# Neural Style Transfer for Natural Language

**Vinitra Swamy, Vasilis Oikonomou**

This is a report detailing our mid-semester research progress in the task of *Neural Style Transfer for Natural Language*. We have conducted extensive research with the goal of understanding this emerging field and developing an execution/evaluation strategy that can help us set up our experimentation. We start by overviewing the problem of Neural Style Transfer, and proceed further with a general literature review in our problem space, followed by our research plan, our datasets, and a second literature review on evaluation strategies.

## Overview

The way we have formulated the problem so far is as follows. Given an input sequence (that could be a sentence or even a paragraph) and a target "style", we want to generate an output sequence. This output sequence, while retaining important information presented in the input sequence, "rewrites" the input sequence in the target "style".

Our investigation so far has focused on two existing areas of research, Neural Machine Translation and Text Summarization. However, one core difference that we have noticed between the task at hand and the aforementioned ones is the absence of labeled data. This has forced us to broaden the scope of our research in areas such as language modeling and think about ways we can formulate the problem as an unsupervised learning one.

## Literature Review

The paper that we found to be most closely aligned to our research goal of style transfer for non-parallel text is Style Transfer from Non-Parallel Text by Cross-Alignment by Shen et. al. The MIT CSAIL team formulates the technical challenge of neural style transfer as separating the content from text characteristics using a variational autoencoder, a discrimination network, and a decoding RNN. The first step in the process is identifying a content representation for a given sentence, which Shen et. al. attempt by minimizing the reconstruction error for a sentence (condensing the representation and then expanding it back to its original form with the highest possible accuracy). The problem with this approach is that the autoencoder might learn style features, and we want to guarantee it only learns content.

The solution proposed by the CSAIL team is a discriminator neural network that takes the latent variable z from the autoencoder and attempts to predict the initial sentence style it was extracted from. If our discriminator network is good at identifying the input style, then we know that z encodes some sort of stylistic representation, which was not our intention. Therefore, we jointly optimize over the variational autoencoder and the discriminator network such that the the latent variable z is effective in minimizing reconstruction error as well as maximizing loss for the discriminator network, and the discriminator network has the highest accuracy it can obtain.

How do we jump from the latent z content variable to stylized output? CSAIL uses a third neural network (decoder RNN) that outputs content in the target style given z and the target style as input parameters.

While the process and the methodology is sound, we are uncomfortable with their data and evaluation strategy. The input data for the CSAIL formulation is a Yelp dataset that has reviews with classified sentiment -- this approach makes a leap in the problem space by equating sentiment with style. Given this assumption, the evaluation strategy used in the paper is a pretrained sentiment classifier that takes the newly generated reviews and classifies the sentiment for that sentence in comparison to the “transfer style” passed into the decoder RNN.

There has been more work in text generation that could add some insight to our pursuit. In particular, work by Hu et al on Controlled Text Generation. Their approach is fairly similar to the one we examined above. Hu is using Variational Autoencoders and holistic attribute discriminators for effective imposition of semantic structures. This means that they can disentangle certain features of the learned latent space and control certain aspects of the text generation process. An example they refer to has to do with controlled binary sentiment (positive or negative). This paper is particularly interesting because it attempts to tackle natural language as a discrete problem as opposed to image generation that occurs in a continuous pixel space. Perhaps the idea of a latent (and continuous) space representation could be useful moving forward. Some more in depth investigation of VAEs needs to happen based on readings such as Carl Doersh's tutorial on variational autoencoders..

Slightly less relevant, but interesting papers

Our initial inspiration for taking on this task is the application of Neural Style Transfer to images as presented by Gatys et al. Most approaches we’ve seen in this space (separate the task into a content network and style network, learning the content vectors for a certain image, and then running it through a learned style network to output an image in a specific style.

In Paraphrasing for Style, Wei Xu et. al. investigate the challenge of paraphrasing Shakespeare into their modern translation counterparts in simple English. This problem addresses the endeavor of comparing Shakespeare to its corresponding sparknotes/enotes translations, providing perfectly aligned data. This enables the authors to formulate the problem as an NMT approach and use BLEU scores as an evaluation metric. However, we unfortunately are trying to solve the problem of non-parallel text, so this paper only serves as idea inspiration.

**Research Plan**

Our initial (rather ambitious thought) was to develop a distinctively novel approach to address our research undertaking. However, we realized that without aligned data or a research funding budget large enough to fund a Mechanical Turk endeavor, we would likely fall short of our lofty goals. A more fruitful idea that we had is to take the MIT CSAIL paper and improve upon the

findings. We see improvement possibilities in replicating results using a more relevant dataset, character level RNNs, and a newly formulated evaluation metric for the unaligned, non-sentiment based task.

## Data

Based on our research, there has been a variety of datasets that have been used for the goal of Style Transfer in Natural Language that vary based on the approach (supervised or unsupervised) that the authors have adopted. In terms of the supervised approaches we have observed, several authors have used collected datasets of (semi)-aligned text. For example, some have collected data for translations between Shakespeare's original work and the English contemporary versions. One dataset that was used in a supervised learning setting is IMDb2 (a dataset of 62,000 movie reviews from 62 different authors) by Serrousi et al. We have already requested (and received) that dataset, and plan to use it for our research.

In terms of unsupervised methods, we believe that retrieving works of authors publicly available from Project Gutenberg should be able to provide varying degrees of stylistic differences. We have already developed scripts for retrieving desired books  after some project details become clearer. It is worth mentioning that Project Gutenberg-based datasets have been the basis for almost all the authorship attribution papers we have read.

## Evaluation

Using the successful paradigm of Neural Style Transfer for Images as our starting point, we have put a lot of effort in exploring existing and formulating new ways of quantifying semantic differences in style and content between our input and target sequences.

### Quantifying Semantic Similarity

The first semantic similarity based method we have considered is using pre-trained word embeddings (Word2vec, GloVe etc.) to generate a "content" vector for the input and output sequences and then using cosine similarity to quantify the semantic differences between the sequences. The reason we liked this method was because it is computationally inexpensive and is insensitive to word order. Also, we have seen that it provides good empirical results on unseen corpuses. For example, Paredes et. al successfully apply this technique on data from LiveJournal. We have considered improving on this idea by using tf-idf to weight the words in the sequence. However, there are certain problems with this approach. One has to do with the fact that, although the two sequences may be similar, it does not guarantee that certain key information from the input sequence will transfer to the output. For example, the sentences, "I fed my cat" and "I fed my dog", are semantically close, however, the difference between cat and dog is very significant and is not fully captured by the distributional hypothesis.

This points to the need for a measure that will provide a more interpretable result by looking at word similarity more closely. Thabet Slimani refers to a collection of similarity metrics based on synonym/hypernym relations as outlined in Princeton's WordNet. This method requires that we either utilize an architecture that implicitly allows us to align the input and output sequences. This method could potentially be used in conjunction with the one outlined above. There is some consideration to be put in terms of the computational effort required as well as the process through which we could identify corresponding words between input and output. Perhaps a scheme similar to the one identified by Li et al could provide a feasible solution.

We have also looked at existing measures like the BLEU and ROUGE scores used in Machine Translation and Text Summarization respectively. Although we have put considerable effort in understanding and thinking of potential modifications, it seem like our ability to use these metrics is conditional on whether we are able to find data for a supervised task.

We have considered working with more advanced methods like compositional semantics as outlined in Chapters 17 and 18 of Martin and Jurafsky and presented by Jacob Andreas as part of his guest lecture. However, this seems like a quite complex process that may not be worth the added complexity to the project.

A final consideration with regards to measuring similarity is utilizing existing seq2seq architectures that already account for content. An approach we felt was particularly relevant was the "coverage vector" presented by Tu et al. The idea comes from the field of MT and modifies the attention mechanism of the network so that the network does not over translate certain words in the input at the expense of others.

Quantifying Stylistic Similarity

In our research, we found input coming from areas the field of authorship attribution to be particularly insightful in terms of the featurization of text. However, one core difference between our sources and our own task is that stylistic comparison takes place primarily on a document level. Our project was initially designed to be a sentence-level mapping. Our research prompted us to consider the possibility of mapping between longer sentences such as multiple sentences or even paragraphs. With this in mind, here is a summary of some metrics we have considered.

One idea that we had was to use perplexity to evaluate the quality of a sentence under a certain language model. As an example, consider the task of "translating" a sentence from modern English to Shakespearean English. Given a corpus of Shakespeare's work, we can clearly learn an associated language model, using a RNN approach. In this case, we could calculate the perplexity of the language model on the generated sequence and use that as a proxy to the difference in style between the target language model and the style of our artificial sequence.

Our survey of the field of author disambiguation showed us that in general, the most discriminative feature is the function words frequencies. We found many papers in which

authors proposed capturing stylistic differences by measuring the similarity between each document's function words frequencies. In particular, we found Marius Popescu's paper on Statistical Similarity Measures for Stylistic Multivariate Analysis to be a very helpful overview of metrics that we can use for this task. Some ways of comparing the frequency vectors include but are not limited to KL divergence and euclidean distance. In our particular case, we think that the best approach should be determined by looking at the quality of the clustering that each metric produces on the data we end up training our model on. The main problem this approach has is that the comparison for stylistic similarity happens on a document level while we intend to find a measure that can work well on a sentence or paragraph level. It could be the case that these are too high variance settings for this approach to be reliable.

On the other hand, Ruder et. al take a different stance towards the problem of authorship attribution that might be more interesting to us. They use a character-level CNN for large scale authorship attribution. In particular, they emphasize that their approach does a really good job at handling multiple classes and, combined with a very short inference time improves on current state-of-the-art by quite a bit. What makes this approach particularly relevant to our problem is that character level cnn capture information about morphemes that other techniques that operate on a word level are blind. This can help us extract more signal from our sequences. An interesting survey of modern techniques for authorship attribution that we used extensively can be found as part of Efstathios Stamatatos's work at the University of the Aegean.

**Next Steps**

After our extensive literature review, our goal is to follow the steps outlined in the "Research Plan" section of this report, as well as discuss possibilities with our friends in NLP Research as well as Professor Bamman. We aim to provide a significant contribution to the problem space of Neural Style Transfer heavily inspired by Shen et. al.'s work.