CS 410 NLP Main Project - A Study of Shakespearean Style Transfer and Text Generation

Darya Haines (darya4@pdx.edu) , Dave Howell (dave9@pdx.edu) , Serena Glick (sglick@pdx.edu) , Sina Bagheri Nezhad (sina5@pdx.edu)

Abstract

The Shakespearean corpus is often used in style transfers that turn modern English into a facsimile of Shakespearean English as well as being used to generate random Shakespearean sentences. Style is defined as patterns of lexical and syntactic choice which are distinct from the content of a sentence, which means having a focus on form while preserving semantics. Style transfer uses an intermediate paraphrasing step to change text from one style to another. The paper used in this project's research to study style transfer used unsupervised learning, which means no paired data exists between any two styles. Shakespearean text generation, a form of Natural Language Generation, uses Gated Recurrent Units, GRUs, to generate text from characters and not words. The input to the model is a text file of Shakespeare's writing that is available online and the output uses a start string as a prompt to generate mock Shakespearean text.

2 Introduction

This group chose to do a project where the main goal was to research different NLP topics that used the Shakespearean corpus and try to implement and improve a preexisting model. Since this corpus is so small and there is not a lot of variety in the models that use it, the group decided to focus on the idea of style transfer from modern English to Shakespearean English and the idea of generating Shakespearean text. The NLP task that this project addresses is the research of style transfer and text generation for the dataset known as the Shakespearean corpus. The main style transfer paper that was examined yielded eleven different style transfers in one paper, not just Shakespearean style transfer. Along with the time it would take to fine-tune that model adding up to forty days, this concept was not chosen as the final model for this project. The final model was chosen to be an improvement upon a Shakespearean text generation that was using GRUs, which are a version of Recurrent Neural Networks (RNN).

3 Related Work

There are many reports and projects that are related to this specific project since Shakespeare's works have been around four hundreds of years and remain a point of study for people in NLP. While there are many reports as well as plenty of research done into style transfer, text generation, and models that use Shakespearean English, not all are relevant enough to discuss in this report. Some of the research done into these topics are no longer viable due to age and modern advances that are incompatible. This report will discuss the research that has been done in the NLP topics of style transfer and text generation as they relate to this project.

3.1 Style Transfer

Giecska and Toshevska (2021) suggested various deep learning techniques that also perform text style transfer. That report discussed a systematic review of style transfer methodologies that use deep learning in order to understand the challenges and opportunities between practical solutions. They based their review around representation learning and sentence generation in a distinctive style while highlighting the similarities and differences between observed proposed solutions. Gangal et al. (2017) wrote about how to use copy-enriched sequence-to-sequence models to change modern English into Shakespearean English. They pretrained embeddings of words by using external dictionaries mapping Shakespearean words to

modern English words. The main part of their research was into automated methods that transform text from modern English into Shakespearean English by using a trainable end-to-end neural model with pointers in order to enable copy action. Their paper was the basis for the supervised Shakespearean style transfer model in this project. Iyyer et al. (2020) reformulated unsupervised style transfer as a paraphrase generation problem. They created a model that changed modern English into eleven distinctive styles, including Shakespearean, romantic poetry, and tweets. Their report also included a survey of twenty-three distinctive style transfer papers in which they show existing metrics and propose fixed variants. This was the main report that this project referenced when discussing style transfer and the feasibility of creating a style transfer model. However, as this report covered eleven distinct styles instead of the one style needed for this project, this report was not implemented and was passed over for a text generation approach.

3.2 Text Generation

Due to the pivot to text generation from style transfer, there was less time to research the topic of text generation. However, there were still plenty of reports to study as this was a well-covered topic in NLP. Ferreira et al. (2019) introduced a systematic comparison between neural pipelines and end-to-end data-to-text for generating text using GRUs and deep learning methods. Their research is similar to the project in this report due to both using GRUs to perform text generation. Their report demonstrated how pipeline models generalize better to unseen inputs and showed how using explicit intermediate steps in the generation process results in better generated texts than texts generated by using end-to-end approaches. Yalcin (2021) discussed how to use NLP to generate artificial Shakespearean text. They discussed RNNs as well as sequential data and how they related to text data in text generation. This report was the main reference this group used when creating the Shakespearean text generation model and is where the corpus that this group used came from.

4 Methodology

The first stage of this project included reading several different reports that covered what constitutes style transfer and language generation for the style of English known as Shakespearean language. These reports were narrowed down to two different reports with code that demonstrated style transfer and Shakespearean text generation. The models that were used were a Shakespearean unsupervised style transfer model that used GPT2 and a Shakespearean text generation model that used GRUs, which are a form of RNN. While the only model that was implemented was the Shakespearean text generation model, both models were heavily researched and the methodology to each approach was the main consideration of this project. The dataset that was used in both of these models was the Shakespearean corpus, which was cleaned and prepared by Karpathy and hosted by the team that created TensorFlow. Most preprocessing steps were taken by that team, so the text generation model simply read in that data and used it as is.

4.1 Shakespearean Style Transfer

The techniques that were used in the Shakespearean style transfer research model include using intermediate paraphrasing to simplify the input, using unsupervised style transfer and GPT2 to train and test the model, as well as using inverse paraphrasing to switch styles. This model took a modern English sentence as an input and would output that sentence translated into Shakespearean English. The research for this model showed an intermediate paraphrasing step that was used to simplify input so it could easily be translated into Shakespearean English. This step was made by using neural machine translation to translate sentences to non-English and then translate back to English. It was also seen that inverse paraphrasing was needed in order to switch the styles. This model would use GPT2, fine-tuned with the ParaNMT-50M paraphrasing dataset, to simplify input and then would use GPT2 with inverse paraphrasing for each individual style to translate the input and produce a translate output. Figure 1 shows a visual representation of the two translation steps that are taken when the model is trained to translate modern English into Shakespearean English as well as what the steps look like during the testing phase.

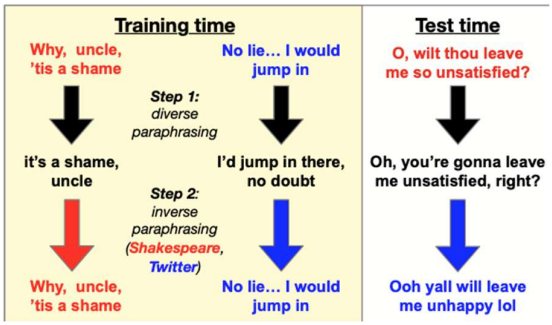

Figure 1: Shakespearean Style Transfer Training and Test Steps. Source: Iyyer et al (2020), Reformulating Unsupervised Style Transfer as Paraphrase Generation

The limitation that prevented this model from being implemented was mainly due to the procured code using eleven diverse styles, not just Shakespearean style, which would have required up to forty days of retraining just to simplify into only performing Shakespearean style transfer. The entirety of this model was gained from research on other papers on this topic as the training time needed to change this model did not fit the time constraints this project had.

In addition to the unsupervised style transfer, we worked on making a supervised style transfer model to implement a version of Shakespearean style transfer. In this model, we used a dataset of 21075 Shakespeare's sentences paired with their modern language paraphrases provided by Jhamtani et al. (2017). In our model, we used the GPT2 and fine-tuned it with our dataset. To train the model, we made a text file that consists of modern language sentences and their related Shakespeare sentence in a line of the text file. We fine-tuned GPT2 in 20 epochs with our dataset to train it to transfer modern language sentences to Shakespearian style.

4.2 Shakespearean Text Generation

The techniques that were used in the text generation model included using GRU to perform machine learning tasks associated with memory and clustering as well as using a dense layer in keras to produce an output of text. This model uses three layers; an embedding layer, a GRU layer, and a dense layer. The embedding layer is the input layer where input values are accepted and converted into vectors. The GRU layer is a RNN layer that is filled with 1,024 gradient descent units. The dense layer is where the result is outputted with outputs sized using the number of unique characters in the vocabulary. The tools used for this model included Python, Google Collab, numpy, tensorflow, and keras. Using Google Collab, the training of thirty epochs took a few minutes and using jupyter notebooks that same training took about an hour and a half. The model was designed using a character-based approach, where the input to the model is a short string containing the name of a character and the output is Shakespearean text containing that character. While the majority of this model came from other papers, the hyperparameters that were used to generate the model, such as temperature, as well as the experiments that were performed on the model to attempt to improve accuracy were implemented by the authors of this paper.

5 Experiments

Since the main goal of this project was to improve the text generation model, several experiments took place to try and improve the quality of the generated text. One experiment was to trim the start and end characters to remove partial words. This experiment was trying to add a preprocessing step that removes any leading and trailing non-space characters and replaces them with spaces. In the end, this experiment showed that avoiding partial words means padding with spaces, which destroys the integrity of the analysis because the extra spaces are considered meaningful. All of the other experiments that took place involved tinkering with the hyper-parameters of the text generation model to determine what configuration would produce the most accurate results. For example, the temperature hyper-parameter was changed multiple times to determine what number would produce the cleanest generated text. While very few of the experiments that took place improved the model, this group gained a better understanding of how the model should work and what the best output might look like.

6 Conclusions

This project was given a time constraint of four weeks for research and implementation. In the end there was not enough time for creating a model from scratch, so most of the code in this project was created by the TensorFlow team as an example on Shakespearean text generation with a few modifications and improvements by the researchers of this project. The open-source code

for the two models in this project can be found at https://github.com/sinaai/shakespearize and the style transfer demo that was created by Iyyer et al. (2020) can be found at http://arkham.cs.umass.edu:8553/.

6.1 Results

The results of this project included a loss of 0.6979 by the thirtieth epoch and a human observed accuracy of about fifty percent in the generated text. This means that the generated Shakespearean text looked accurate from a distance, but when examined was illegible due to misspelled words and incorrect grammar. Further evaluation metrics were unable to be completed in the time constraints given for this project.

6.2 Insights

Some interesting insights that came out of this project were that some reports leave no room for improvement, or if there is room it might be a lengthy process to improve the code. For example, the reason that this group moved away from the topic of Shakespearean style transfer is because the report that was being used had very little room for improvement and it would take forty days to train the model with any edits. Another interesting insight that came from this project was that using the limited Shakespearean corpus will mean that the model will be limited in its ability to learn and generate text that is legible and makes sense. When trying the experiment to trim the start and end characters to remove partial words, an interesting error occurred where the output would suddenly add larger spaces then were there before. It was determined that this error was occurring because the model was trying to overcompensate for the removed spaces by adding even more spaces, resulting in a much larger output with less words. While that change did not work as expected, it was still an interesting learning experience.

7 Future Directions

There are several possible directions that this project could take in the future if there had been more time allotted for this project. One possible future direction is to switch the model from character-based generation to word-based generation. This means that instead of giving the model the name of a person to start off the text generation, the model would be given a word prompt from which to generate Shakespearean text. Character-based text generation has some limitations that affect the output such as cutting off words as input which affects the generated output. Another possible future direction could also be taken using sub-words instead of normal words or characters, which makes it possible to encode words that were not even in the training data in order to improve the accuracy of the model. This might help overcome some of the limitations of the smaller Shakespearean corpus. While there was not enough time allotted for this project to make major modifications to the existing Shakespearean text generation model, these were some of the possible directions that were suggested toward the end of the research and modeling process.

8 Ethical Considerations

There are many ethical considerations in any type of project. Any project that's main focus is an NLP task has unique ethical considerations since it uses real data collected from some source in order to train a model that can be used to do a wide variety of tasks that affect the average person. Who benefits from this data and how could it be misused? How was the data collected? What are the models' ethical limitations? All of these questions and more need to be considered in order to evaluate if a model is adhering to a certain code of conduct in order to protect the researchers, enhance the validity of the research that was done, and to make sure that scientific integrity is being maintained. Our project focused on Natural Language Generation, so of course there is a concern regarding the purpose of this generation; one would not want it to be used for malicious or deceptive purposes, especially since it can be difficult if not impossible to tell the difference between human generated and computer-generated texts

8.1 Dataset Considerations

For this type of project, where most of the code and data came from preexisting sources, the biggest ethical considerations come from the source of the data. Was this data collected ethically and how big was the pool of information that the data was collected from? It is well known that the Shakespearean corpus is comprised of all of Shakespeare's plays, so the data was collected as ethically as possible. The issue is from the data that

the corpus was collected from. Since Shakespeare wrote these plays around 1589, they could contain outdated information that could be considered inappropriate and that is an ethical risk that comes from using this corpus. Another ethical issue that stems from the use of this corpus is the size of the corpus. Since the data is limited, any model that uses this data will also be limited in the amount of training and testing it can do with the data. Therefore, any model that uses this corpus will be unable to produce truly accurate results due to the limited nature of its training.

Another issue is the dataset used in the Style Transfer paper, ParaNMT-50M. This dataset was introduced in the 2018 paper: "Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations" by John Wieting and Kevin Gimpel. The authors describe generating the dataset via "translate[ing] the Czech side of a large Czech-English parallel corpus" without specifying the name of that corpus. Furthermore, the authors of the paper make no mention of Ethical Considerations, so we have no way of knowing whether there is any bias, hate speech, or other undesirable elements present in the original corpus and subsequent dataset.

8.2 Misuses and Benefits

There are only a few kinds of potential misuse that might arise from this type of project since it has a limited amount of data and is straightforward in its design. There are also not a lot of benefits from this model since it is a not well used language for anything outside of fun. Some possible misuses could be using this model to create fake plays that are sold as real or by using this model the generate gibberish spam emails that bother people. These tend to be the normal issues with text generation models. As for who will actually benefit from this model and how, most people will only benefit from using this for humorous purposes, generating mini plays with friend's names to give them or to create little fake plays for personal enjoyment.

9 Project Ownership

This group was comprised of four members who split the work evenly so as to ensure that everything got done in a fast and efficient manner. The work was split into three parts: the coding, the presentation, and the report. Each section was then assigned to the members of the group who were most knowledgeable about those sections.

Serena Glick was the point person who was responsible for the group's formation and kept everyone accountable by setting dates and times for group zoom meetings as well as facilitating each group zoom session. She participated in idea formation and research and contributed articles towards our project's implementation details. Finally, she designed the entire Google Slide presentation from scratch and presented it to the class on Thursday December 1st, 2022.

Darya Haines was the main writer of the project report and kept everyone accountable to what the report required. She participated in idea formation and report requirement research as well as helped brainstorm ideas for the presentation. In addition to that, she also actively participated in the group discussions and report discussions that took place during the weekly group zoom meetings. Finally, she also helped present the project's corpus limitations to the class on Thursday December 1st, 2022.

Dave Howell mainly worked on improving the Shakespearean text generation model. He was also a participant in the idea formation and research that took place as well as contributed articles towards our project's implementation details. He was a vocal participant in the weekly group zoom meetings and helped finalize the presentation concepts. Finally, he also helped present the project's technical details to the class on Thursday December 1st, 2022.

Sina Bagheri Nezhad mainly worked on improving the Shakespearean text generation model and the supervised style transfer model as well as researching supervised and unsupervised style transfer. He also participated in idea formation and research and contributed articles towards our project's implementation details. Finally, he also helped present the project's technical details to the class on Thursday December 1st, 2022.

10 References

Orhan Yalcin. 2021. Create Your Own Artificial Shakespeare in 10 Minutes with Natural Language Processing. In *Towards Data Science*.

Thiago Castro Ferreira, Emiel Krahmer, Chris van der Lee, and Emiel van Miltenburg. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures.In *Proceedings of the*

2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). https://doi.org/10.48550/arXiv.1908.09022.

Mohit Iyyer, Kalpesh Krishna, and John Wieting. 2020. Reformulating Unsupervised Style Transfer as Paraphrase Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.18653/v1/2020.emnlp-main.55.

Kevin Gimpel and John Wieting. 2018. ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations. https://doi.org/10.48550/arXiv.1711.05732.

Sonja Gievska, and Martina Toshevska. 2021. A Review of Text Style Transfer using Deep Learning.In *IEEE Transactions on Artificial Intelligence (TAI), (2021)*. https://doi.org/10.48550/arXiv.2109.15144.

Varun Gangal, Eduard Hovy, Harsh Jhamtani, Eric Nyberg. 2017. Shakespearizing Modern Language Using Copy-Enriched Sequence-to-Sequence Models.In *Proceedings of the EMNLP 2017 Workshop on Stylistic Variation*. https://doi.org/10.48550/arXiv.1707.01161.