**Abstract**

A chatbot is a program that provides conversational output in response to user input. They have many applications such as customer support interfaces, general question answering services, translation apps and virtual assistants. A common goal for chatbots is to simulate a human like interaction for the user. To this end many researches have investigated creating chatbots which can do more than just factually answer questions. That is, they have investigated creating chatbots which have "personality" or "identity." In this work we propose attempting to train a neural network to act as a chatbot which simulates a known personality. Namely, we want to train a neural network to react as TV show personalities.

**Dataset**

F.R.I.E.N.D.S was chosen as the TV show for character dialogue extraction. The initial dataset consisted of the raw script of the TV show.

The raw script was parsed to extract a character’s responses for a given input. Joey was chosen as the character to study as he has a distinctive personality in the show. This gave us the input and output data for a single character to train and test on.

In addition to the characters responses, a separate generic question and answer dataset was used to give the Machine Learning models an understanding of how questions are answered.

**Transformer**

The model used in this task draws inspiration from the structure of the Seq2Seq model with the additional change being the use of Transformers. The model employs the use of Scaled dot product attention and Multi-head attention. The scaled attention provides a basic attention mechanism using encoder hidden states and previous hidden state of the decoder. The multi-head attention mechanism runs multiple scaled attention computations parallelly. These attentions are concatenated and linearly transformed to act as an ensembling method.

The transformer contains Encoder and Decoder units. The encoder generates an attention-based representation with capability to locate a specific piece of information from a potentially infinitely-large context. The encoder contains a multi-head attention mechanism followed by a fully connected feed-forward network. The Decoder unit deals with retrieval from encoded representations, containing a masked multi-head attention, a multi-head attention and a fully connected feed-forward network.

The model we use, contains several such encoder-decoder units formed by transformers. A final fully connected Dense layer uses decoder outputs. The loss function used is a Sparse Cross Entropy loss function available in keras.losses. The model uses the Adam optimizer and a learning rate decay scheduler to decrease the learning rates with respect to number of epochs.

**Evaluation**

The problem of evaluating a chatbot or dialogue machines is well known to be hard. The problem becomes only harder when adding the constraint that the chatbot should have some human like characteristics such as personality or memory [Liu et. al., Radziwill and Benton, Xing Fernández]. A key obstacle faced is that given any statement there could be a large number of appropriate responses. That is, there is no one ideal map from statements to responses that a chatbot would want to learn. Thus, it is hard to construct a test dataset that contains all possible valid responses for every tested statement. Thus, to efficiently evaluate, one must find a way to grade responses without being able to simply compare it against a list known answer. We made two attempts at such metrics.

While they are known to not perform well, it is still common in the literature to attempt to use automatic metrics of some kind. Most of these metrics were originally designed to evaluate translation bots. Such bots face similar problems to those faced by chatbots; there are many appropriate translations of the same phrase for example. These metrics do have some success by comparing then sentence structure and n-gram pattern of generated responses and reference responses. We considered a few common metrics which do this: BLUE, METEOR, ROGUE and WER.

Another metric is human evaluator. This is often used as a much more informative measure of chatbot success as humans tend to have natural ability to tell if something sounds human and/or embodies a specific personality trait. For this test we collected individual familiar with the strong personality which are trying to model with our chatbot (Joey from Friends) and asked them to answer a series of questions. These questions listed a statement from the TV show and asked the tester to pick which of two responses was the most appropriate. Here appropriate would mean both a reasonable response and ideally the most which is similar to the character in question. One response was the one generated while the other was from the original script. They were presented in the same format but in random order. Finally, a total percentage of selected generated responses was calculated and averaged over all testers.