Shakespearean to Formal Modern English: A Stylized Neural Machine Translation Approach

Avantika Borikar, Sharanya Kamath, Yamini Kashyap, Dipti Lohia
{aborikar,sskamath,ykashyap,dlohia}@cs.umass.edu

University of Massachusetts Amherst

1 Problem statement

The purpose of machine translation is to produce semantically correct translations from one language to another. There are several factors which can be used to evaluate the translation to be 'good'. One such factor is to evaluate the reading level. Archaic texts are hard to understand for modern readers, and direct translation into modern language often results in grammatical mistakes or unprofessional sentence structures. The readability of a text is assessed by the ease with which a reader can understand it. In this project, we are interested in improving the readability of archaic Shakespearean texts by implementing neural machine translation of text followed by style transfer to convert the input into formally correct modern English. We have two constituent models - a Neural Machine Translator and a text style transfer model, and perform multiple experiments to improve the readability metric of the input text.

For the scope of this project, we treat Shakespearean English and Modern English as two different languages, hence all models converting between them will be referred to as translation models. Also, we treat Formal and Informal writing styles as two different styles, hence all models converting between them will be referred to as style-transfer models.

2 What you proposed vs. what you accomplished

Our proposed stylized Neural Machine Translation (NMT) model aims to translate sentences into target language while changing the style of the source sentences and preserving their style-independent content. In our use case, the machine translation occurs from Shakespearean English to Modern English(possibly informal), and style transfer occurs from Informal Modern English to Formal Modern English. By the traditional sequential approach (Cohn and Lapata, 2007), we aim to conduct this in two steps: First step, is to translate the sentence from source language to target language and second step is to apply style transfer on the target translated language sentence.

Rather than bidirectional knowledge transfer between Machine translation model and Style Transfer model, we constructed synthetic data from the output of each of these distinct models and trained more models with different training objectives, which we then analyzed and evaluated. We call these models Stylized NMT and Augmented-Data Stylized NMT respectively.

- ~~Collect and preprocess dataset~~
- ~~Build and train (specific baseline model) on collected dataset and examine its performance~~
- ~~Build and train neural machine translation model(s) (using T5)~~
- ~~Build and train style transfer model (using BART)~~
- *Build and train bidirectional knowledge transfer model*
- ~~Build and train data augmentation model~~
- ~~Calculate readability metrics and BLEU scores, and do comparitive analysis across models~~

3 Related work

RNN based encoder-decoder methods, first introduced by (Cho et al., 2014b), have been the most common way to approach the machine translation problem in the past. (Cho et al., 2014a) uses this method alongside a gated recursive convolutional network (grCNN) for translating French text to English. Even though the grCNN learns the grammatical sentence structure automatically, it faces a bottleneck as the input sentence size increases. (Kabbara and Cheung, 2016) uses an LSTM based encoder-decoder architecture for the style transfer task. This method is capable of automatically capturing stylistic nuances instead of hard-coding the feature vectors. The three criteria they use for evaluation is soundness, coherence, and effectiveness of the output text. This is mainly done by human-evaluation based on Likert ratings. In our approach we plan to use automatic evaluation metrics such as BLEU along with human-evaluation.

A much newer study proposed by (Hu et al., 2020) gives an exhaustive insight into controllable text generation, neural machine translation, and text style transfer. They also explain the major challenge faced in these problem statements - the unavailability of parallel corpus i.e. parallel sentences with same semantic meaning but different styles. (Shen et al., 2017) and (Fu et al., 2018) approach the challenge of non-availability of parallel data by using adversarial networks to separate content and style representations. They also introduce two new metrics for evaluation - content preservation and transfer strength. (Sudhakar et al., 2019) solves the same problem by using Generative Style Transformer (GST) which is part of a larger 'Delete Retrieve Generate' framework.

(Wu et al., 2021a) deals with first translating a source language to a target language and then applying certain style transfers to the output. They perform this stylized neural machine translation (NMT) by implementing bidirectional knowledge transfer and distillation between an NMT model and an informal-to-formal style transfer model. We have followed this paper as a reference for most of our experiments.

Since parallel data is so scarce in the style transfer domain, researchers have come up with various ways of dealing with this. (Zhang et al., 2020) proposes several methods of data augmentation which includes back translation, formality discrimination, and multi-task transfer for pre-training a transformer-based (Vaswani et al., 2017) model. (Shang et al., 2019) combines a small set of parallel data with large amounts of non-parallel data by using a projection function, and projecting into the latent space.

4 Datasets

4.1 Shakespearean texts scraped data

We used the dataset curated by (Xu et al., 2012) for training our NMT model. This dataset is created by crawling the `nfs.sparknotes.com` website and downloading the full texts of Shakespeare's sonnets and plays with parallel modern English translations. More text is scraped from `www.enotes.com` and combined with the sparknotes data after performing appropriate sentence alignment. This preprocessed dataset is available openly at `https://github.com/cocoxu/Shakespeare`.

4.2 GYAFC (Grammarly's Yahoo Answers Formality Corpus)

GYAFC (Rao and Tetreault, 2018) is a dataset containing 110K pairs of informal/formal sentence pairs on topics Entertainment and Music, and Family and Relationships. This dataset has already undergone a few preprocessing steps which include removal of questions, URLs, and too short or too long sentences. We will be using this dataset to train our text style transfer model. We requested Joel Tetreault to provide us with access to this dataset for our research purposes.

4.3 Data preprocessing

We first split the Shakespeare dataset in the ratio 85-15 for training and testing. At runtime, the training dataset is further split in the ratio 80-20 for training and validation. BLEU score and readability metrics have been reported on the held-out test split of the dataset. We did preprocessing of the data before training all our translation models, as we have used T5-small for each of them. The preprocessing involved addition of a prefix particular to our training objectives, as has been proposed in (Raffel et al., 2019).

5 Baseline

We use GPT2 Shakespeare style transfer paraphraser by (Krishna et al., 2020) as our baseline model. This model is trained on the same datasets that we use in our approach - Shakespearean texts (Xu et al., 2012) and GYAFC (Rao and Tetreault, 2018). The approach creates parallel data by inputting text into a paraphraser and applying different style transfers to it to get the output text. An inverse paraphraser is trained to convert these outputs back to the original style

inputs. This inverse paraphraser is the final model which can be used to perform any kind of style transfer. This implementation is available at https://huggingface.co/filco306/gpt2-shakespeare-paraphraser
We used this baseline model to generate Formalized Modern English text from the input Shakespearean text and report its readability scores. We then used these scores to compare the efficacy of our approach.

6 Model and Variable Notations

This section defines the notations that we will use throughout the report.

- θ_F: This model is trained on the GYAFC and aims to formalize the modern English input text that is fed into it.
- A_x: Original Shakespearean text, directly taken from Shakespeare's novels.
- B_x: Ground truth human-written English translation for the corresponding A_x, obtained from https://github.com/cocoxu/Shakespeare.
- θ_T: This model is trained with the objective $(A_x) \rightarrow (B_x)$. Its output is hereafter denoted using b_x.
- b_y: The formalized English text obtained by passing b_x through our formality model θ_F.
- B_y: The formalized ground truth English obtained by passing B_x through our formality model θ_F.
- θ_T': This is a stylized NMT model trained with the objective $(A_x) \rightarrow (B_y)$. It aims to directly convert input Shakespearean text to Formal Modern English. Its output is hereafter denoted using b_y'.
- θ_T'': This is an augmented-data stylized NMT model trained with the objective $(b_x, b_y) \rightarrow (B_y)$. It aims to take intermediate translation and stylized outputs from θ_T and θ_F respectively, to output the formalized modern English equivalent. Its output is hereafter denoted using b_y''.

In summary, the ground truth training data for machine translation are pairs of (A_x, B_x), where x is a style and A/B are two languages. The formality style transfer model is trained on pairs of (B_x, B_y) where y is the target(formal) style.

7 Approach

7.1 Building Blocks

The framework we propose consists of two models: (1) A Text Style Transfer Model and (2) A stylized Neural Machine Translation (NMT) Model. Our stylized NMT model aims to convert sentences from Shakespearean English (source language) to Modern English (target language) while changing the style of the source sentences (informal to formal).

Based on the outputs from these two models, we create synthetic data and train other models.

7.1.1 T5 : Neural Machine Translation Models

T5, an acronym for Text-to-Text Transfer Transformer, has been found to outperform a variety of pretrained models for the task of text generation. Hence, we chose T5 for our translation tasks. We start with the T5-small model pretrained on the "Colossal Clean Crawled Corpus" (C4) (Raffel et al., 2019) for all our translation tasks. The transfer learning utilities of T5 are of utmost benefit to us due to the limited nature of Shakespearean text. We have used T5-small available at https://huggingface.co/t5-small for all three of our translation models, namely $\theta_T, \theta_T', \theta_T''$ described in the previous section.

7.1.2 BART : Formality Style Transfer Model

BERT (Devlin et al., 2018) models use bidirectional encoders. GPT (Radford et al., 2018) models use autoregressive decoders. BART (Lewis et al., 2020) combines both of these i.e. it is a sequence-to-sequence model consisting of a bidirectional encoder followed by an autoregressive decoder.
We follow the approach introduced in (Lai et al., 2021) where the authors use two types of reward mechanisms (style classification reward and BLEU score reward) to improve the performance of formality style transfer task. The advantage of using BART models is that they work well with limited amount of parallel data.
The authors of (Lai et al., 2021) have made their model architecture open source at https://github.com/laihuiyuan/

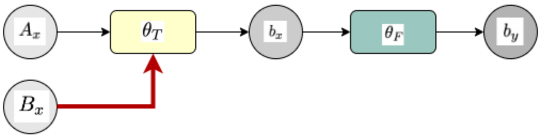

Figure 1: Sequential Model(bold red arrow denotes the passage of target text during training of respective model)

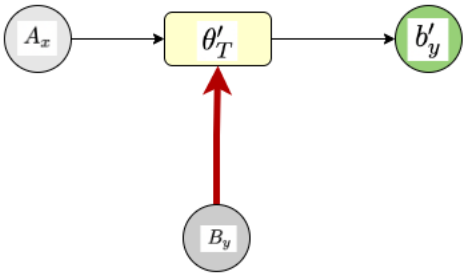

Figure 2: Stylized Machine Translation Model(bold red arrow denotes the passage of target text during training of respective model)

pre-trained-formality-transfer. We pretrain this model on the GYAFC (Rao and Tetreault, 2018) dataset which consists of parallel data for informal to formal English texts of topics Entertainment and Music, and Family and Relationships. This pretraining is necessary because we don't have parallel data for Shakespearean English to formal English.

7.2 Experiments with different models

Examples of training and results can be found in 5.

7.2.1 Sequential Model

Training Objective: In the sequential model architecture (Figure 3) the models θ_T and θ_F, built on top of T5-small and BART respectively, are placed one after another in a sequential manner. The T5 sub-model is trained on (A_x, B_x) as input and target respectively. This T5 sub-model learns to translate Shakespearean text to modern English. But this English is not very readable, as it has just been translated without any kind of refinement or style transfer. Hence its output b_x is hereafter passed through θ_F to convert into more readable formal English b_y. In summary,

$$b_y = \theta_F(\theta_T(A_x))$$

Table 6 shows that the readability metric is higher for b_x and lower for b_y. Since lower metric score means better readability, we can say that formality transfer has improved the readability of Shakespearean text, rather than just translation. Training parameters of these models are described in Table 1 and Table 2.

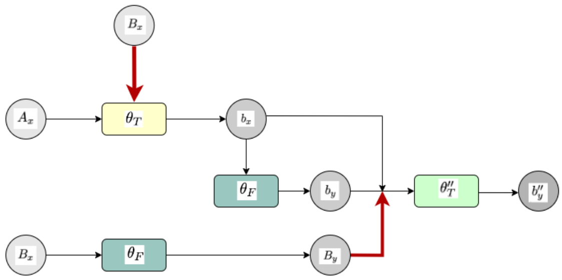

Figure 3: Augmented-Data Stylized NMT Model(bold red arrow denotes the passage of target text during training of respective model)

Name	θ_T
Pretrained model	T5-small
Prefix	*translate Shakespearean English to modern English:*
Learning Rate	2e-5
Epochs	50
Optimizer	AdamW
Weight Decay	0.01
Truncation	128
Batch size	16
Input	b_x
Training Output	B_y
Evaluation Output	b_y

Table 1: Training parameters of translation model

7.2.2 Stylized Machine Translation Model

Training Objective: In this approach, rather than training two separate models, we trained a single T5 model for translating Shakespeare text to formal modern English text. This T5 stylized machine translation model, θ'_T, is trained with A_x as

Name	θ_F
Pretrained model	BART
Learning rate	1e-5
Training steps	1000
Optimizer	Adam
Weight Decay	0.1
Batch Size	32
Dataset	GYAFC

Table 2: Training parameters of Formality Style Transfer Model

Name	θ'_T
Pretrained model	T5-small
Prefix	*translate Shakespearean English to modern formal English:*
Learning rate	5e-5
Epochs	50
Optimizer	AdamW
Weight Decay	0.01
Truncation	128
Batch size	16
Input	A_x
Training Output	B_y
Evaluation Output	b'_y

Table 3: Training parameters of Stylized Machine Translation Model

input and B_y as ground truth. At test time, this model will give output b'_y. In summary,

$$b'_y = \theta'_T(A_x)$$

Table 3 explains the hyperparameters used to train this model.

7.2.3 Augmented-Data Stylized NMT

Training Objective: After evaluating the sequential model and the stylized MT model described above, we thought we could further improve the readability and content-retention of our outputs. Inspired by the bidirectional training objective proposed in Algorithm 1 of (Wu et al., 2021b), we designed a novel training objective which involves training on triplets $(b_x, b_y) \rightarrow (B_y)$. As described in the notations section, $b_x = \theta_T(A_x)$, $b_y = \theta_F(b_x)$, $B_y = \theta_F(B_x)$. b_x and b_y are concatenated together using a new token $< cat >$ for preparing the input of this model. In summary,

$$b''_y = \theta''_T(b_x < cat > b_y)$$

The intuition here is to provide intermediate translation and stylized translation pairs as multi-contextual input to aid the model in generating a perfect stylized translation with minimal content loss. In the method 7.2.2 described above, we noticed that there was significant data and context loss while trying to directly predict from A_x, hence hypothesized that providing intermediate contexts would improve the output quality. This hypothesis turns out to be correct, as shown in 6, evident from the readability scores and BLEU score denoted under b''_y column. Training parameters of this model are described in 4. This model

Name	θ''_T
Pretrained model	T5-small
Prefix	*augmented ShEn to formal En:*
Learning Rate	5e-5
Epochs	50
Optimizer	AdamW
Weight Decay	0.01
Truncation	128
Input	$b_x < cat > b_y$
Training Output	B_y
Evaluation Output	b''_y

Table 4: Training parameters of Augmented Data Stylized NMT

achieves significant content-retention with desirable stylization as shown in the example in table 5.

7.3 Evaluation

7.3.1 BLEU score

Bi-Lingual Evaluation Understudy (BLEU) was proposed by (Papineni et al., 2002) which calculates the similarity between the ground truth reference translations and the machine translated text. It is a position-independent metric that computes the number of token matches between the ground truth and the target text. We will use this metric to evaluate the translations of our model, that performs machine translation from Shakespearean English to modern English. Since our task in neither a pure language translation task nor a pure stylization task, its evaluation is not trivial. There is never a black and white when it comes to translation as it is in case of classification. Multiple equivalent translations may be generated using the building blocks(synonyms/clauses) acceptable in a language but these translations might not show equivalence with the ground truth at word level. Some examples below to justify the same:

1. **Original**: Thus then in brief: The valiant Paris seeks you for his love.
   Human Translation: Well then, I'll say this quickly: the valiant Paris wants you as his bride.
   Model Translation: So, in short, the valiant Paris wants you to marry him.

2. **Original**: Thy dukedom I resign and do entreat Thou pardon me my wrongs.
   Human Translation: I surrender your dukedom and beg you to forgive me all my crimes.

Model	Input	Training Target	Evaluation Output
θ_T	Anon we'll drink a measure The table round.	Soon we will toast around the table.	We'll drink a measure round the table.
θ'_T	Are all thy conquests, glories, triumphs, spoils, Shrunk to this little measure?	Have all your conquests, victories, triumphs, achievements come to so little?	Are all your conquests, glories, triumphs, spoils, all stricken to this little measure?
θ''_T	It's time we threw out the field and assembled ourselves to the ground. <cat> It is time we threw out the field and assembled ourselves to the ground.	It is time that we brought our armies into the field.	It is time that we pulled ourselves out to the field and assembled ourselves to the ground.

Table 5: Sample results from various T5-derived models

> **Model Translation**: I'm leaving your dukedom and begging you to pardon me my wrongs.

In cases like this, the model's BLEU score metric suffers and ceases to be reliable. Hence, we also perform human evaluation on a sample of our translations with the help of our peers.

7.3.2 Readability

Readability is a concept that involves both textual qualities and individual readers' abilities.

In the subject of computational linguistics, readability and text simplification have been extensively investigated, with numerous metrics and methodologies offered in the literature. Global text features such as type/token ratios, lexical consistency, and the proportion of long vs short words are used as indications in common readability measures.

Checking the readability level of the text will tell how well the model has performed the formal stylized translation.

Formalizing the text, requires the translated language to be concise with readable vocabulary and syntactic structure. To evaluate reading level of the translated stylized English language we use Dale-Chall Readability metric (DC), Automated Readability Index (ARI) and Flesch-Kincaid Grade Level score (FK).

The Dale-Chall (DC) readability score is based on a list of 3000 commonly used English terms that captures text lexical information (Chall and Dale, 1995). Words that are not on the list are deemed "difficult." This metric identifies four factors that she believes are important in determining readability: word load, sentence structure, concept richness, and human interest.The percentage of difficult words and the average number of words per phrase are used to calculate the measure, as shown below:(Chall and Dale, 1995)

$$0.1579 \times (\frac{\#difficultwords}{\#words} \times 100) + 0.0496(\frac{\#words}{\#sentences}) \quad (1)$$

Low score indicates readable text by a student of lower grade. Thus, low score indicates better readability.

Automated Readability Index (ARI) is used to measure the understandability of text. It generates an approximation of the US grade level required to grasp the material. The formula for calculating the automated readability index is given below: (Wikipedia contributors, 2021)

$$4.71 \times (\frac{\#characters}{\#words}) + 0.5(\frac{\#words}{\#sentences}) - 21.43 \quad (2)$$

where characters is the number of letters and numbers, words is the number of spaces, and sentences is the number of sentences. A low score implies that the text was read by a student of a lower grade level. As a result, a lower score suggests better readability.

Flesch-Kincaid Grade Level (FK) is used by the US Army for assessing the difficulty of technical manuals. Many other states in the United States utilize the Flesch-Kincaid Grade Level to evaluate other legal documents like as business rules and financial paperwork. This is calculated as below.(Kincaid et al., 1975) Table 6 shows the evaluation metric scores for all the models.

$$0.39 \times (\frac{\#words}{\#sentences}) + 11.8(\frac{\#syllables}{\#words}) - 15.59 \quad (3)$$

7.3.3 Human Evaluation

The most commonly used metric is human evaluation. The second part of our model that performs style transfer and gives a formal touch to the intermediate Modern English text can be evaluated with the help of our peers. We have picked the Act 2, Scene 3 from Merchant of Venice by William Shakespeare and used all our approaches to generate its translations, and compiled the transformations in section A.1 of Appendix.

List of Guidelines given to evaluators:

1. Score each sentence based on three categories on a scale to 1 to 5.

2. Category 1: Grammar - Is the sentence grammatically correct.

3. Category 2: Readability - Is the context of the sentence clear?

4. Category 3: Similarity - How similar is the model generated output from the ground truth.

Five evaluators were asked to rate the sentences and the results are summarised in the table below. The scores obtained from our evaluators can be found in Human Evaluation.

From table 7 we can observe that the average rating provided to the outputs produced by the Augmented Data Stylized NMT is higher that that of the outputs provided by the T5 model in the first stage.

8 Error Analysis and Challenges

In this section, we will talk about the patterns of the errors and the underlying challenge(s) that lead to them. The task of translation is neither equivalent to pure machine translation nor to pure summarization because the vocabularies of Modern English and Shakespearean English differ greatly while simultaneously having a large intersection of words. The challenges that contribute to errors are described in detail here.

8.1 Separate vocabulary

8.1.1 Frequent words from the Shakespearean vocabulary

There are certain words that appear very frequently in Shakespearean vocabulary, such as thee/thy/twas as they are equivalents of commonly used pronouns/conjunctions/prepositions. These words are easily understood by the model and seamlessly translated into their modern English equivalents. In this scenario, the model sometimes behaves like a dictionary mapper. Some examples below:

1. **Original**: Age, thou art shamed!
   Human Translation: Our era should be ashamed!
   Model Translation: Age, you're shamed!

2. **Original**: And do it with all thy heart.
   Human Translation: Then do so, with all your heart.
   Model Translation: And do it with all your heart.

8.1.2 Infrequent words of the Shakespearean vocabulary

Now, there are certain infrequent words in the Shakespearean vocabulary which the model fails to learn because of lack of examples. Example below:

1. **Original**: Benedicite.
   Human Translation: God bless you.
   Model Translation: Benedicite.

8.1.3 Interjections from the Shakespearean vocabulary

The interjections that were prominent in Shakespearean style of writing cannot be easily learned by the model. This is because the same interjection is translated in multiple ways by a human translator based on the context, and this only increases the perplexity of the model for those interjections. Most common example that we came across was the word "Zounds!" Examples below:

1. **Original**: Zounds, who is there?
   Human Translation: Damn it, who's there?
   Model Translation: Zounds, who's there?

2. **Original**: Zounds, sir, you're robbed!
   Human Translation: For God's sake, sir, you've been robbed.
   Model Translation: You're robbed, sir!

From the examples above, it can be inferred that the model sometimes fails to even identify and translate the word "Zounds", or considers it to be a proper noun.

Scores	Baseline	A_x	B_x	B_y	b_x	b_y	b'_y	b''_y
BLEU	0.136	-	-	0.206	0.151	0.191	0.199	0.195
DC	8.458	7.248	8.032	6.665	8.351	6.734	7.190	6.624
ARI	1.991	1.153	0.8590	1.101	0.425	0.733	0.8707	0.484
FK	4.215	3.244	3.213	2.875	2.909	2.531	2.674	2.377

Table 6: Evaluation metric for the outputs obtained by all the models. BLEU scores calculated against B_x

Evaluators	Grammar	Readability	Similarity	Grammar	Readability	Similarity
Output	b_x			b''_y		
Evaluator 1	4.538	4.615	4.307	4.923	4.846	4.538
Evaluator 2	4.307	4.384	4.615	4.923	4.846	4.692
Evaluator 3	4.846	4.846	4.615	4.923	4.923	4.692
Evaluator 4	4.153	4.153	4.230	4.538	4.538	4.692
Evaluator 5	4.384	4.615	4.307	4.923	4.846	4.538
Average Ratings	4.4	4.5	4.4	4.8	4.7	4.6

Table 7: Ratings provided by human evaluators

8.2 Metaphors

When Shakespeare uses metaphors, the model fails to understand what they really stand for and generates messy translations. Example:

1. **Original**: Be the Jacks fair within, the Jills fair without, the carpets laid, and everything in order?
   Human Translation: Are all the cups and glasses in their places, the tablecloths laid out—everything in order?
   Model Translation: Are the Jacks in the house, the Jills in the house, the carpets laid, and everything in order?

2. **Original**: An he speak any thing against me, I'll take him down, an he were lustier than he is, and twenty such Jacks.
   Human Translation: If he says anything against me, I'll humble him, even if he were stronger than he is—and twenty punks like him.
   Model Translation: If he says anything against me, I'll take him down, if he were lustier than he is, and twenty such Jacks.

8.3 Loss of context

We see that the model loses context at times, which a reader is expected to have owing to the previous few lines from the verse, because we are training on sentences chosen from Shakespearean writings rather than whole acts/scenes. Examples below:

1. **Original**: To the gates of Tartar, thou most excellent devil of wit!
   Human Translation: I'd follow you to the gates of Hell, you sneaky little devil!To the gates of Tartar, thou most excellent devil of wit!
   Model Translation: To the gates of Tartar, you excellent devil of wit!

9 Contributions of group members

- Sharanya Kamath:
  - worked on building and training BART model, θ_F, for formality style transfer
  - helped in collecting GYAFC dataset used for training formality style transfer model
  - contributed in report writing and designing model architecture figures
- Yamini Kashyap:
  - worked on building and training the translation model, θ_T and augmented-data stylized NMT model, θ''_T
  - helped in collecting Shakespeare-English parallel dataset.
  - contributed in report writing and designing examples explaining results from different models used in this study.
  - performed error analysis on the results.
- Avantika Borikar:

- worked on T5 stylized translation model, θ'_T
  - helped in evaluating and analyzing readability score for the the output texts from each architecture used in this study.
  - contributed in report writing and analysis of the results obtained.

- Dipti Lohia:
  - evaluated BLEU score for each architecture in this study
  - worked on collecting and analyzing human evaluation for the output texts obtained from each architecture
  - contributed in report writing and analysis of the results obtained.

10 Conclusion

We aimed to generate a translation model to convert Shakespearean text to modern day English language. For this purpose, we used a T5 model. Further, we improved the model by adding a style transfer model to produce a more formal modern text, more understandable by the reader. Hereafter, we evaluated the results and felt like the readability could be further enhanced and then we augmented the data of former models and trained a new model with a novel training objective. We find that the model trained on augmented data triplets, θ''_T outperforms all other approaches on all evaluation metrics. This is because the model now has additional intermediate context that it can retain while simultaneously performing translation and style transfer.

Through our experiments conducted, we learnt that machine translation and style transfer have extensive applications in understanding as well as improving the readability of archaic texts. As a future study, we can extend our approach to produce a style transfer model for changing writing styles across authors.

Another future direction that this study has lies in the usage of models that support processing of long-form texts, such as Longformer (Beltagy et al., 2020). These models might help us retain the context across sentences while training, which is one of the major errors we analysed.

11 Acknowledgement

We would like to thank Prof. Mohit Iyyer and Andrew Drozdov for guiding us, and providing valuable resources and feedback for this project. We would like to thank all the evaluators - Nikhil, Veda, Praneeth, Juie, Chirag for participating in our human evaluation study.

References

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Chall, J. and Dale, E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014a). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cohn, T. and Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 728–735, Prague, Czech Republic. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. (2018). Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Hu, Z., Lee, R. K.-W., Aggarwal, C. C., and Zhang, A. (2020). Text style transfer: A review and experimental evaluation. *arXiv preprint arXiv:2010.12742*.

Kabbara, J. and Cheung, J. C. K. (2016). Stylistic transfer in natural language generation systems using recurrent neural networks. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 43–47.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Krishna, K., Wieting, J., and Iyyer, M. (2020). Reformulating unsupervised style transfer as paraphrase generation. In *Empirical Methods in Natural Language Processing*.

Lai, H., Toral, A., and Nissim, M. (2021). Thank you bart! rewarding pre-trained models improves formality style transfer.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Rao, S. and Tetreault, J. (2018). Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Shang, M., Li, P., Fu, Z., Bing, L., Zhao, D., Shi, S., and Yan, R. (2019). Semi-supervised text style transfer: Cross projection in latent space. *arXiv preprint arXiv:1909.11493*.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Sudhakar, A., Upadhyay, B., and Maheswaran, A. (2019). Transforming delete, retrieve, generate approach for controlled text style transfer. *arXiv preprint arXiv:1908.09368*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wikipedia contributors (2021). Automated readability index — Wikipedia, the free encyclopedia. [Online; accessed 12-May-2022].

Wu, X., Liu, J., Li, X., Xu, J., Chen, Y., Zhang, Y., and Huang, H. (2021a). Improving stylized neural machine translation with iterative dual knowledge transfer. In *Proceedings of the 30th international joint conference on artificial intelligence (IJCAI), Montreal, QC, Canada*, pages 19–27.

Wu, X., Liu, J., Li, X., Xu, J., Chen, Y., Zhang, Y., and Huang, H. (2021b). Improving stylized neural machine translation with iterative dual knowledge transfer. In *IJCAI*.

Xu, W., Ritter, A., Dolan, W. B., Grishman, R., and Cherry, C. (2012). Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914.

Zhang, Y., Ge, T., and Sun, X. (2020). Parallel data augmentation for formality style transfer. *arXiv preprint arXiv:2005.07522*.

Appendix

A.1 - Merchant of Venice Excerpt with translations

The following 5 diagrams are the outputs of our experiments run on Act 2, Scene 3 of Merchant of Venice by William Shakespeare.

I am sorry thou wilt leave my father so.
Our house is hell, and thou, a merry devil, Didst rob it of some taste of tediousness.
And Launcelot, soon at supper shalt thou see Lorenzo, who is thy new master's guest.
Give him this letter.
Adieu!
If a Christian do not play the knave and get thee, I am much deceived.
But adieu.
These foolish drops do something drown my manly spirit.
Adieu.
Farewell, good Launcelot.
Alack, what heinous sin is it in me To be ashamed to be my father's child!
But though I am a daughter to his blood, I am not to his manners.
O Lorenzo, If thou keep promise, I shall end this strife, Become a Christian and thy loving wife.

Figure 4: A_x

I'm sorry you'll leave my father like that.
Our house is hell, and you, a merry devil, robbed it of some tediousness.
And Launcelot, you'll see Lorenzo soon at supper.
Give him this letter.
Goodbye!
If a Christian doesn't play the knave and get you, I'm very wrong.
But goodbye.
These foolish drops drown my manly spirit.
Goodbye.
Goodbye, good Launcelot.
Oh, what a heinous sin I'm having to be ashamed of being my father's child!
But even though I'm a daughter to his blood, I'm not to his manners.
Oh Lorenzo, if you keep your promise, I'll end this fight, become a Christian and your loving wife.

Figure 5: $b_x = \theta_T(A_x)$

I am sorry you will leave my father like that.
Our house is hell, and you, a merry devil, robbed it of some tediousness.
And Launcelot, you will see Lorenzo soon at supper.
Give him this letter.
Goodbye!
If a Christian does not play the knave and get you, I am very wrong.
But goodbye.
These foolish drops drown my manly spirit.
Goodbye.
Goodbye, good Launcelot.
What a heinous sin I am having to be ashamed of being my father's child.
Even though I am a daughter to his blood, I am not to his manners.
If you keep your promise, I will end this fight and become a Christian and your loving wife.

Figure 6: $b_y = \theta_F(b_x)$

I am sorry you will leave my father so.
Our house is hell, and you, a merry devil, have robbed it of some taste of tediousness.
You will soon, at supper, see Lorenzo, who is your new master's guest.
Give him this letter.
Adieu!
If a Christian does not play the knave and you get to, I am very deceived.
But adieu.
These foolish drops drown my manly spirit.
Adieu.
Good Launcelot.
What a heinous sin I am to be ashamed of being my father's child?
I am a daughter to his blood, but I am not to his manners.
If you keep promise, I will end this fight, become a Christian and your loving wife.

Figure 7: $b'_y = \theta'_T(A_x)$

I am sorry you leave my father like that.
Our house is hell, and you, a merry devil, robbed it of some tedious work.
And Launcelot, you will soon be able to see Lorenzo at supper.
Give him this letter.
Farewell!
If a Christian does not play the knave and get you, I am very wrong.
But goodbye.
These foolish drops drown my manly spirit.
Farewell.
Farewell, Launcelot.
What a heinous sinner I am having to be ashamed of being my father's child.
I am a daughter of his, but I do not like his manners.
Oh, Lorenzo, if you keep your promise, I will end this fight, become a Christian and marry you.

Figure 8: $b''_y = \theta''_T(b_x, b_y)$

I am sorry thou wilt leave my father so.
Our house is hell, and thou, a merry devil, Didst rob it of some taste of tediousness.
And Launcelot, soon at supper shalt thou see Lorenzo, who is thy new master's guest. '
Give him this letter.
Adieu!Farewell!"
If a Christian do not play the knave and get thee, I am much deceived.
But adieu. Farewell.Adieu to you.
These foolish drops do something drown my manly spirit.
Adieu. Farewell.
Farewell, good Launcelot.
O, how vile a sin it is in my heart To shame my mother to bear me the name of her father!
But though I am a daughter to his blood, I am not to his manners.
If thou hold'st promise of it, let me end the strife; Become thou Christian, and thine own wife, For this is the cause of this quarrel."

Figure 9: Baseline translation

A.2 - Word Cloud Visualization

Word clouds are attached below to give you an idea of the most common words that we processed in our experiments- from Shakespearean vocabulary, modern English vocabulary, and formalized modern English vocabulary respectively.

Figure 10: Shakespearean text

(a) Model translated text

(b) Translated and formalized text