Personality-based AI Chatbot

Team Number: 14

Team Members:

Ganesh Raj Kyatham (grk62)

Smrithi Shankar(4350)

Vidisha Kotnala (vk434)

INTRODUCTION

Chatbots have become ubiquitous in customer care, gaining even more popularity with the advent of modern language models. Despite their widespread use, they face significant challenges, including their:

1. tendency to favor responses with the highest likelihood,
2. a tendency for generating nonspecific answers, and
3. the lack of a consistent personality due to training on diverse dialogs with different speakers.

We aim to solve a few of these issues and make chatbots friendlier and more useful to interact with. For the confines of this project we will focus on the following goals with respect to improving chatbots:

1. to make the chatbot's replies more consistent to make the interactions more reliable.
2. to provide information specific to the personality that the chatbot is trained on, instead of defaulting to responses with the highest consensus
3. to create a more human-like interaction, by assigning a consistent personality to the chatbot.

Through these improvements, we aim to elevate the overall friendliness and utility of chatbots, fostering a more seamless and satisfying user experience.

For this project our goal is to build a chatbot that emulates the personality of the popular TV show FRIENDS character, Chandler Bing.

DATASET

The dataset utilized for this project is the Friends Corpus, encompassing conversations spanning 10 seasons of the popular American TV sitcom "Friends," which aired during the 1990s. The original dataset is available here.

The dataset is structured and processed using the Convokit library, which provides a convenient framework for managing conversational data. The format includes:

- **Speaker-level Information:** Characters are identified by their names, with an additional dummy user named "TRANSCRIPT_NOTE" for non-character utterances.
- **Utterance-level Information:**
    - **id:** Index in the format sAA_eBB_cCC_uDDDD (e.g., s01_e18_c05_u021).
    - **speaker:** Speaker's name (e.g., Monica Geller).
    - **conversation_id:** Id of the first utterance, assuming conversations begin at the start of a new scene.
    - **reply_to:** Id of the utterance being replied to (None if the utterance is the first in a conversation).
    - **timestamp:** Not available in the dataset.
    - **text:** Textual content of the utterance.

Metadata associated with utterances may vary by season and include character entities, emotion labels, tokenized text, transcript notes, tokens with notes, and caption information.

- **Conversation-level Information:**

Conversations represent scenes of the show. They are indexed by the id *sXX-eYY-cZZ*, where *XX* denotes the season (e.g. 01), *YY* denotes the episode (e.g. 01), *ZZ* denotes the conversation (e.g. 01).

We shall split the dataset into training, validation and test sets.

EXPERIMENTS WITH CONVERSATION MODELS

Hugging Face offers a range of conversational models suitable for various use cases. In our evaluation, we narrowed down our choices to gpt2 and blenderbot-400m-distillbot, opting to deploy them without any additional fine-tuning as a starting point. Our assumption was that, owing to their training on extensive open text datasets, these models might possess a small level of understanding of the TV show FRIENDS and its character, Chandler Bing.

Specifically, GPT-2 underwent pre-training on WikiText2, while Blenderbot was trained on publicly available internet text.

We conducted tests to check the models' knowledge of the TV show FRIENDS, and specifically Chandler Bing.

Example responses while tested for the knowledge of FRIENDS:

```
TEXT PROMPT ross, rachel, monica, chandler, phoebe, Joey were in the hit TV show F.R.I.E.N.D.S
 **********BLENDERBOT output**********
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
["<s> Wow, I didn't know that.  Ross was my favorite character on that show.</s>"]
**********GPT2 output**********
[{'generated_text': "ross, rachel, monica, chandler, phoebe, Joey were in the hit TV show F.R.I.E.N.D.S. - they've been in a mi
llion episodes since 2008.\n\n\n"}]


TEXT PROMPT so no one told you life was gonna be
 **********BLENDERBOT output**********
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
["<s> I didn't think so, but I guess I was wrong. I was so disappointed in myself.</s>"]
**********GPT2 output**********
[{'generated_text': 'so no one told you life was gonna be the same every day as they\'re now," said one former member. "And all
I can tell you is that one time last week, [David] told me, \'You should have told me this'}]
```

Example responses when prompted to respond like Chandler Bing.

```
TEXT PROMPT Ross, Monica, Joey and chandler are having a coffee at central perk. Ross says I just want to be married again and
rachel walks in wearing a wedding dress. Chandler says
 ***********BLENDERBOT output***********
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
["<s> I love Ross! He's such a great character. I wish I could dress up like him.</s>"]
***********GPT2 output***************
[{'generated_text': "Ross, Monica, Joey and chandler are having a coffee at central perk. Ross says I just want to be married a
gain and rachel walks in wearing a wedding dress. Chandler says he's not even sure how the divorce should be handled."}]


TEXT PROMPT Joey tribbiani gets locked in his own closet because of his stupidity, when chandler comes home to discover they ha
ve been robbed he
 ***********BLENDERBOT output***********
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
['<s> Joey Tribiani is an American singer, songwriter, and actor.</s>']
***********GPT2 output***************
[{'generated_text': 'Joey tribbiani gets locked in his own closet because of his stupidity, when chandler comes home to discove
r they have been robbed he gets mad enough to turn his back on Chantal and run around the house to see who he can'}]
```

Our findings revealed unsatisfactory outcomes.

For both the models 10 prompts were given to test the model's knowledge regarding FRIENDS and 10 prompts to make the model respond like Chandler Bing. Out of the total 10 prompts only 2 outputs were relevant and that was when tested on the knowledge about FRIENDS.

Fine-tuning these models using the FRIENDS corpus must yield better and more relevant results.

FUTURE STEPS

We plan to fine-tune our pre-trained gpt2 and blenderbot-400m-distillbot models on the FRIENDS corpus and specifically on the interactions of Chandler Bing with rest of the characters in order to train the model on his speaking style.

Further, we plan to use the BLEU (Bilingual Evaluation Understudy) and the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores to evaluate our model.

The BLEU score is mainly used in translation tasks, but based on several sources on the Internet, it is now a popular metric for generative models as well. It has shortcomings with regards to evaluating translation task performance, but we found that it serves as a reasonably good metric for generative chatbot applications. Using this score we check how much the words in the machine-generated responses appear in the human references, specifically in our case, responses by the same character to similar queries in the test set. This is precision-focussed.

In a similar fashion, we also intend to use the ROUGE score. There are 3 subtypes of the ROUGE score, namely ROUGE-N, ROUGE-L and ROUGE-S. ROUGE-N doesn't explain much as it simply measures the number of matching n-grams between the model-generated text and the human reference (again, in our case, the responses by the same character to similar queries in the test set). We speculate that ROUGE-L and ROUGE-S better capture a character's style and way of speaking, and hence feel they may be better suited to test the performance of our chatbot. We shall use all three in our project. On the whole, ROUGE is recall-based and evaluates how much the words in the human references appear in the model-generated output.

Additionally, we also came across papers that evaluate the performance of a personality-emulating chatbot based on personality tests used for humans. One such paper (cited in the References section) makes use of the OCEAN personality evaluation method to evaluate the performance of a personality-emulating chatbot. We plan to investigate this methodology further, possibly using other personality tests such as the MBTI (Myers-Briggs Type Indicator) framework.

REFERENCES

- GPT2 - huggingface
- Blenderbot-400m-distill huggingface
- GPT-3 Trained To Impersonate By: Alexander Castañeda, Patrick Brown, Rais Kazi, Landyn Moreno, Christian Tomah, Phillip Peng, Michael Hildner https://medium.com/@patrickbrown5530/gpt-3-trained-to-impersonate-e0a801810245
- A Persona-Based Neural Conversation Model Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, Bill Dolan https://arxiv.org/abs/1603.06155
- Personalizing Dialogue Agents: I have a dog, do you have pets too? Saizheng Zhang†,1, Emily Dinan‡, Jack Urbanek‡, Arthur Szlam‡, Douwe Kiela‡, Jason Weston‡ https://aclanthology.org/P18-1205/
- TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents: Thomas Wolf, Victor Sanh, Julien Chaumond & Clement Delangue https://arxiv.org/abs/1901.08149
- Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset Hannah Rashkin, Eric Michael Smith, Margaret Li, Y-Lan Boureau https://arxiv.org/abs/1811.00207
- Gaikwad, Susmit, "Chatbots with Personality Using Deep Learning" (2019). Master's Projects. 678. DOI: https://doi.org/10.31979/etd.w5wa-vujn https://scholarworks.sjsu.edu/etd_projects/678
- Xing, Y. and Fernández, R. (2018). *Automatic Evaluation of Neural Personality-based Chatbots*. [online] Association for Computational Linguistics, pp.189–194. Available at: https://aclanthology.org/W18-6524.pdf.