Talk with Shakespeare: A Chatbot with Style Transformation

Manqiu Liu
Georgia Institute of Technology
Atlanta, Georgia, USA
mliu437@gatech.edu

Qian Huang
Georgia Institute of Technology
Atlanta, Georgia, USA
qhuang301@gatech.edu

Venetis Pallikaras
Georgia Institute of Technology
Atlanta, Georgia, USA
vpallikaras3@gatech.edu

Siyan Cai
Georgia Institute of Technology
Atlanta, Georgia, USA
scai70@gatech.edu

Jia Shi
Georgia Institute of Technology
Atlanta, Georgia, USA
jshi351@gatech.edu

ABSTRACT

Chatbots are a popular subfield of research within natural language generation (NLG) which has seen an increasingly wide range of real-life applications. Their success depends not only on an accurate capture of the intended purpose or personality, which determines the content of the generated responses, but also on the linguistic style of those responses. We aim to build a Shakespearean chatbot whose responses are aligned with Shakespearean characters both in content and in style. We divide this task into two parts, one of content and one of style. We will explore several state-of-the-art methods to implement these two parts. For content, we use TransferTransfo to build a chatbot that responds with Shakespearean content; for style, we experiment with two different methods, Seq2Seq with attention as well as the DeleteOnly model, to complete the task of text style transfer. Finally, combining the two parts together, we have a chatbot that responds in Shakespearean content and style which is then evaluated against the baseline using various evaluation metrics.

1 INTRODUCTION

The primary goal of this project is to create a chatbot that speaks in a Shakespearean style. A chatbot – or more formally, a generative data-driven dialogue system - takes a user input and generates a message in response in hopes of mimicking, for example, a given personality. Chatbots have seen a wide variety of applications in recent years, from people's everyday lives to technologically advanced usages. Text style transfer is another field of natural language generation (NLG) closely related to dialogue generation, machine translation, and so on. Unlike chatbot, which is concerned with the content of its generated sentences, text style transfer is concerned with altering specific attributes of the input sentence while keeping the original content unchanged.

Our method combines chatbot and text style transfer to create a chatbot that speaks in a stylized fashion. The first part is a chatbot with specified personalities that generates sentences in response to user input, and the second part is a style transfer model that alters the style attribute of the generated sentences, changing it from modern English into Shakespearean language. Thus, the chatbot is able to carry a conversation with the user and responds in a Shakespearean style. We explore some of the state-of-the-art models for these two separate tasks. After embedding the input using pretrained GloVe embeddings [7] and retrofitted embeddings, we build a chatbot model using TransferTransfo proposed by Wolf et al. [11], a method which effectively improves the shortcomings of existing methods, such as their inconsistent responses and inability to sufficiently leverage dialogue history. Then, we explore two methods for text style transfer. First, Seq2Seq [9] is a widely used method for a myriad of natural language processing tasks, and we incorporate attention structure into our Seq2Seq model to improve its performance. We also employ the DeleteOnly model proposed by Li et al. [4], a neural model that separates an input sentence into its style attributes and the content-only sequence, aiming to alter only the former of the two. Finally, we evaluate our method using various evaluation metrics, including edit distance, BLEU score, and PPL (perplexity). Overall, our method demonstrates competitive performance compared to baseline methods.

2 DATA DESCRIPTION

Our dataset consists of two corpora, one for the original Shakespeare text and the other for its modern English translations. The two corpora have sentence-by-sentence alignment. The parallel data comes from Xu et al.'s 2012 paper "Paraphrasing for Style" (Xu et al., 2012) [12], where the authors scraped modern translations of 17 Shakespeare plays from http://nfs.sparknotes.com, and additional translations of 8 Shakespeare plays from http://enotes.com. After data preprocessing, the SparkNotes data yields 21,079 sentence pairs and the eNotes data yields 13,640 pairs. Xu et al. note that the modern translations from these two sources are qualitatively different; the SparkNotes translations are closer to modern English, while the eNotes translations are more conservative and loyal to the original text. We combined both sources of translations to yield our dataset for training and testing.

3 METHODS

3.1 Model Overview

Our project consists of two parts: the style transformation part and the chatbot part. We first leveraged a pre-trained chatbot model to generate responses given the input messages, then performed style transfer for the returned message from the chatbot model. For the style transformation model, we experimented on two main types of models: the Seq2Seq model with GRU encoder and GRU decoder, and the DeleteOnly model with RNN structure. We also introduced pre-trained embeddings, retrofitted embeddings, and attention mechanics into the baseline model to improve the performance. For the chatbot model, we used TransferTransfo with transformers encoder and decoder layers. We adjusted personality

settings in the model to get stylized responses. We will choose the top two style transfer model and incorporate them into the chatbot model, to finalize our Shakespeare chatbot.

3.2 Embedding Method

3.2.1 Pre-trained GloVe Embedding. GloVe is an unsupervised algorithm [7] to learn word vectors from a word-word co-occurrence matrix. It is proved to have high performance in NLP tasks. As is mentioned in the paper [3], pre-trained embeddings can help the model better understand the training data, which can shorten the training process and improve the model performance. Instead of using random normal embeddings as in the baseline model, we used pre-trained GloVe embeddings with 300 dimensions in our model, which were further used in our training process.

3.2.2 Retrofitted Embedding. A word in different contexts may have different meanings and synonyms, and this has a big impact on our project. The words in Shakespeare's works can have very different usage and meanings compared to modern English. Thus, creating embeddings to incorporate the word similarity pattern, such as the synonym pair of "you" and "thou", between modern English and Shakespeare's work, is important in our training process. We leveraged the retrofitting method [8] to learn word representations on paraphrased contexts. We expect this method to bring improvement into our downstream models.

3.3 Style Transformation Model

3.3.1 Seq2Seq. Seq2Seq models [9] can solve a plethora of problems, such as machine translation, text generation, chatbots (reply-answer problems), text summarization etc. The main building blocks of this architecture are the GRU/LSTM blocks and the embeddings. On a higher level it consists of the encoder and decoder modules. The encoder "reads" the whole input and summarizes into the hidden state; this is the context vector and it contains the information of the whole input text. On the other hand, the decoder, whose initial state is the final state of the encoder, has access to the whole context of the previous sentence. The input of each LSTM/GRU cell is the maximum value of the previous cell output after passing through a softmax layer. In the end, we combine each cell output (max value only) and convert those ids into words. This sentence is the output of our Seq2Seq model, and in our case it is the text after being converted into a Shakespearean one.

3.3.2 Attention. There are two types of Attention [10], and in both cases they tend to describe the inner relations between the two components. In the case of general attention, it shows the relation between the input and output elements, while in the case of self-attention, it shows the relationship between the input elements. The main goal of attention is to solve the issue of vanilla seq2seq models, namely that they can't process long input sentences. In vanilla seq2seq the output of the encoder is passed into the decoder, which means it only "focuses" on the last words/elements, while attention considers (combines) the whole input elements (all hidden states). This allows the decoder to have access to the whole input information and can select the parts that it finds useful.

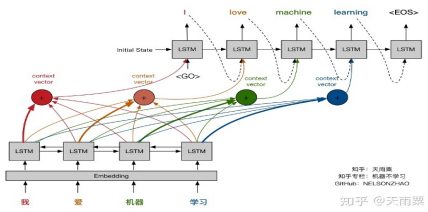

Figure 1: Attention

3.3.3 DeleteOnly Model. In their 2018 paper, Li et al. proposed a set of methods that show substantial improvement on existing approaches to text attribute transfer [4] (Li et al., 2018). Traditionally, adversarial networks are employed to complete the task of altering specific attributes of a sentence such as style, sentiment, and tense. Li et al. observe that we only need to change a few attribute markers — words or phrases indicative of a particular attribute — while leaving the rest of the sentence unchanged. Motivated by this intuition, the authors build four different systems for text attribute transfer: two baselines, "RetrieveOnly" and "TemplateBased", and two neural models, "DeleteOnly" and "DeleteAndRetrieve". The model architecture is shown in the graph below:

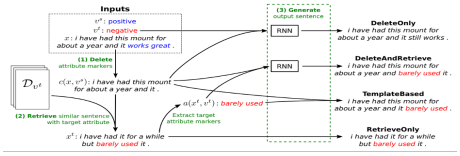

Figure 2: Delete Model Structure

Our project adopts the "DeleteOnly" model. This model removes the source attribute markers and replaces them with the target attribute markers in such a way that it ensures that the original sentence remains fluent. The procedure of "delete" means that the text is separated into a set of attribute markers $a(x, v^{src})$ and a sequence of content words $c(x, v^{src})$. The content words are embedded using an RNN. Next, it concatenates the final hidden state with a learned embedding of the target attribute v^{tgt}. This is then fed into an RNN decoder to generate a new sentence y with the desired target attribute. The training of the model is done by reconstructing the sentences in the training corpus given their content and original attribute value by maximizing

$$L(\theta) = \sum_{(x,v^{src}) \in D} logp(x|c(x, v^{src}), v^{src}; \theta)$$

These systems have been evaluated on three datasets (Yelp reviews, Amazon reviews, and image captions) using various evaluation metrics and their respective performance is as below:

In figure 4, the "Classifier" shows the percentage of sentences labeled as the target attribute by the classifier. BLEU measures content similarity between the output and the human reference.

	Yelp		Captions		Amazon	
	Classifier	BLEU	Classifier	BLEU	Classifier	BLEU
CrossAligned	73.7%	3.1	74.3%	0.1	**74.1%**	0.4
StyleEmbedding	8.7%	**11.8**	54.7%	6.7	43.3%	10.0
MultiDecoder	47.6%	7.1	68.5%	4.6	68.3%	5.0
TemplateBased	81.7%	**11.8**	92.5%	**17.1**	68.7%	**27.1**
RetrieveOnly	**95.4%**	0.4	**95.5%**	0.7	**70.3%**	0.9
DeleteOnly	85.7%	7.5	83.0%	9.0	45.6%	24.6
DeleteAndRetrieve	88.7%	8.4	**96.8%**	7.3	48.0%	22.8

Figure 3: Model Automatic evaluation results

3.4 Chatbot Model: TransferTransfo

The generative data-driven dialogue system (chatbot) algorithm chosen for this project is TransferTransfo [11], which is a state-of-the-art model developed by Wolf et al. from HuggingFace (2019). Previous works in this area, such as the seq2seq model from our course homework, face multiple challenges in generating intelligent conversations. Due to limitations of model architectures, those previous methods produced inconsistent responses and could not capture much of the dialogue history. TransferTransfo focuses on improving these two perspectives and shows good results. Since we have already seen the performance of baseline seq2seq in the homework and in the paper, we chose to directly adopt TransferTransfo over a comparison.

The architecture of TransferTransfo involves an encoder and a decoder model. The encoder used is adopted from the multilayer Transformer of Radford et al., who also provided a basis for the training process. The decoder transformer has 12 layers and masked self-attention heads to account for previous context. TransferTransfo has been pre-trained with the BooksCorpus dataset, which is helpful for learning long-range information by using a document-level corpus over a shuffled sentence-level corpus. Since the model competed in the Conversational Intelligence Challenge 2 (ConvAI2), the fine-tuning step is performed with the PERSONA-CHAT dataset. This allows the model to generate conversations with a consistent personality. For our purpose, we changed the randomized personalities to several options related to Shakespeare's work, including Shakespeare himself, Hamlet, Romeo, and Juliet. By inputting personal descriptions of these characters, we can further create a more immersive and intelligent environment for conversations.

4 EXPERIMENTAL SETTINGS AND EVALUATION

We will use our Shakespeare dataset to perform style transformation task and incorporate our style transfer model with pre-trained chatbot. For style transformation task, we will use BLEU score and edit distance between the predicted transformation sentences and the true sentences to evaluate the transformation performance. For BLEU score's baseline, we will use the unchanged modern sentences as the baseline prediction, and then compare it with BLEU scores from model's output. We use PPL to evaluate the response of chatbot, and finally we will manually evaluate the similarity and performance of the combined Shakespeare chatbot. On our original data, we will perform train/test splits with 100 or 1000 pairs of sentences to be our test data.

Model	Settings(Embeddings/Epoch/Batch Size/Learning Rate/ Layers/Hidden States/Model Specific
Seq2Seq	GloVe(300d)/15/64/0.001/2 layer/128 hidden dimension
Seq2Seq+retrofit	GloVe(300d)/15/64/0.001/2 layer/128 hidden dimension/10 iteration of retrofitting
Seq2Seq+Attention	GloVe(300d)/15/64/0.001/2 layer/128 hidden dimension
Seq2Seq+Attention+retrofit	GloVe(300d)/15/64/0.001/2 layer/128 hidden dimension/10 iteration of retrofitting
DeleteOnly Model	Embedding(64d)/60/32/0.001/2 layers/128 hidden dimension

Table 1: Hyperparameter Settings

4.1 Edit Distance

Edit distance [5] is a metric to quantify the dissimilarity of two strings by counting the minimum number of operations required to transform one string to another. We will allow operations of removal, insertion, and substitution of a character in the string. The edit distance can help us compare the differences between our prediction and the true sentences.

4.2 BLEU

BLEU (BiLingual Evaluation Understudy) [6] is a metric for automatically evaluating machine-translated text. It measures the similarity of machine-translated text to a set of reference translations. BLEU score ranges from 0 to 1. The lower the BLEU score is, the lower the quality of the translation. A value of 1 means there is a perfect overlap with the reference translations.

4.3 PPL

PPL [1] is the Perplexity of gold prediction tokens from the chatbot's next token probability predictions. This metric is used to evaluate the chatbot for the language modeling task.

Hyperparameters settins can be found at table 1.

5 RESULTS

5.1 Style Transformation Evaluation

We can see from Table 2 that all the model's performances are worse than the as-it-is baseline. Among different models, the performances are discussed as follows:

5.1.1 Effect of Embedding. We experiment on random or pretrained embeddings, variable or fixed embeddings, and retrofitted and unrevised pretrained embeddings. We found that using pretrained embeddings in our model could improve the model performance both in metrics of BLEU score and edit distance. As further tuning pretrained embeddings might lead to overfitting, we also observed that fixed pretrained embeddings performed better than unfixed pretrained embeddings. The retrofitted embeddings brought improvement to the BLEU score and edit distance in Seq2Seq model

Model	Embedding	BLEU Score	Edit Distance
Null Model	/	66.93%	30.07
Seq2Seq	Random Initialization (Not fixed)	64.78%	35.10
Seq2Seq	GloVe (fixed)	65.44%	34.91
Seq2Seq	Glove + Retrofit (fixed)	65.74%	32.78
Attention Seq2Seq	Random Initialization (Not fixed)	57.61%	38.43
Attention Seq2Seq	GloVe (fixed)	62.10%	34.11
Attention Seq2Seq	GloVe + Retrofit (fixed)	59.32%	43.94
DeleteOnly Model	/	64.20%	31.78

Table 2: Model Performance

but made the situation worse when it came to the Attention Seq2Seq model.

5.1.2 Effect of Attention Mechanism. We observed that the basic Seq2Seq performed better than the Seq2Seq2 attention model both in metrics of BLEU score and edit distance. We deduced that the attention mechanism can work well on large dataset, but with limited parallel dataset, it can confuse the learning process and bring worse performance.

5.1.3 Comparison between DeleteOnly and Seq2Seq. We found that the DeleteOnly model achieved similar performance with Seq2Seq model but has much lower edit distance. As the DeleteOnly model has large number of parameters, we believed the limited dataset also impaired the model performance.

5.2 Chatbot Model Personality Evaluation

Shakespeare	Hamlet	Romeo	Juliet

>>> Hi.
hi. how are you doing?
>>> Who are you?
i'm william shakespeare.
>>> What do you do for a living?
i write poems.
>>> Where are you from?
i am from england.
>>> Do you believe in love?
i do, but not in my life.
	>>> Hi.
	hello. how are you today?
	>>> Who are you?
	i am a young prince from denmark.
	>>> What do you do for a living?
	i am a student.
	>>> Where are you from?
	i am from the south.
	>>> Do you believe in love?
	i don't think so, but i am also a prince of denmark.
		>>> Hi.
		hello. how are you doing today?
		>>> Who are you?
		i am romeo.
		>>> What do you do for a living?
		i'm a student at the moment.
		>>> Where are you from?
		my parents are catholic and they were killed by the capulets.
		>>> Do you believe in love?
		i'm in love with my girlfriend, and i'm also the son of a pride
			>>> Hi.
			hi! i am juliet. my name is juliet. how are you?
			>>> Who are you?
			i am juliet. i love to read.
			>>> What do you do for a living?
			what do you do for a living?
			>>> Where are you from?
			i'm in new york city.
			>>> Do you believe in love?
			i'm a young girl, a virgin.

Figure 4: Chatbot Personality Comparison Results

For the chatbot's personality feature, we wrote out four different paragraphs as additional inputs about each person or character's distinct information. Then, we asked the chatbots the same questions and found that they were able to pick up the details from the personality and use them intelligently in the responses. Especially for the question about belief in love, each character provided a reasonable answer. This shows that we have created immersive environments for users to generate interesting conversations.

5.3 Shakespearized Chatbot Evaluation

Null Model	Seq2Seq + GloVe	Seq2Seq + GloVe + Retrofit

>>> Hi.
how are you mad?
>>> Who are you?
I am a very boy and a beast sir.
>>> What do you do for a living?
Nay an an worthy mistress.
>>> Where are you from?
Must in the things of these Must?
>>> Do you believe in love?
Nay but a worm is a tale but I am much nothing.
	>>> Hi.
	my lord and my sinews i will do my tent.
	>>> Who are you?
	i am a prince.
	>>> What do you do for a living?
	i am a gentleman.
	>>> Where are you from?
	i am dromio in the tower?
	>>> Do you believe in love?
	i think i do not think i think it is so well.
		>>> Hi.
		what do you do?
		>>> Who are you?
		i am a man and?
		>>> What do you do for a living?
		i am a gentleman.
		>>> Where are you from?
		i am bound in france.
		>>> Do you believe in love?
		i have a brother and i have been a lover and a soldier s eldest.

Figure 5: Shakespearized Chatbot Comparison Results

We chose and compared three typical Shakesperized chatbot results. The three style transfer models are baseline model, Seq2Seq + Glove and Seq2Seq+glove+Retrofit. The results from the baseline model seems not very reasonable, while the Seq2Seq + Glove and Seq2Seq + glove + Retrofit models produce better and more reasonable results. The above two advanced models perform equally well from our human evaluation.

6 DISCUSSION AND FUTURE WORK

From our style transformation outcome, we can see that the performance doesn't match our initial expectations. Based on our literature survey, we have the following hypothesis:

6.1 Limited Parallel Data

We only have about 20k paired sentences for training, which is not enough comparing to the large number of parameters from our models. We observed from papers with similar tasks [2] and found that to achieve a good performance with simple Seq2Seq models, they trained on large datasets with over millions of sentences. To further improve our model's performance, we could include more versions of modern translations to the Shakespeare's work to create a larger dataset.

6.2 Model Complexity

With limited data, simple models such as classic Seq2Seq and attention mechanism that required heavy training cannot learn the pattern well; instead, it might end up causing model overfitting or underfitting. To achieve better performance, we would like to try on models such as pointer models to incorporate the synonym relationships among words.

7 CONCLUSION

In this project we built a Shakespeare chatbot by combining style transformation models with pretrained chatbot model. We experiment on the Shakespeare paired data and evaluated our model using BLEU score, edit distance and PPL score. We found that the pre-trained embedding with retrofitting method could have better performance in style transformation task, and also that with a limited dataset, simpler models such as Seq2Seq achieved better metrics comparing to the model with attention mechanics.

8 CONTRIBUTION

All team members have contributed a similar amount of effort.

REFERENCES

[1] Peter F. Brown, Vincent J. Della Pietra, Robert L. Mercer, Stephen A. Della Pietra, and Jennifer C. Lai. 1992. An Estimate of an Upper Bound for the Entropy of English. *Comput. Linguist.* 18, 1 (mar 1992), 31–40.

[2] Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the Bible. *Royal Society Open Science* 5, 10 (Oct. 2018), 171920. https://doi.org/10.1098/rsos.171920

[3] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing Modern Language Using Copy-Enriched Sequence to Sequence Models. In *Proceedings of the Workshop on Stylistic Variation*. Association for Computational Linguistics, Copenhagen, Denmark, 10–19. https://doi.org/10.18653/v1/W17-4902

[4] Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer. (June 2018), 1865–1874. https://doi.org/10.18653/v1/N18-1169

[5] Frederic P. Miller, Agnes F. Vandome, and John McBrewster. 2009. *Levenshtein Distance: Information Theory, Computer Science, String (Computer Science), String Metric, Damerau?Levenshtein Distance, Spell Checker, Hamming Distance*. Alpha Press.

[6] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*.

[7] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[8] Weijia Shi, Muhao Chen, Pei Zhou, and Kai-Wei Chang. 2019. Retrofitting Contextualized Word Embeddings with Paraphrases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1198–1203. https://doi.org/10.18653/v1/D19-1113

[9] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 3104–3112.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[11] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. (01 2019).

[12] Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for Style. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, Mumbai, India, 2899–2914. https://aclanthology.org/C12-1177