TEXT GENERATION FOR STAR WARS CHARACTER

Keshvi Srivastava
University of Illinois, Chicago
ksriva5@uic.edu

Varunya Yanamadala
University of Illinois, Chicago
vyanam3@uic.edu

Abstract

Natural Language Generation is a subfield of Natural Language Processing. The application of neural networks for movie dialogue generation is still a nascent field. With only a few tested models, there is still much to be discovered. It is difficult to grasp the characteristics of an individual playing a role in a movie, however, movie scripts are a good place to start. An individuals' dialogues tend to capture their distinct vocabulary and speech patterns. The paper attempts to generate dialogues for Star Wars characters - ObiWan Kenobi, Yoda, and C3PO. These individuals differ from one another in the sense - Obiwan is a human, Yoda is an alien and C3PO is a droid. Each has different speech patterns as well vocabularies that the model captures. Along with speech patterns, the model also captures the mood of the seed sentence and continues the generation of text in correspondence to that mood.

1 Introduction

Natural Language Generation can be defined as a process of producing meaningful phrases and sentences in the form of natural language[5]. Our projects aim to divert attention towards generating movie dialogues through the study of past scripts. We use a bi-directional LSTM model to predict the next most likely word in a seed sequence. The input sequence consists of five words spoken by the character understudy from which the model must predict the next most likely five words. The aim of the paper is to identify the distinct vocabularies of each individual while simultaneously picking on the speech patterns. For instance, while Obiwan speaks in human interpretable English, Yoda speaks in reverse. The trained model should be able to identify these differences and generate the dialogue accordingly.

We also study the mood of the input and output dialogues. Since the model learns from movie scripts, it should be able to identify the mood and hence predict the next likely word in correspondence to it. To check how well the generated sentences have taken the mood into consideration we did a mood detection on the generated sentence using a pipeline setup considering NRC, Bing, and Afinn lexicon files.

2 Related Work

2.1 Text Generation

Natural Language Generation (NLG) is a diverse field aimed at generating text based on structured data. This data could be pre-defined sentences or documents. For dialogue or script generation, NLG plays a vital role to obtain logical and semantically sound sentences. Our research shows the use of Markov Chain, Recurrent Neural Networks, LSTM, and GRUs for generating novel sentences[5]. These models are the basic tools for NLG and have been seen to work for script generation with Shakespeare drama and Harry Potter books.

Another technique used in conjunction with the above models is the use of sliding windows to break the input sequence[16,17,18,23]. The input is broken into pre-defined window sizes, with a step size. This helps to collect the input and output training set. The models also take two different approaches for this case - predict individual letters (characters) in the dialogue or the whole word itself.

Our model works with the basic NLG tool (LSTM) with modified input data. We observe some semantically and logically correct dialogues too. The model is described in detail in Section 4.1.

2.2 Mood Detection

Mood detection or sentiment classification has many challenges involved based on the domain of the task. From basic methods or pipelines, which doesn't require any pre-training on the data would require pre-processing to get frequent words and to match these words from the lexicon files to obtain the final sentiment.

Latest techniques in this field include the leverage of deep learning[32,33] concepts to train over labeled data to predict the sentiment or mood of the sentence. But, for this to be achieved we need to have data sufficient enough to train the model with moods given to each of the sentences as labels. These would be classes into which the sentences would get classified to. For a mood detection task like this, there was no previous work done with data labeled.

Hence, the method we incorporated considers a simple pipeline to capture the keywords that get generated from the text generation model to detect the mood conveyed in the sentence. We shall discuss more about this procedure that best worked for the ten-word length sentences generated in 4.2 section.

3 Dataset

In this project, we used Star Wars movie scripts. The data is derived from Kaggle Star Wars Dataset[34] and IMSDB[35] (site to get full movie scripts). The text files from Kaggle Dataset consist of Episode V, VI, and VII where most of the main characters are already filtered out. To add to the dataset, we also extracted movie scripts for the same characters from Episode I, II, and III from the IMSDB HTML Hyperlinks. Together we extracted dialogues of characters Obiwan Kenobi, Yoda, and C3-PO. We have selected these characters specifically for their individual unique features:

- Obiwan - Kenobi : Human character with an English vocabulary. He speaks in the same format as an average human like us.
- Yoda : Alien with an English vocabulary. However, different grammar than what we speak. Yoda speaks in reverse English, meaning that the sentence begins with the subject and ends with pronouns. For example, "afraid are you?" which means "are you afraid?"
- C3-PO : A Droid with artificial intelligence assistance vocabulary. For example, "you must repair him sir if any of my circuits or gears will help ill gladly donate them"

On filtering we got the following dataset:

Character	Number of Dialogues
Obiwan-Kenobi	583
Yoda	168
C3-PO	358

3.1 Data Processing

Working with textual data includes pre-processing to remove all irrelevant characters. When working with dialogues, there are multiple characters used to denote silence and pause which must be removed too. We performed the following pre-processing tasks on the data:

- Remove repeated punctuation denoting continued silence (Example, "...")
- Remove all other punctuation
- Remove all numbers and non-alphabetic characters (Many characters in Star Wars have numeric digits in their names too)
- Reduce all word to lower case
- Split all dialogues into individual words

The pre-processed data is fed to the model. However, there is one additional step which is to convert these words to their integer values.

3.2 Data Tokenization

All neural networks require numerical inputs, for which we convert our tokens into integer values. To do so, a word vocabulary is created for each character and each word in this vocabulary is assigned a unique integer value. We used these integer values to convert our dialogues to a sequence of integer tokens.

3.3 Data Splitting

The tokenized dialogues are now ready to be split into training input and output sets. We observed that the length of the dialogues across the characters can vary between 1 to 137 words each, where an average of 10 words/dialogue is the mode. To create a final dialogue of 10 words, the model had to be trained in a way to predict a minimum of five words; if a five-word input sequence was provided. Hence, we decided to split each dialogue into a sequence of five words where the sixth word in the dialogue would be the required predicted output[23].

Example, an Obiwan dialogue is: **"offhand id say this mission is past the negotiation stage"**, then the input(X) and output(Y) training set would be:

X	Y
offhand id say this mission	is
id say this mission is	past
say this mission is past	the
this mission is past the	negotiation
mission is past the negotiation	stage

This is done to each dialogue and all dialogues are appended to a final input set X. If a dialogue has a length less than five then it is padded with null values. Each of these dialogues is split into individual words and tokenized before being fed to the model.

4 Methodology

The project is split into two steps. The first is the text generation part (explained in Section 4.1) and the next is the mood detection part (explained in Section 4.2).

4.1 Text Generation

Given a seed sentence, the model predicts the next most likely word in the sequence. We make use of a Bi-directional LSTM network that has a higher capacity of holding information over long sequences of data. The model's sequence-to-word architecture allows us to generate a final sequence of the most probable words. The aim of the project is to produce dialogues that are grammatically correct, capture the individuals' speech idiosyncrasies, and also depict the mood of the seed dialogues themselves.

For training, the model is fed a five-word sequence with predicted output as the sixth word. To put high-level concepts in context, each word in the input sequence is also attached to its GloVe embedding. Glove is an unsupervised learning algorithm to learn vector representation. In this project, we have made use of the 300-dimensional GloVe vectors for the input words. The training samples are sequentially fed to the Bi-directional LSTM (Figure 1).

We create a single Bi-directional LSTM layer with 200 neurons. The GlobalMaxPooling1D layer down-samples the input sequence, hence, reducing its dimensionality. The output of an LSTM is the "hidden layer" so we apply a dense layer to further capture linear relationships. The final dense layer converts the output of the above layers into the actual word probabilities across our entire vocabulary. We have used a RELU and Sigmoid activation function here[15]. Since the output word can be any one of the unique words, our problem is a multi-class classification problem, hence the categorical_crossentropy loss function is used here[29].

We fit the model on our data to train the same. To generate the output, a random seed sentence is selected from the training collection. This seed sentence is fed to the trained model and the next word is predicted. This is done iteratively with each newly generated word appended to the seed sentence(Figure 2). We achieve the final dialogues of ten words each as hypothesized in our problem statement.

4.2 Mood Detection

Given a seed sentence, the model predicts the next most likely word in the sequence. We make use of a Bi-directional LSTM network that has a Given the generated sentence the task of this model is to detect the mood of the sentence. For the given dataset, we might have different moods that can be conveyed at different levels. To start with, we considered the Ekman classification[37] containing 6 different sentiments - anger, disgust, joy, fear, sadness, surprise, and two other classes trust and anticipation. Since all sentences generated might not be classified into any of the above 8 classes, we included positive and negative classes as well. If none of these conditions are met due to some gibberish sentence that would have been generated, we classify the sentence as neutral. Since the work we did is closely related to sentiment analysis we interchangeably use sentiment and mood in our next sections.

As in Figure 3, the generated sentence is first pre-processed and then along with lexicon files, the sentence mood is detected. Before coming up with this process, we experimented with different ways to formulate the mood detection part. Few among them were to consider the intensity values for each word (this didn't help much because it

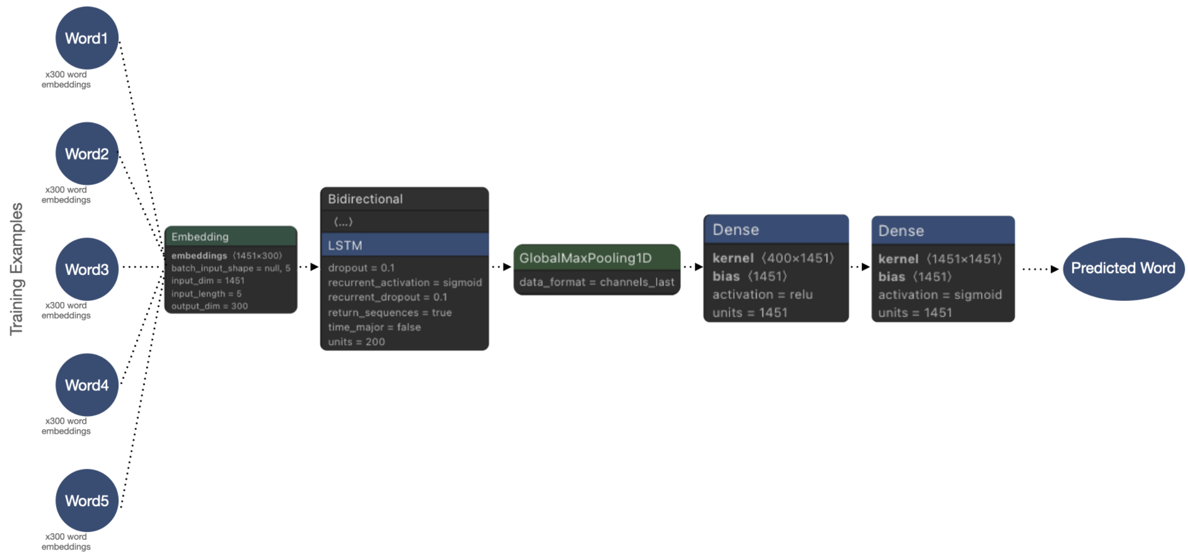

Figure 1: Bidirectional LSTM network.

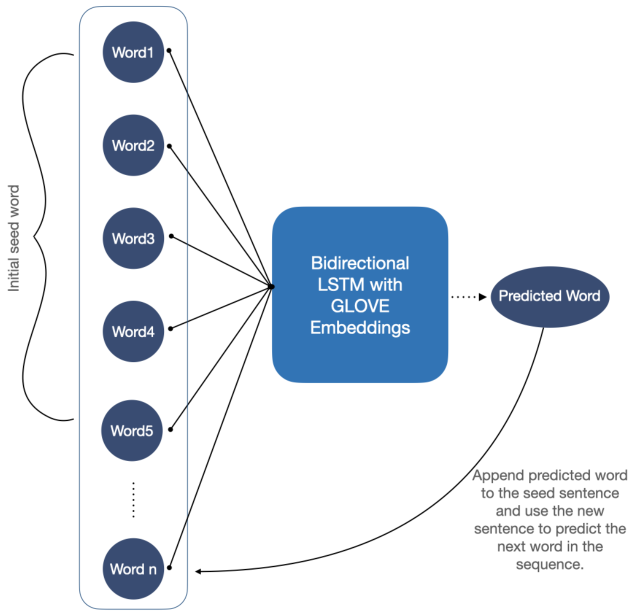

Figure 2: Dialogue Generation

was numerical and converting to a class label from this wasn't quite fruitful), the numbers obtained varied from a range of -10 to 10 (which might lead to generic mood distribution than a specific class), considering the mood of only the generated sentence (sometimes negated the sentiment conveyed). So, we finally considered the moods of the words obtained after pre-processing and calculated frequencies of each such label. This was classified as the sentence mood. This worked in our case as the number of words generated from the model is the same as the seed words count. If at all the mood of the seed sequence is joy, to sustain the balance in the generated sequence, it had to contain words of class joy, else it would be classified as neutral (due to a mix of different sentiments) or a new mood class. This approach majorly worked because of the small length of the sentence considered for our experiment. Had it been sentences with a higher number of words, the distribution would be random and this approach might not work well.

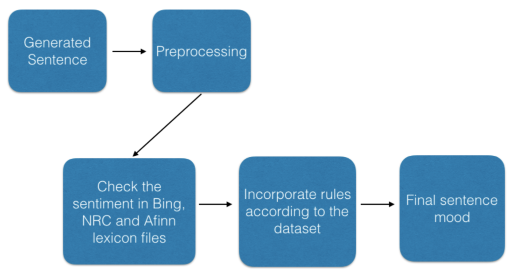

Figure 3: Mood detection pipeline

5 Evaluation

When working with text generation, it is very difficult to analyze if the sentence is actually a valuable one or not. On the first glimpse, we can see that the sentences are not necessarily grammatically correct. The model also tends to overlearn certain words such as - the, with, he, she; which makes sense as they are the most repeated words. The generated sentences for each character can be viewed in the Appendix section of the paper.

To obtain a more unbiased evaluation of the generated dialogues we decided on the following approaches:

- We checked for the BLEU (Bilingual Evaluation Understudy) Score on the 10 generated dialogues for each character. The BLEU Score is calculated in reference to the training data used for the model. It measures the similarity between the generated and original dialogues[2]. The BLEU score helps us understand if the generated dialogues capture the features of the individual they are trained for. The similarity score checks for n-gram similar words. A score closer to 1.0 would indicate a perfect match. However, BLEU score is not the ultimate measure and does not consider the semantic meaning of the sentence structure[2], hence, we also performed human evaluations.

- For human evaluation, 5 out of 10 generated dialogues for each character were selected and sectioned in the survey. Each section included 3 questions for each dialogue.

  Q1: How likely is the dialogue being spoken by that character?

  Q2. Does the dialogue sound natural?

  Q3. Does the dialogue make sense?

- We finally checked for the mood of the seed dialogue to see if it matches the generated dialogue. We compared the two moods to check if the model is able to pick the mood from the context of the seed sentence and generate the dialogue accordingly. For this, we also obtained feedback via a survey to see if the classified moods for both seed and generated dialogues were correct.

6 Results

6.1 Text Generation

The results obtained from the BLEU score and human evaluation for the generated text can be seen in Figure 4. We observe from our table that C3-PO performed the best across all evaluations. The sentences for C3-PO obtained a high BLEU score because of C3-PO's unique vocabulary which was captured very well by the model. The human evaluation was also higher for the same reason. However, Obiwan and Yoda both have simple vocabularies because of which their BLEU Scores were lower but also did not convey much to validate the generated dialogues. The human evaluation showed Yoda to be on the higher side because most of the generated sentences seemed grammatically reversed. We also believe that the highly contradictory results may be caused due to the pool of individuals selected for the survey. The individuals could have been from opposite spectrums - people who have watched and follow star wars, other who do not.

	Likeliness (How likely is the character to say the dialogue generated?) 1-5	Naturalness (How natural does generated dialogue sound?) 1-5	Sensibility (Is the sentence grammatically sensible?) 1-5	BLEU Score %
Obi-wan	2.34	2.17	2.34	70.51
Yoda	3.14	2.89	2.91	56.36
C3PO	3.69	3.7	3.61	76.85

Figure 4: Text Generation Result

6.2 Mood Detection

To evaluate how well the mood detection approach performed on the generated sentences, we considered two approaches. First, we compared the classified moods of the seed and generated sequences (through the trained model). Second, we rolled out a survey to get feedback on the classified moods and verify if the classifications were correct. The results of the survey are as in Figure 5.

	Accuracy (Comparing the moods of generated sentences to the seed words)	From Survey
Obi-wan	55.70%	33%
Yoda	66.70%	33%
C3PO	77.80%	44%

Figure 5: Mood Detection Result

The accuracy is based on the percentage match between the seed and generated dialogue mood, obtained through the classification model. The survey included options for mood selection for each generated dialogue. We then compared the survey feedback with the seed dialogue mood to obtain our survey accuracy. The high accuracy for C3PO can be attributed to the grammatically correct sentences that got generated. The survey results presented to you are as of the results considered until 04/26/2021.

7 Future Work

From our evaluations and data analysis we observed a few issues which we believe could be improved for future research:

Lack of abundance data : We observed that due to the limited amount of dialogues for each character, the model as a whole was not able to churn out the most meaningful sentences. While the sentences sounded like movie dialogues, they were not always grammatically sound. It would also be interesting to add more features to the model other than simple dialogues which could add additional data and add more value to the personality of the character.

Better models : GPT-2 (Generative Pre-trained Transformer -2) and BART are two models which are newer to the field of text generation and would be interesting to work with for further development. These models are more complex and customizable but require heavy computation power that we lacked. We could also append a pre-specified weight matrix to help curb the over-learning of common words.

A different mood text generation approach: Our proposed model considers the contextual mood of the seed sentence and expects the model to pick the same when generating the dialogues. However, we see from our results that this is not always successful. The classified mood for the seed differs from the generated one. This might be due to a lack of keywords that help identify the specific mood. If the seed sentence does not contain the keyword for the mood and generated sentence does, then the two classified moods would not match. It would have been more appropriate to label each sentence of the given dataset for training so as to check the relevance for the generated sentence mood classification. As an alternative, even a dataset that was labeled on a similar domain might be better.

8 Conclusion

Generating human-like scripts is still far from perfection. The field of NLG still lacks linguistic capabilities and cannot be relied on for generating complex high-quality dialogues[5]. Through our project, we have shown how bi-directional LSTMs can be used as a basic tool to obtain dialogues that capture the individuals' personalities in some capacity. However, we have not obtained optimum results. We require more analysis into what makes the character whole and helps obtain the perfect emotion with semantically correct dialogue. The study can be expanded into obtaining more mood-labeled data and personality features that make a character stand out from others. Obtaining more domain-specific features could also help the model churn out semantically correct sentences.

References

1 Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin and Yejin Choi. 2017. Connotation Frames of Power and Agency in Modern Films. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.

2 Asli Celikyilmaz, Elizabeth Clark, Jianfeng Gao. Evaluation of Text Generation: A Survey, 2020.

3 Josep Valls-Vargas, Jichen Zhu, Santiago Ontañón. 2019. Towards End-to-End Natural Language Story Generation Systems. AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 2019.

4 "GoEmotions PyTorch", https://github.com/monologg/GoEmotions-pytorch

5 "A Comprehensive Guide to Natural Language Generation", https://medium.com/sciforce/a-comprehensive-guide-to-natural-language-generation-dd63a4b6e548.

6 "Natural Language Generation (NLG)", https://medium.com/@b.terryjack/natural-language-generation-nlg-fe8844db6f01.

7 "Star Wars: An NLP Adventure",

https://medium.com/swlh/star-wars-an-nlp-adventure-5e0deb70cd75.

8 "How to write with artificial intelligence", https://medium.com/deep-writing/how-to-write-with-artificial-intelligence-45747ed073c#.n9zpfkh60.

9 "The Bidirectional Language Model", https://medium.com/@plusepsilon/the-bidirectional-language-model-1f3961d1fb27.

10 "Text generation with an RNN — TensorFlow Core", https://www.tensorflow.org/tutorials/text/text_generation#the_prediction_task.

11 "Module: tf.keras.preprocessing.text — TensorFlow Core v2.4.1", https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text.

12 "Generating Short Star Wars Text With LSTM's", https://towardsdatascience.com/generating-short-star-wars-text-with-lstms-c7dc65e252c4.

13 "Film Script Generation With GPT-2", https://towardsdatascience.com/film-script-generation-with-gpt-2-58601b00d371.

14 "How to evaluate Text Generation Models? Metrics for Automatic Evaluation of NLP Models", https://towardsdatascience.com/how-to-evaluate-text-generation-models-metrics-for-automatic-evaluation-of-nlp-models-e1c251b04ec1.

15 "Simple Text Generation", https://towardsdatascience.com/simple-text-generation-d1c93f43f340.

16 "A Gentle Introduction to Calculating the BLEU Score for Text in Python", https://machinelearningmastery.com/calculate-bleu-score-for-text-python/.

17 "How to Develop a Word-Level Neural Language Model and Use it to Generate Text", https://machinelearningmastery.com/how-to-develop-a-word-level-neural-language-model-in-keras/.

18 "Text Generation With LSTM Recurrent Neural Networks in Python with Keras", https://machinelearningmastery.com/text-generation-lstm-recurrent-neural-networks-python-keras/.

19 "A Story Generator", https://github.com/aherriot/story-generator.

20 "Text Generation using BiLSTM PyTorch", https://github.com/FernandoLpz/Text-Generation-BiLSTM-PyTorch.

21 "Text Generator for Harry Potter", https://github.com/amisha-jodhani/text-generator-harry-potter.

22 "Weighted word autocomplete using Star Wars dataset", https://www.kaggle.com/hamishdickson/weighted-word-autocomplete-using-star-wars-dataset.

23 "Mimicking Star Wars Characters using A.I (RNN)", https://www.kaggle.com/guidant/mimicking-star-wars-characters-using-a-i-rnn#2.-Data-Preparation.

24 "Classic NLP exploration of Star Wars dialogues", https://www.kaggle.com/liopic/classic-nlp-exploration-of-star-wars-dialogues.

25 "Starwars Chatbot - NLP application", https://www.kaggle.com/andrebarros1/starwars-chatbot-nlp-application/output.

26 "Analyzing Star Wars Movie Scripts", https://www.kaggle.com/xvivancos/analyzing-star-wars-movie-scripts/execution.

27 "Chatbot Star Wars", https://www.kaggle.com/eriktrs/exerc-cio-chat-bot-star-wars.

28 "Beginners Guide to Text Generation using LSTMs", https://www.kaggle.com/shivamb/beginners-guide-to-text-generation-using-lstms.

29 "Python for NLP: Deep Learning Text Generation with Keras", https://stackabuse.com/python-for-nlp-deep-learning-text-generation-with-keras/.

30 "Text Generators — Text Generation Using Python", https://www.analyticsvidhya.com/blog/2018/03/text-generation-using-python-nlp/.

31 ”Analyzing Star Wars Movie Scripts”, https://www.kaggle.com/xvivancos/analyzing-star-wars-movie-scripts/.

32 ”Emotion Detection in Reddit Comments”, https://github.com/google-research/google-research/tree/master/goemotions/.

33 ”NLP - Emotion Detection with Multi-class”, https://rian-van-den-ander.github.io/Multi-Label-Emotion-Classification/.

34 ”Star Wars Movie Scripts Kaggle Dataset”, https://www.kaggle.com/xvivancos/star-wars-movie-scripts.

35 ”Movie Scripts Dataset”, https://imsdb.com/.

36 ”Project Code Repository”, https://github.com/keshvi-srivastava/star-wars-dialogue-generation.

37 ”Universal Emotions”, https://www.paulekman.com/universal-emotions/.

Appendix A

A.1 Obiwan-Kenobi Generated Sentences and Bleu Score

```
to relieve you of chancellor palpatine not join him have
name obiwan since oh before you were born you two
you not been so impatient to defeat vader then you
artoo switch on the comlink do artoo it certainly cant
a morecivilized time for over over the left and left
it surrounds us and penetrates us us grind us killed
brokenand dont forget shes apolitician theyre apolitician all betrusted over
as well thats not all very proud that you two
by rushing back prematurely i fear something back and you
course but it was a good friend which reminds he

BLEU score for ben -> 0.7051203366541631
```

A.2 Yoda Generated Sentences and Bleu Score

```
remains vader you must confront you become ready you you
transform into the force mourn you not it your you
hurry careful timing we will need i are are you
times nothing is what itappears to the dark side side
twilight is upon me and the dark side the dark
then a jedi will you have for i old to
of the emperor or suffer the force the dark side
itappears to be but the force you are be with
train yourself to let go it it is will be
a fully trained jedi knight the deepest force become you

BLEU score for yoda -> 0.5636460841171014
```

A.3 C3-PO enerated Sentences and Bleu Score

```
where are we going i cant see know should you
gone with master luke than stay here with you well
told you artoodetoo you know better than to trust trust
were coming out of the asteroid field has direct to
this thats why ive also been programmed to make up
communicate but it has the most peculiar dialect i believe
has been rather a longtime on you master luke is
get me onto one of those youre going to get
oh my id forgotten how much i hate space travel
wait oh dear artoo artoo artoo artoo you might to

BLEU score for threepio -> 0.7684978670123553
```

A.4 Obiwan-Kenobi Mood Comparison

sentence	sentiment	seed_words	seed_words_mood
to relieve you of chancellor palpatine not join him have	negative	to relieve you of chancellor	positive
name obiwan since oh before you were born you two	neutral	name obiwan since oh before	neutral
you not been so impatient to defeat vader then you	anticipation	you not been so impatient	anticipation
artoo switch on the comlink do artoo it certainly cant	neutral	artoo switch on the comlink	neutral
a morecivilized time for over over the left and left	neutral	a morecivilized time for over	anticipation
it surrounds us and penetrates us us grind us killed	fear	it surrounds us and penetrates	fear
brokenand dont forget shes apolitician theyre apolitician all betrusted over	neutral	brokenand dont forget shes apolitician	neutral
as well thats not all very proud that you two	sad	as well thats not all	negative
by rushing back prematurely i fear something back and you	fear	by rushing back prematurely i	surprise
course but it was a good friend which reminds he	joy	course but it was a	neutral

A.5 Yoda Mood Comparison

sentence	mood	seed_words	seed_words_mood
remains vader you must confront you become ready you you	anticipation	remains vader you must confront	anger
transform into the force mourn you not it your you	negative	transform into the force mourn	negative
hurry careful timing we will need i are are you	positive	hurry careful timing we will	anticipation
times nothing is what itappears to the dark side side	sadness	times nothing is what itappears	neutral
twilight is upon me and the dark side the dark	sadness	twilight is upon me and	neutral
then a jedi will you have for i old to	neutral	then a jedi will you	neutral
of the emperor or suffer the force the dark side	sadness	of the emperor or suffer	negative
itappears to be but the force you are be with	fear	itappears to be but the	neutral
train yourself to let go it it is will be	positive	train yourself to let go	positive
a fully trained jedi knight the deepest force become you	negative	a fully trained jedi knight	positive

A.6 C3-PO Mood Comparison

sentence	mood	seed_words	seed_words_mood
where are we going i cant see know should you	negative	where are we going i	neutral
gone with master luke than stay here with you well	positive	gone with master luke than	positive
told you artoodetoo you know better than to trust trust	trust	told you artoodetoo you know	neutral
were coming out of the asteroid field has direct to	anticipation	were coming out of the	anticipation
this thats why ive also been programmed to make up	neutral	this thats why ive also	neutral
communicate but it has the most peculiar dialect i believe	negative	communicate but it has the	trust
has been rather a longtime on you master luke is	positive	has been rather a longtime	neutral
get me onto one of those youre going to get	anticipation	get me onto one of	neutral
oh my id forgotten how much i hate space travel	disgust	oh my id forgotten how	negative
wait oh dear artoo artoo artoo artoo you might to	neutral	wait oh dear artoo artoo	neutral

A.7 Links to Google Forms

Generated Dialogue Evaluation :
https://docs.google.com/forms/d/1TnN4HxC6i1gJHcKvJDVfkTTbktq6GAcIyxjW00aaysA/edit?usp=sharing.

Detected Mood Evaluation :
https://forms.gle/6qFY6z8wCLbjqRvS6.