Vader to Potter: A Neural Machine ChatBot

Ishaan Agrawal
Columbia University
Computer Science (B.S)
ia2341@columbia.edu

Sunand Iyer
Columbia University
Computer Science (B.S)
sri2117@columbia.edu

Viraj Rai
Columbia University
Computer Science (B.S)
vr2736@columbia.edu

Purpose and Background to the Investigation

The ability to model a person's speaking style, grammatical structure, personality and tone can provide us with the ability to create and design a variety of communication platforms. The wide range of applications extends to electronic recorded communications for customer service firms, mental health chat bots tuned for individual patients, machine translation that preserves tone and personality etc. The aim of this investigation is to be able to model character-influenced conversation in a chat-box environment that can preserve a fictional character's individual eccentricity, personality and tone.

1. Introduction and Goal

The goal of this research endeavour is to create a Seq2Seq model that models conversation between 2 fictional characters from movie scripts in a chat box environment. The model will be trained on multiple movie transcripts with dialogues between 2 characters. The chatbot should be able to speak in coherent sentences and be able to respond appropriately to dialog. Furthermore, another aim of the model will be to preserve character tone and personality in unexplored conversational domains. For example, a conversation between Darth Vader (Star Wars series) and Ron Weasley (Harry Potter series) would begin with a topic question or statement and the characters will venture on a conversational back-and-forth with the same style as defined in the movie scripts. We hope to have a chatbot that can respond to input especially similar to movie dialogues. The reason behind training the model on characters from movie transcripts is twofold:

- There is abundant training data available on transcripts of movie series. For example, series such as Star Wars and Harry Potter can provide character dialogues of up to 10 hours (movie time). This gives us the ability to focus purely on the model architecture rather than worrying about data mining.
- If we can accurately model conversations between fictional characters with solely movie transcripts, then we can confidently state that one can mimic a real individual's speaking style and tone with enough conversational data.

2. Related Work and References

The inspiration for our initial approach to the chatbot model was the work put forth in the paper "A Neural Chatbot With Personality" [5]. In this published paper, Huyen Nguyen demonstrates the ability to train a neural chatbot solely with move transcripts. While the conversation is not as freely flowing as one would expect, the output of the chatbot model were grammatically coherent, even though the responses were irrelevant to the input utterance [5]. In order to start with a more conversationally coherent model, we used the published article "A Neural Conversational Model" [8] as inspiration for our model architecture. We will be treating the model provided in [8] as a starting point, and will also provide a good baseline to evaluate our improvements to the model.

In the paper "Enterprise to Computer: Star Trek chatbot", Grishma Jena approaches the neural chatbot with a similar strategy [4]. Both [5] and [4] use the Cornell Movie Dataset [3] as the base dataset with which they train their neural models. However, [4] takes it one step further by fine tuning the generic model on Star Trek movie transcripts, thereby injecting a science-fiction personality to the chatbot. We will be using a similar approach when trying to ensure that our chatbot model that mimics a movie character does so by enforcing the character's perks, tone and eccentricity.

In order to tackle the problem of conversational coherence that plagues the models in [5] and [4], we will use some findings and approaches from a paper titled "Topic Aware Neural Response Generation" [2]. This approach uses joint attention vectors from the forward pass of previous conversational statements to create a topic-aware attention vector that provides context to the model for future responses, as was also used in [8].

3. Dataset

For all the work performed so far, the major dataset used to train our neural network is the Cornell Movie Dataset [3] as used by [5] and [4]. This dataset contains 220,579 conversational utterances, along with a vocabulary of around 9,000 unique words that occur at least 6 times (words that appear once are treated as unknown $< UNK >$). This dataset provides a good starting point for model training. The reason for this is that the end goal of the project is to simulate conversation with a fictional character. Therefore, a generic movie transcript dataset of fictional characters will provide a good guideline for character to character interaction in a fictional setting. For example, a chatbot with Star Wars cahracter Yoda would be difficult to mimic using a generic data corpus, as the grammatical structure of the majority of Yoda's responses are inverted to conventional English. Therefore, the Cornell Movie Dialog Corpus provides a nice starting point to train the model.

While [3] will be used to train a generic conversational bot, we will use character specific dialog corpuses for major movie characters. Keeping the scope of the project in mind, we would like to create a model for (at most) 3 major fictional characters. Examples of these characters could include Darth Vader (Star Wars Original Trilogy), Harry Potter and Sheldon Cooper (The Big Bang Theory). While [3] does provide for dialogues for some of these characters, the corpus is not exhaustive. Therefore, we will also resort to "The Hogwarts Library" [1], a Kaggle based Star Wars Dataset [9] etc. to fine tune our generic model.

4. Current Model Architecture and Algorithm

The proposed project takes place in 3 phases. The first phase is to generate and train the baseline model described in [8]. The correctness and shortcomings of the model will be taken into account when finalizing the architecture of our novel model. The second stage would be to modify the baseline model with certain architectural changes with an expectation in conversational improvements. These improvements should include conversational coherence along with grammatical correctness. The third phase of the project is to fine tune the generic model on character specific datasets in order to create a chatbot that mimics the tone and style of a certain fictional character.

Given below is a list of technical specifics used in the current model as well as improvements that are to be made:

1. Encoder - Decoder Seq2Seq Model
2. Attention Decoder
3. Model with Buckets
4. Google Word Embeddings

4.1. Encoder - Decoder Seq2Seq Model

The current model is a variation on a Seq2Seq model; however, the major architecture of the model is a classic Seq2Seq model. It consists of two sub-models, an encoder model and a decoder model that takes the input utterance and the desired response respectively. The encoder model will generate a vector representation of the input statement. Meanwhile the decoder will try to estimate the probability of the next token based on the encoded vector and the input of the decoder.

4.2. Attention Based Decoder

It is the opinion in the academic community that attention provides significant improvement to classical deep learning approaches to deep learning tasks. This is because the attention mechanism allows the decoder units to only focus on the desirable portions of the encoded vector generated by the encoder model. The current model also has an attention mechanism inbuilt into it. This should allow for a coherent conversational structure. The attention mechanism mimics the description suggested in "End-to-End Continuous Speech Recognition" by Bahdanau et. al [6].

4.3. Model With Buckets

This is a technical aspect that is to be incorporated into the current model for better inference generation. For testing, we would like to incorporate a bucketing system. The bucket system works as follows. A bucket can be defined as a 2-tuple (x, y), where x is the length of the longest input statement that can fit into that bucket, and y is the maximum length of the output statement in that bucket. Buckets of different sizes accept non-intersecting sets of input and output sentences. The bucketing system allows us to evaluate the model based on input and output sentences of different length. Therefore, information regarding various loss values at these different buckets can give us information regarding the accuracy of the model with varying input lengths.

4.4. Google Word Embeddings

One shortcoming of the baseline model proposed in [5] is the fact that it uses a hashing technique to represent individual tokens of an utterance. These tokens are then passed through an embedding layer where word embeddings are generated. Since the baseline model is trained specifically on fictional movie dialog corpuses, the generated word embeddings may not accurately reflect the semantic information stored in the token. Moreover, these embeddings are learned on the fly. However, since the dataset used to train the embeddings is on the order of 200,000 data points, the generated embeddings may be prone to errors. Therefore, we will look to improve on the baseline by using a pre-trained Google word embedding. We experimented with a variety of word embeddings before arriving at the 300 dimensional word embedding, including a variety of different Stanford Glove embeddings [7]. Given below are a list of all the word embeddings that were experimented with before finazlizing the choice.

- 6B - 50d, 100d, 200d and 300d Glove Embeddings
- 42B-300d Glove Embeddings (Common Crawl)
- 27B-200d Glove Embeddings (Twitter Corpus)
- Google News Word Embeddings

While judging each pre-trained word embedding, we looked at the vocabulary quality and vocabulary size. Although the dimensionality of the Google embeddings was the largest (300 dim vectors), it resulted in a larger overall vocabulary size when combining it with the Cornell Movie Dialog corpus. Moreover, the large dimension size may provide better information regarding the semantics of the word.

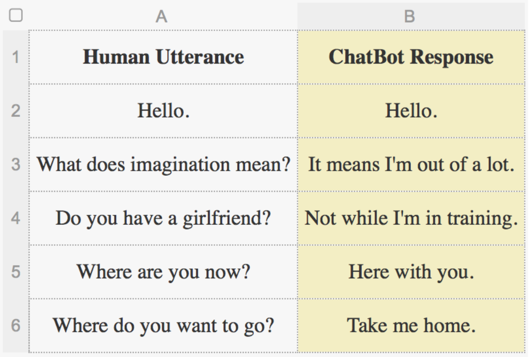
4.5. Overall Model

The overall model can be summarized as follows:

- Seq2Seq Model
- Encoder - Attention Decoder: Legacy Tensorflow contributions. 2 LSTM Layers each
- Google Word Embeddings
- TODO: Inference using a Bucketing System [5]
  - Small - (19, 19), (28, 28), (33, 33)
  - Large - (40, 43), (50, 53), (60, 63)

5. Preliminary Results

As of the milestone date, we have performance results from the baseline model derived from the description in [8]. Given below are some sample conversational exchanges with the baseline chatbot.



Human Utterance	ChatBot Response
Hello.	Hello.
What does imagination mean?	It means I'm out of a lot.
Do you have a girlfriend?	Not while I'm in training.
Where are you now?	Here with you.
Where do you want to go?	Take me home.

The model seems to always output grammatically coherent statements if the output is a complete sentence. Otherwise, it takes liberty in using ellipsis (...) to keep the sentence incomplete. Smaller input sentences tend to yield reasonable output (grammatically, maybe not semantically); however, the model is too large to support longer sentences. Therefore, we kept the model limited to an utterance of 20 words. This is a major problem and we would like to rectify this for our final model. We can do so by adding the bucketing system to reduce paddings for shorter sentences.

Given below are a few statistics for the output results of the model:

- Perplexity: 1.87
- BLEU Score: 0.4476

References

[1] The hogwarts library. http://www.hogwartsishere.com/library/.

[2] Y. W. J. L. Y. H. M. Z. W.-Y. M. Chen Xing, Wei Wu. Topic aware neural response generation. http://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14563/14260.

[3] C. Danescu-Niculescu-Mizil and L. Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[4] A. B. L. U. J. . S. Grishma Jena, Mansi Vashisht. Enterprise to computer: Star trek chatbot. https://arxiv.org/pdf/1708.00818.pdf.

[5] T. C. Huyen Nguyen, David Morales. A neural chatbot with personality. http://web.stanford.edu/class/cs224n/reports/2761115.pdf.

[6] K. C. Y. B. Jan Chorowski, Dzmitry Bahdanau. End-to-end continuous speech recognition using attention-based recurrent nn. 2014.

[7] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[8] O. Vinyals and Q. V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015.

[9] Xavier. Star wars movie scripts. https://www.kaggle.com/xvivancos/star-wars-movie-scripts, 2018.