CS584: A DRAMATIC DIALOGUE

Daniyal Mufti
Stevens Institute of Technology
dmufti@stevens.edu

ABSTRACT

NLP chat-bots learn languages in a similar way that children learn a language. After having learned a number of examples, they are able to make connections between questions that are asked in different ways. In this way, the bot understands what the question is about without being precisely programmed for it and an appropriate answer can be given. In a conversation form, this is also called Conversational AI. This is an important problem in conversational AI development and is the clearest form of the Turing test.

1 Introduction

A generative chat-bot is an open-domain chat-bot program that generates original combinations of language rather than selecting from pre-defined responses. The purpose of this project is to build numerous versions of a generative conversational chat-bot on the Cornell large dialogue data-set using LSTM Seq2Seq RNN Encoder - Decoder architecture and evaluate it based on a self developed human evaluated relative scale quantitative Turing Test. We will use different model parameters to train different versions of the chat bot. We will then evaluate our different model versions on a human evaluated Turing Test with a self developed relative scale. 10 Questions will be devised to test our model versions. This project will shed light on the affect of using more or less data, the result of running more or fewer epochs and the viability of our suggested human evaluated relative scale quantitative Turing Test.

2 Model Variations Explored

- LSTM Seq2Seq RNN Encoder - Decoder architecture - The method we will explore is an LSTM Seq2Seq RNN Encoder Decoder architecture with 10,000 conversations used using 1 epoch. this will be a baseline model; we expect this to do poorly.
- LSTM Seq2Seq RNN Encoder - Decoder architecture - The method we will explore is an LSTM Seq2Seq RNN Encoder Decoder architecture with 10,000 conversations used using 300 epochs.
- LSTM Seq2Seq RNN Encoder - Decoder architecture - The method we will explore is an LSTM Seq2Seq RNN Encoder Decoder architecture with 15,000 conversations used using 150 epochs.
- LSTM Seq2Seq RNN Encoder - Decoder architecture - The method we will explore is an LSTM Seq2Seq RNN Encoder Decoder architecture with 15,000 conversations used using 300 epochs.

3 Data-Set

We will use the Cornell Large Dialogue data-set located at https://www.cs.cornell.edu/ cristian/Cornell_Movie-Dialogs_Corpus.html to train our model. This corpus contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts. Since we are training different model variations for evaluation, we will keep our data-set static. This is an extremely large corpus with 220,579 total conversations; Because of space and time limitations, we will limit ourselves to 5,000-15,000 conversations.

4 Evaluation

We will evaluate our different model variations on a self developed, relative scale Turing Test. The scale will be from 1 - 10 where 1 indicates the model does poorly when interacting with the participant and 10 indicates the model does very well on the Turing Test when interacting with the participant. The score will be given by question and then averaged out over the 10 questions which will be our final quantitative metric.

A blurb about the Turing Test:-

The Turing Test is a method of inquiry in artificial intelligence (AI) for determining whether or not a computer is capable of thinking like a human being. The test is named after Alan Turing, the founder of the Turing Test and an English computer scientist, cryptanalyst, mathematician and theoretical biologist. Turing proposed that a computer can be said to possess artificial intelligence if it can mimic human responses under specific conditions.

This method of evaluation is something I am proposing as a new way to quantitatively measure the performance of a chat bot. A quantified relative scale Turing Test that we can test the performance of one chat bot against a baseline measure. While my research into the subject is relatively limited, I have not come across such a metric. In the future I will work on refining the metric for future research.

5 Other Related Works

There are, no doubt, many exceptional papers and bodies of work out there exploring conversational AI and generative chat-bots. Due to the limitation in time, I have only been able to explore a handful of these which I will cite below. I will strive to do more research in this area in the near future.

- How to build a State-of-the-Art Conversational AI with Transfer Learning by Thomas Wolf (https://medium.com/huggingface/how-to-build-a-state-of-the-art-conversational-ai-with-transfer-learning-2d818ac26313)
- Building a Conversational Chatbot with NLTK and TensorFlow by Bamigbade Opeyemi (https://heartbeat.fritz.ai/building-a-conversational-chatbot-with-nltk-and-tensorflow-part-2-c67b67d8ebb)
- The Second Conversational Intelligence Challenge (ConvAI2) by Emily Dinan et al. (https://arxiv.org/abs/1902.00098)

6 Experimental Design

- I used Python for implementing the code.
- Keras high level API library was used to build and fit our models.
- Self built functions were created to clean the data.
- Our architectural design consisted of using an embedding matrix(hence an embedding layer in the creation of our neural network),1 LSTM layer to train on questions and then use the weights of that layer as the inital state to train on the answers using a second LSTM layer. Hence the first LSTM layer will serve as an encoder while the second LSTM layer will serve as a decoder. The final step of the model is to design an inference model that will be used to decode unknown input sequences and let the participant interact with the chat bot. We will use a similar encoder decoder architecture to design this.

```
Model: "model"
_________________________________________________________________________________________________
Layer (type)                 Output Shape              Param #     Connected to
=================================================================================================
input_1 (InputLayer)         [(None, None)]            0
_________________________________________________________________________________________________
input_2 (InputLayer)         [(None, None)]            0
_________________________________________________________________________________________________
embedding (Embedding)        (None, None, 200)         2118800     input_1[0][0]
_________________________________________________________________________________________________
embedding_1 (Embedding)      (None, None, 200)         2118800     input_2[0][0]
_________________________________________________________________________________________________
lstm (LSTM)                  [(None, 200), (None,      320800      embedding[0][0]
_________________________________________________________________________________________________
lstm_1 (LSTM)                [(None, None, 200),       320800      embedding_1[0][0]
                                                                   lstm[0][1]
                                                                   lstm[0][2]
_________________________________________________________________________________________________
dense (Dense)                (None, None, 10594)       2129394     lstm_1[0][0]
=================================================================================================
Total params: 7,008,594
Trainable params: 7,008,594
Non-trainable params: 0
```

```
_________________________________________________________________________________________________
Inference decoder:
Model: "model_1"
_________________________________________________________________________________________________
Layer (type)                 Output Shape              Param #     Connected to
=================================================================================================
input_2 (InputLayer)         [(None, None)]            0
_________________________________________________________________________________________________
embedding_1 (Embedding)      (None, None, 200)         2118800     input_2[0][0]
_________________________________________________________________________________________________
input_3 (InputLayer)         [(None, 200)]             0
_________________________________________________________________________________________________
input_4 (InputLayer)         [(None, 200)]             0
_________________________________________________________________________________________________
lstm_1 (LSTM)                [(None, None, 200),       320800      embedding_1[0][0]
                                                                   input_3[0][0]
                                                                   input_4[0][0]
_________________________________________________________________________________________________
dense (Dense)                (None, None, 10594)       2129394     lstm_1[1][0]
=================================================================================================
Total params: 4,568,994
Trainable params: 4,568,994
Non-trainable params: 0
_________________________________________________________________________________________________
Inference encoder:
Model: "model_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         [(None, None)]            0
_________________________________________________________________
embedding (Embedding)        (None, None, 200)         2118800
_________________________________________________________________
lstm (LSTM)                  [(None, 200), (None, 200) 320800
=================================================================
Total params: 2,439,600
Trainable params: 2,439,600
Non-trainable params: 0
_________________________________________________________________
```

- We have 4 model variations. Model 1 has 10,000 conversations with 1 epoch, Model 2 has 10,000 conversations with 300 epochs, Model 3 has 15,000 conversations with 150 epochs and Model 4 has 15,000 conversations with 300 epochs. The first model variation took approximately 15 minutes to run, the second model variation took approximately 10 hours to run, the third model variation took about 14 hours to run and the fourth model variation took 26 hours to run. These models were run on a CPU I7-6700K quad Core processor.

7 Experimental Results

```
Model 1

1.Enter question : how are you
 i
score:1

2.Enter question : what is your name
 i i
score:1
3.Enter question : what do you do
 i i
score:1
4.Enter question : what day is it
 i i
score:1
5.Enter question : are you happy
 i
score:1
6.Enter question : are you sad
 i
score:1
7.Enter question : what did you do today
i i i
score:1
8.Enter question : what

score:1
9.Enter question : why

score:1
10.Enter question : how

score 1

Average score 1
```

```
Model 2
1.Enter question : how are you
 i what to that with the goddamn that i am i that not what to put do
 here

score: 3
2.Enter question : what is your name
 leeloo minai lekararibalaminaitchai twombley ekbat de sebat

score:4
3.Enter question : what do you do
 no i that not know

score:4
4.Enter question : what day is it
 i am not

score: 3
5.Enter question : are you happy
 when you are me the times

score: 4
6.Enter question : are you sad
 yeah my ah

score:3
7.Enter question : what did you do today
 you know i what been thinking about tell you about the girls my
 father and then i am gonna bust you

score:5
8.Enter question : what
 i am not really hungry but i would like to what a bloodthirsty mob

score:4
9.Enter question : why
 because it is my duty

score 9
10.Enter question : how
 just ask hal it is conceivable they might keep something from
us but they would never keep anything from hal

score 9

Average Score : 4.8
```

```
Model 3
Enter question : how are you
 been better i read about what is been happening i want

score: 7
Enter question : what is your name
 leeloo minai lekararibalaminaitchai ekbat de sebat

score:4
Enter question : what do you do
 i am a f***ing lawyer how we f*** bob

score: 7
Enter question : what day is it
 do he be a lot of red things

score:5
Enter question : are you happy
 i was gonna bring the back off of two cops debbie of the way i
 am getting a sick what were you

score:5
Enter question : are you sad
 saw your dad was here he is talking to you let me down

score:6
Enter question : what did you do today
 this is i have

score:3
Enter question : what
 i have got about i am not so much things real gave me to happen again

score:4
Enter question : why
 i am not asking you to marry me

score:7
Enter question : how
 you never disappointed me

score:8

average score: 5.6
```

```
Model 4
Enter question : how are you
 been better i read about what is been happening with you i should have called

score 9
Enter question : what is your name
 leeloo minai lekararibalaminaitchai ekbat de sebat

score:4
Enter question : what do you do
 i am a revolutionary

score:9
Enter question : what day is it
 time to find you one old mangs actually david

score:7
Enter question : are you happy
 keep back going back take more

score:6
Enter question : are you sad
 with us right we can see him

score:6
Enter question : what did you do today
 what i do is an orderly life you are an alcoholic i just say
that i am a big room in my way

score:6
Enter question : what
 you are kidding

score:7
Enter question : why
 i have selected you to return with me along with several other
specimens for study we will examine you and the others discover desirable features
to incorporate in our earthcolonizers

score:9
Enter question : how
 you never disappointed me

score:8

Average Score: 6.3
```

8 Experimental Results Continued

Quantitative Analysis	Model 1	Model 2	Model 3	Model 4
Score	1	4.8	5.6	6.3
Score(Model 2 Re-based)	-	1	1.167	1.3125
Scoring Scale Criteria:				
1. Comprehension level of output sentence				
2. Relevance of output sentence to the question				
3. Relative comparison of above criteria to baseline model				

The first model was used as a baseline but does not really do a great job of predicting anything. The Average score of 1 is extremely poor, as expected(however we would set this to 1 in any case when we use it as a base model). Our next models do progressively better jobs and we are able to get coherent semi coherent sentences out of the dialogue. We are able to get a relative 4.8 score for our second model, 5.6 for the third model and 6.3 for the fourth model. However it is must be noted that the baseline was extremely poor. Setting another baseline would probably lower the score of the second model. In the Quantitative Analysis done, we re-base the models and set the second model as the base. In conclusion it is promising that the models get progressively better as we increase data size and the number of epochs run. it can also be noted that model 3 performs better than model two with more conversations but fewer epochs. This is a good indication that increasing data size would most likely yield better results in the future than increasing the number of epochs run.

9 Conclusion and Future works

Conclusion, final thought and future works

- As of today our self developed human evaluated relative scale quantitative Turing Test seems to give us decent quantitative measure of relative performance. A problems with this method is that the relative scoring is performed individualistically by the scorer and will have bias in it. A problem with we might be having also is the baseline model we compare our other models to.I.e If the initial model is really terrible then all other models will score high in relation. We might want to pick a better model as a baseline; maybe give negative scores to models that perform worse than the new baseline model. Future works on the suggested evaluation method might be attempting to have a qualitative score rather than a quantitative one or perhaps using absolute scores. Another idea of work that could be done with this is once scoring a significant number of models, an ensemble learning model can be built on top of the many model variations with the greatest score for a specific question being used.
- Future improvements to our process would be to use more powerful hardware. It took a long time to train these models even with a limited numbers of conversations being used, Using a GPU for this purpose would be the next step to improve performance. Another step could be to try and run or redesign the model to run on a distributed system like spark to improve performance.
- In conclusion this was a very interesting project to explore and work on. I plan on exploring if my human evaluated relative scale quantitative Turing Test method can be be a topic of research and maybe try to see if publication is a possibility with it. I am excited to dig deeper into the world of NLP and chat bots and look forward to contributing to the field in the near future,

References

[1] How to build a State-of-the-Art Conversational AI with Transfer Learning by Thomas Wolf. (https://medium.com/huggingface/how-to-build-a-state-of-the-art-conversational-ai-with-transfer-learning-2d818ac26313)

[2] Importance of a Search Strategy in Neural Dialogue Modelling by Ilya Kulikov, Alexander H. Miller, Kyunghyun Cho, Jason Weston (http://arxiv.org/abs/1811.00907)

[3] Correcting Length Bias in Neural Machine Translation by Kenton Murray, David Chiang (http://arxiv.org/abs/1808.10006)

[4] Breaking the Beam Search Curse: A Study of (Re-)Scoring Methods and Stopping Criteria for Neural Machine Translation by Yilin Yang, Liang Huang, Mingbo Ma (https://arxiv.org/abs/1808.09582)

[5] Hierarchical Neural Story Generation by Angela Fan, Mike Lewis, Yann Dauphin (https://arxiv.org/abs/1805.04833)

[6] Language Models are Unsupervised Multitask Learners by Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (https://openai.com/blog/better-language-models/)

[7] The Curious Case of Neural Text Degeneration by Ari Holtzman, Jan Buys, Maxwell Forbes, Yejin Choi (https://arxiv.org/abs/1904.09751)

[8] Retrieve and Refine: Improved Sequence Generation Models For Dialogue by Jason Weston, Emily Dinan, Alexander H. Miller (https://arxiv.org/abs/1808.04776)

[9] The Second Conversational Intelligence Challenge (ConvAI2) by Emily Dinan et al. (https://arxiv.org/abs/1902.00098)

[10] Building a Conversational Chatbot with NLTK and TensorFlow by Bamigbade Opeyemi (https://heartbeat.fritz.ai/building-a-conversational-chatbot-with-nltk-and-tensorflow-part-2-c67b67d8ebb)

[11] Natural Language Processing (almost) from Scratch, 2011 by Ronan Collobert

[12] Lemmatization in Natural Language Processing (NLP) and Machine Learning by Sunny Srinidhi (https://contactsunny.medium.com/)

[13] A Survey on Evaluation Methods for Chatbots by Wari Maroengsit, Thanarath Piyakulpinyo, Korawat Polyiam and Suporn Pongnumkul (https://www.researchgate.net/publication/333524709$_{A}S{urvey}_{o}n_{E}{valuation}_{M}{ethods}_{f}{or}_{C}{hatbots}$)