Style text-to-transfer: Shakesperean Formal English to Modern English style transfer

CSCI 5800: Natural Language Processing and Generative AI – Fall 2024

Project Proposal

Team Member:

Name	Student ID	Email Id	Work division
Anmol Singhal	111021672	Anmol.singhal@ucdenver.edu	Data Prep: Clean and prepare data for training Model Building: Develop and test the model.
Bhavishya Vudatha	111091008	bhavishya.vudatha@ucdenver.edu	Model Refinement: Enhance and fine-tune the model. Data Collection: Gather and process the dataset.
Pooja Kulkarni	111024438	pooja.kulkarni@ucdenver.edu	Evaluation: Apply and analyze performance metrics. Optimization: Improve model efficiency and accuracy.
Sai Krishna Karnam	110942676	saikrishna.karnam@ucdenver.edu	Frontend/Backend: Build user interface and backend services. Testing: Ensure integration and functionality of the system.

Introduction:

In school, many students read classic literature by authors like Shakespeare, Charles Dickens, etc. But understanding the traditional, formal English can be difficult. This makes it hard for modern readers to fully appreciate these texts.

Our project aims to create a system that converts Shakesperean English into modern English sentences. This will help people understand these classic texts more easily while keeping the original meaning and tone. We will use natural language processing (NLP) techniques focusing on style text-to-transfer branch to make this possible. We plan to use both pre-trained models and train our own models to build a small hybrid system for converting these texts.

This project has the potential to grow, with the possibility of making a system that can handle various kinds of formal texts and turn them into modern, easier-to-read languages. Our goal is to make classic literature more relatable and understandable for today's readers. We will also evaluate how correct the conversions are using standard NLP metrics and may add more features if time allows.

The motivation comes from our own experiences as students, struggling to understand complex texts. This project has real-world value by showing how AI and NLP can help make difficult, old language easier to understand.

To conclude, the idea is to create a tool in which you put a sentence in Shakespearean English and our developed model will convert this classic text in modern English. The plan is to train the model to learn this behavior by analyzing the patterns of the style change and then convert text accordingly.

Tools & Methodology:

Our primary platform will be Google Collab as it has the most accessible and powerful computational resources like GPU support. Using this we have planned to **develop our own model**, and use models like sequence-2-sequence, and fine-tune existing mT5 model or some other models to translate classic Shakespearean English text to Modern English text. We have planned to use several libraries for data processing, model training and evaluation like transformers, Py Torch and many more. We will be using Kaggle and web scrapping to develop our dataset collection. We will perform multiple preprocessing steps to prepare the data for training and ensure consistency. This will include tasks such as cleaning the text by removing unnecessary punctuation, converting it to lowercase, eliminating extra spaces, and filtering out non-alphanumeric characters. We will apply word and subword tokenization techniques (like Byte-Pair Encoding) to break the text into manageable units for transformation models. Additionally, spelling normalization will be conducted to replace archaic words with their modern counterparts (like "thou" to "you").

Dataset**:**

- **https://www.kaggle.com/datasets/garnavaurha/shakespearify**
- **https://www.kaggle.com/datasets/pronox/englishtoshakespearedict**

The Shakespeare Style Transfer dataset consists of dialogues from Shakespearean plays paired with their modern English equivalents. Each dialogue has been translated to reflect contemporary language while preserving the meaning and tone of the original text. The dataset captures a wide range of styles, from poetic and formal to conversational, ensuring that the essence of Shakespeare's work is preserved. This dataset aims to support research in style transfer, helping to bridge the gap between classical literature and modern readers. It is offered for research use under a non-commercial license.

We will also be using beautiful soap objects to scrap the data from different web pages. For this we will be finding Shakespeare data from different websites like: Read Modern Hamlet Translation, Scene By Scene (nosweatshakespeare.com). Post collecting the data we will be performing data processing steps.

Goal:

Accuracy of Translation: The translation model should be able to achieve high translation accuracy for input that also contains paraphrases and synonyms.

- **Success Metric**: Achieve a BLEU score of 35 or above on surface-level evaluation. METEOR score between 0.6 and 0.7
- **Process:** Through large and quality datasets, finding good hyperparameters for the model, including mechanisms that handle paraphrases and synonyms, use contextual embeddings.

Efficiency of the Model: Ensure the model translates text efficiently, making it suitable for real-time applications.

- **Success Metric**: Limit processing time to at most 3 seconds per sentence on a basic computing setup.
- **Process:** Pruning, Caching common words and phrases, code optimization

Interactivity and Usability: Develop a website that allows users to input Early English text and receive Modern English translation in real-time. Implement APIs that allow users to perform translation.

- **Process:** Develop a UI using React for Frontend, Flask/ Django for Backend and API handling.

Expected Products:

The style text to transfer project can develop a variety of items. Firstly, it can generate a language transformation tool that can translate formal, historical Shakesperean English into present-day English. This could lead to the development of applications such as an interactive chatbot that helps users in breaking down complex words and phrases into simpler terms. Additionally, an automated text simplification system could be integrated into educational platforms, helping readers or students engage with difficult texts more easily. These innovations could be applied in digital libraries, learning apps, or academic settings.

Moreover, this project could enable users to modernize formal content seamlessly through writing aids and text editors. Such products would be extremely useful in fields like education, content creation, and writing.

References:

1. https://medium.com/@ageitgey/build-your-own-google-translate-quality-machine-translation-system-d7dc274bd476
2. https://arxiv.org/pdf/2010.11934
3. https://www.kaggle.com/datasets/garnavaurha/shakespearify
4. https://www.kaggle.com/datasets/pronox/englishtoshakespearedict
5. Read Modern Hamlet Translation, Scene By Scene (nosweatshakespeare.com)
6. Roudranil/shakespearean-and-modern-english-conversational-dataset · Datasets at Hugging Face