment classification and transfer, the presence of a negative clause may directly compete with a positive sentiment in another clause muddying the ultimate classification and transfer solution.

Setup Given that a large non-parallel dataset of error corrections is not publicly available, I artificially introduce grammatical usage errors and typos to the Yelp dataset using a very simple pattern matching procedure.

4.3.1 Error Correction Task Analysis

To verify that issues with divergent lexicon do not become an issue, and to provide a baseline level for saliency values in cases where the vocabulary is mostly overlapping, I again provide the saliency-ranked vocabulary below, in Table 4.10. Because the sentences are sampled from the same origin corpus, the saliency values here are much lower than in the other subtasks. This definition of style is hence not based on word choice (with the exception of common mistakes skewing the term frequency distribution slightly). The most salient words in the correct category ("whom" and "an") make sense given that the errors were artificially generated, and mistakes like "who" vs "whom" are common. The odd selection of words salient to the "Error" set may be partially due to the fact that the error-generation procedure only works on sentences that have a matching pattern, and this can introduce a bias; crucially the raw saliency values here are still relatively low.

The most desirable aspects of the error correction task are those related to evaluation. In terms of human evaluations, the task is both easier more straightforward: humans need only evaluate content preservation and overall fluency, as the latter represents both attribute and style correctness. Further, the distinction between con-

Correct	Errors
('I', 22620)	('abordable', 55)
('whom', 35)	('zart', 45)
('an', 15)	('sowas', 41)
("'italian", 14)	('echte', 39)
('honest-to-god', 14)	('weinauswahl', 39)
('limitword', 14)	('crois', 39)
('underdeveloped', 13)	('parfaitement', 36)
('omlett', 12)	('flott', 36)
('flattery', 12)	('vaut', 36)
('marbles', 12)	('abordables', 35)

Table 4.10: Most salient words across correct and incorrect styles. The salience of words like "whom" occur because the artificially generated typo corpus frequently replaces "whom" with incorrectly used "who." Most saliency scores are low, which makes sense because both sub-corpora sample from the same underlying Yelp dataset with only the possible addition of artificial errors.

tent and style is very well defined and simple for humans to recognize, increasing the likelihood of consistent evaluation results. Finally, automatic evaluation methods are likely to work well because there is usually a very small number of possible solutions; normally BLEU suffers as a reliable metric because it cannot capture diverse solution spaces.

Chapter 5

Conclusions and Contributions

Overall, I have identified the most popular evaluation techniques in style transfer in natural language, identified key problems with those approaches, and extracted important principles and analysis techniques from those problems. Finally I successfully applied all of these techniques to the most popular style transfer task, sentiment transfer, and concluded with possible future alternative tasks.

Analyzing the State of the Art The first objective of this thesis is to survey the methods and evaluation techniques currently used in style transfer today to establish how well the proposed tasks are defined and how well they allow researchers to evaluate and compare different models. Many interesting and complex treatments of "style" in natural language have actually relied heavily on what most humans would consider "content." Although the distinction is somewhat arbitrary, a good intuitive definition of the two facilitate reliable human evaluations downstream. A simpler but very popular task used in every contemporary study to date is sentiment transfer. Because of its prevalence I focused on it extensively, and formulate a number of principles that articulate the problems with sentiment transfer as a task.

Important Properties and Tools for Analysis I identified several major pitfalls in the sentiment transfer task including the difficult distinction between content and style, the phrase-level granularity (rather than sentence level), and presence of

unclassifiable and untransferable cases. In order to detect when those properties may or may not pose a challenge to models performing the task and humans evaluating it, I introduced a few simple tools for inspecting a dataset, including vocabulary rankings, automatic classifier predictions and human performance on the classification and transfer tasks.

Results Most of the identified problems with the sentiment transfer task were corroborated by the results. The presence of words like "refund" and "cockroach" illustrated the blurry line between style and content in terms of vocabulary, while poor human results on the transfer task—3.70 in content preservation—solidified this challenge as a real issue. The limited percentage of Yelp sentences that both annotators identified as positive or negative suggests that unclassifiability interferes to some extent; the low overall human performance on transfer suggests that there are indeed untransferrable cases like the anectdotes used in section 3.2.3

Automatic measures proved to correlate reasonably well in the sentiment transfer case, demonstrating that with a properly defined task, quick and acceptable evaluations of a model can be readily produced via BLEU scores and automatic classification. It should be noted that BLEU's effectiveness is limited as the overall quality of results increases, so as the current state of the art improves this evaluation technique may become less meaningful.

Finally, a new possible baseline task was proposed: non-parallel error correction. This task would be especially easy to identify and judge. The success metrics would be simpler, as "target attribute correctness" and "grammaticality" can be collapsed to a single overall fluency judgment. The task inherently has most of the desirable properties outlined in this thesis, and limited biases like disjoint content-bearing vocabulary sets. The best model to date [1] has demonstrated exceptional proficiency in grammaticality success, so it is likely that it will succeed in this new task setting.

Appendix A

Tables

Correct	Substituted Error
worst	worse
might have	might of
you 're	your
their	there
too	to
whose	who 's
whom	who
used to	use to

Table A.1: A selection of the types of typos and mistakes introduced to the Yelp dataset for the Error Correction task.

NRA	gun	guns	firearm	firearms
concealed carry	national rifle association	abortion	planned parenthood	pro-life
pro-choice	trump	russia	russian	russians
putin	tax	taxes	tariff	tariffs
immigration	immigrant	immigrants	migrant	border wall
american	americans	deport	deported	democrat
democrats	democratic	republican	republicans	FBI
CIA	china	military	korea	white house
congress	senate	GOP	pentagon	supreme court
senator	representative	conservatives	government	sanctions
feds	politics	taxes	legislation	campaign
tax	campaign	lobbying	union	unions
trump	president	obama	house of representatives	legalize

Table A.2: Keywords used in filtering Webhose data to reduce the amount of divergent content. All queries are lowercased because the dataset is tokenized and lowercased before running queries; 3-letter organizations are left capitalized for readability.

Bibliography

[1] Iterative matching and translation for non-parallel style transfer, 2018.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL `http://arxiv.org/abs/1409.0473`.

[3] Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. Universal dependencies for learner english. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 737–746. Association for Computational Linguistics, 2016. URL `http://www.aclweb.org/anthology/P16-1070`.

[4] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. *CoRR*, abs/1711.06861, 2017. URL `http://arxiv.org/abs/1711.06861`.

[5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *ArXiv e-prints*, June 2014.

[6] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567, 2018. URL `http://arxiv.org/abs/1803.05567`.

[7] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. Controllable text generation. *CoRR*, abs/1703.00955, 2017. URL `http://arxiv.org/abs/1703.00955`.

[8] Yoon Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014. URL `http://arxiv.org/abs/1408.5882`.

[9] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043, 2017. URL `http://arxiv.org/abs/1711.00043`.

[10] Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. Rationalizing neural predictions. *CoRR*, abs/1606.04155, 2016. URL `http://arxiv.org/abs/1606.04155`.

[11] J. Li, R. Jia, H. He, and P. Liang. Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer. *ArXiv e-prints*, April 2018.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://doi.org/10.3115/1073083.1073135`.

[13] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

[14] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W Black. Style Transfer Through Back-Translation. *ArXiv e-prints*, April 2018.

[15] Alec Radford, Rafal Józefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment. *CoRR*, abs/1704.01444, 2017. URL `http://arxiv.org/abs/1704.01444`.

[16] Marek Rei, Mariano Felice, Zheng Yuan, and Ted Briscoe. Artificial error generation with machine translation and syntactic patterns. *CoRR*, abs/1707.05236, 2017. URL `http://arxiv.org/abs/1707.05236`.

[17] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. Style transfer from non-parallel text by cross-alignment. *CoRR*, abs/1705.09655, 2017. URL `http://arxiv.org/abs/1705.09655`.

[18] Webhose.io. `https://webhose.io`.

[19] Adam Yala. Text nn. `https://github.com/yala/text`$_n n$, 2018.

[20] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 180–189. Association for Computational Linguistics, 2011.

[21] Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. Adversarially regularized autoencoders for generating discrete structures. *CoRR*, abs/1706.04223, 2017. URL `http://arxiv.org/abs/1706.04223`.

[22] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017. URL `http://arxiv.org/abs/1703.10593`.