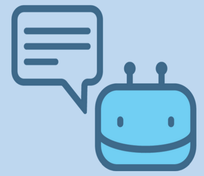
Mimic: Character Based Chatbot

Nirvan S. Theethira
nith5605@colorado.edu

Charlie Carlson
chca0914@colorado.edu

Ketan Ramesh
kerk1228@colorado.edu

Pramod Vankatesh Kulkarni
prku8035@colorado.edu



Introduction

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." The goal of this project was to train a neural network to act as a chatbot that simulates the known TV personality Joey from F.R.I.E.N.D.S.

Dataset (*F.R.I.E.N.D.S*)

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 for all seasons.

```
The One Where Monica Gets a New Roommate (The Pilot-The Uncut Version)
Written by: Marta Kauffman & David Crane
Transcribed by: guineapig
Additional transcribing by: Eric Aasen
(Note: The previously unseen parts of this episode are shown in blue text.)

[Scene: Central Perk, Chandler, Joey, Phoebe, and Monica are there.]

Monica: There's nothing to tell! He's just some guy I work with!

Joey: C'mon, you're going out with the guy! There's gotta be something wrong with him!

Chandler: All right Joey, be nice.  So does he have a hump? A hump and a hairpiece?

Phoebe: Wait, does he eat chalk?

(They all stare, bemused.)
```

The raw script was parsed to extract statement response pairs from each scene. We then collected all pairs where Joey was the responder. Joey was chosen as the character to study as he has a distinctive personality in the show. For those not familiar the show, Joey can be described as naive, sarcastic, loving, misogynistic and loud.

Charachter Input	Charachter Output
Ooh . . . I love Barbados !	<START> Ooh . . . I cant believe Im kissing yo...
Hey , you know , before you said that nothing ...	<START> Well , I only said that because of Ros...
Rach , you there ?	<START> Oh my God , its Ross . What are we gon...
Oh , ju ju just stay calm . Just be calm . For...	<START> No idea what it means . <END>
Yeah , sure . . .	<START> I dont believe this . . . Have you guy...

In additional to having Joey like personalities, we wanted our chatbot to in general respond like a human. To this end we considered pretraining our models on an additional dataset of general question and answer pairs. The goal of introducing this data set was to help the model understand generic a question answering format. We reserved 20% of both datasets to be used as testing data.

Generic Questions	Generic Answers
What s up ?	<START> Not much . <END>
you are not put together	<START> i have always thought whoever did it c...
Tell me a joke	<START> what do you get when you cross a road ...
Who uses super computers ?	<START> Anybody who wants to work with large n...
how far is the sun	<START> the sun is about million miles from ea...
leo tolstoy	<START> Is one of Russia s greatest writers . ...
what is chemistry	<START> the science of mixing chemicals . <END>
Bend over	<START> That s personal ! <END>

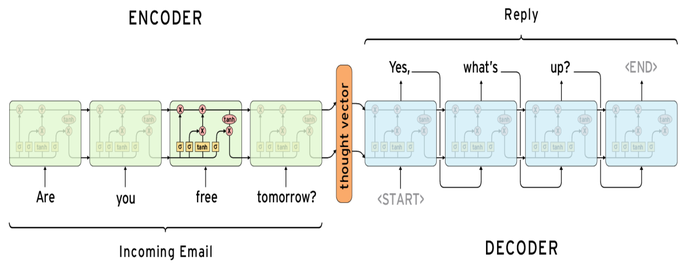
Seq2Seq Model

A *Seq2Seq model* consists of an *Encoder* and a *Decoder*.

All input text is preprocessed, tokenized and converted to Pre-trained GloVe embeddings with the help of a Keras embedding layer. Word vectors are fed to encoder *long short-term memory (LSTM)* one word at a time.

The encoder LSTM tries to capture the essence of the encoder input sequence in two state vectors that are passed to the decoder LSTM. The decoder LSTM, which is fed output text word embeddings is trained to use a dense SoftMax output layer to predict the most probable output word given two input state embeddings and an embedded word.

Once the model is trained an *inference encode decoder model* is created using the trained model weights. The inference encoder model takes in a user input question and generated two state vectors. The input state vectors are fed into the inference decoder model along with a sentence start token and all words predicted by the SoftMax layer are captured until an end token is generated.



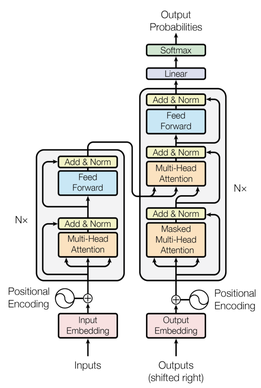
Transformer Model

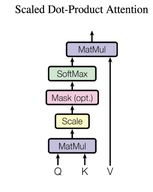
The model is pre-trained on a Question-Answering dataset. The model contains an Encoder-Decoder mechanism with transformers.

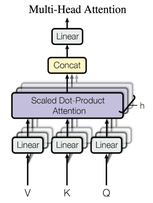
The Encoder contains a multi-head attention mechanism which runs several scaled dot product attention procedures in parallel. This acts as an ensemble method and enables self attention.

The Decoder deals with retrieval from the encoder representations using a masked multi-head attention, a multi-head attention and a feedforward network.
A fully connected Dense layer uses decoder outputs with the loss function being sparse cross entropy loss function, the accuracy metric being sparse categorical accuracy.

The following diagrams illustrate the structure of the abovementioned components.







Evaluation

Evaluating the quality of a chatbot's responses requires some creativity. One obstacle is that any statement can have many appropriate responses. Another is that deciding if a response embodies a specific personality is highly subjective. Consider the following examples of responses to the statement "Hello" Which is the most appropriate?

A. Howdy B. Hello C. Hey D. Yo E. Greetings

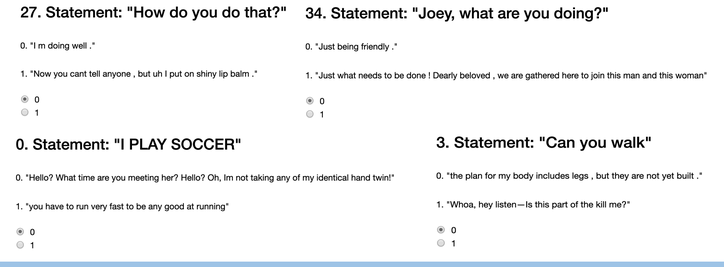
Automatic Metrics

The *Automatic metrics* we considered compared the sentence structure and word patterns of our generated responses against test responses for the same statement. Most of these metrics were designed to test translation bots which often use similar models and face similar problems. In the table below you can see data for a range of automatic metrics such as BLEU, METEOR, ROUGE and WER. With the exception of WER each of the scores is between 0 and 1 with 1 being the best. For WER, higher scores are worst and 0 is the best possible score.

Metric:	BLEU:	ROUGE-1:	ROUGE-*l*:	METEOR:	WER Avg:
S2S (Q&A)	0.055	0.213	0.204	0.092	1.554
S2S (Joey)	0.037	0.141	0.131	0.064	1.922
Trans (Q&A)	0.001	0.102	0.094	0.035	2.412
Trans (Joey)	0.100	0.270	0.261	0.137	1.218

Human Metrics

The *Human Metric* we used was a blind human evaluation. A human tester was given a collection of statements from each data set along with their original response and generated response (in random order). The tester asked to select the responses which they felt was most like the character in question. Finally, we average the ratio of generated responses selected over all testers. In an ideal setting, a human tester would not be able to distinguish between the generated response and the test response. So, an ideal score would be 0.5. The Seq2Seq model got a score of **0.413** for the question and answer dataset and **0.35** for the Joey dataset. The Transformer model got a score of **0.263** for the question and answer dataset and **0.325** for the joey dataset.



Conclusion

Overall, both of models were able to generate responses that were somewhat Joey like. Poor performance with the automatic metrics was somewhat expected; even "perfect" generated response could differ greatly from a specific test response. This has also been observed in many pervious research projects that have designed chatbots with similar goals. The human metric did demonstrates that in some cases human testers familiar with Joey could not distinguish between generated responses from each model and real test responses. To improve upon these results it is suggest that one more careful do data extraction, allow for additional training time or, of course, consider more complex models with additional layers.