Unsupervised Style Transfer for Poetry

Anthony Maltsev [1] Michael Young-jin Cho [1]

[1]Department of Computer Science, Stanford

Introduction

Textual style transfer is the problem of preserving meaning while changing textual style. Use cases include: language translation, text simplification, or debiasing offensive texts. For example:

```
Source Sentence (informal):  Come and sit!
Target Sentence (formal):  Please consider taking a seat.
```

We investigate translating authorship for poetry, specifically translating works by Shakespeare into the style of New York School poets like Frank O'Hara and vice versa. There are two primary challenges associated with such tasks:

- There is a lack of parallel data; in other words, given a Shakespearean poem, there is no known "right answer" for how Frank O'Hara would've written it.
- There is a lack of data in general; poets simply don't write that many poems in their lives.

A survey on deep learning for textual style transfer [2] identifies 3 primary approaches for our problem: disentanglement, prototype editing, and pseudo-parallel corpus construction.

We will focus on the Iterative Back-Translation [1] method of pseudo-parallel corpus construction, trying to address the challenges raised in the introduction.

Dataset

We are using subsets of the PoetryFoundationData dataset available on huggingface: one for all avaliable Shakespeare poems and another is all available poems by prominent New York School poets (John Ashbery, Barbara Guest, James Schuyler, Kenneth Koch, and Frank O'Hara).

Around **1000 training sentences** is available for each of the two.

References

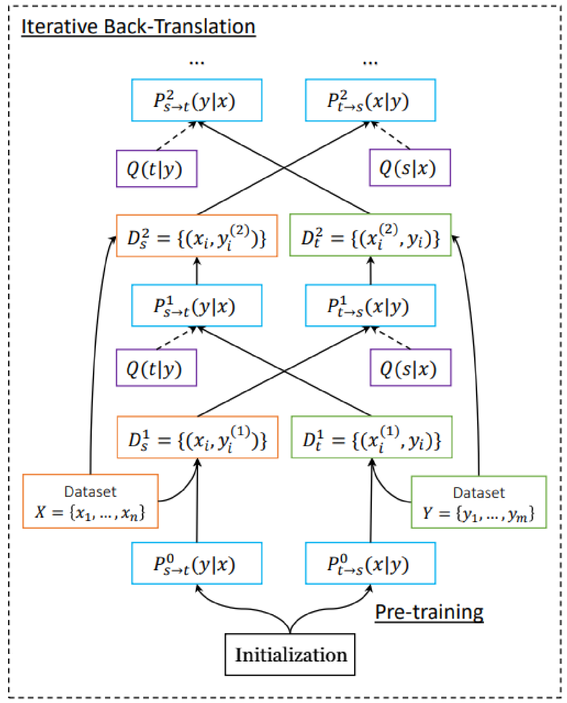
[1] Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In Alexandra Birch, Andrew Finch, Thang Luong, Graham Neubig, and Yusuke Oda, editors, *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[2] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205, 04 2022.

[3] Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. Style transfer as unsupervised machine translation, 2018.

Methods

Iterative Back-Translation creates a synthetic corpus of parallel data, which is much easier to train on than two separate mono-style corpora for each author. We then train on this synthetic data and iterate:



1. We finetuned OpenAI's GPT2 model to do text classification between the two styles as the discriminator model. Then, we introduced an additional loss component based on this discriminator, following [3]; this encourages the translators remain in the right style.
2. We initialize each translator by using general language models: a 220M parameter model based on Gemini and fine-tuned on the poetry corpus, and Google T5-small from huggingface.

Overall, the final loss function was a weighted sum of the discriminator loss and the training loss from training the language model on the synthetic parallel corpus data.

We used the huggingface transformers library to train the initial models based on GPT2 and Gemini as we described above. We implemented the Iterative Back Translation training loop using a custom gradient function in tensorflow.

Experiments and Results

We trained two different types of models: a discriminator and the translators.

Discriminator: very good results in classifying whether a text was written by Shakespeare or by a more modern New York School poet. On a validation set of 100 text examples by either Shakespeare or a New York School poet, the discriminator achieved 100% accuracy.

Translators: unfortunately, bad results. Evaluation was mostly based on qualitatively observing input-output pairs, since it did not seem worthwhile to use quantitative metrics on the resultant outputs.

We tried two different baseline models for the translators: a fine-tuned Gemini and T5-small.

Fine-tuned Gemini: we fine-tuned Gemini to initially output the same text as was inputted. However, even as we trained the IBT procedure, they continued to output the same text as the input:

```
input(NY): Elm Grove, Adcock Corner, Story Book Farm?
output(S): Elm Grove, Adcock Corner, Story Book Farm?
```

T5-small: the results of this training were largely nonsensical as the training did not overcome the intialization of the model (mainly trained on other sequence-to-sequence tasks like text translation).

However, there were some examples that the model was learning some level of translation, for example:

```
input:  to the nose, a resinous thought, to the eye,
a lacquered needle green...

output:  a squeezed green, a resinous thought, to the eye
```

Conclusions

- Our results are likely due to overfitting (for Gemini) and a lack of data to overcome system noise (T5-small).
- We also had an error in calculating gradients for the discriminator loss, causing the gradients to all be zero for this term.
- IBT was an interesting solution to not have a structured parallel corpus.
- Through this project we confirm that the task of style transfer without proper parallel corpora is very challenging.

Big thanks to CS230 staff for the support, AWS credits, and learning!