Shaamer Kumar
Ryusuke Suehiro

Creating a Realistic, Airline Chatbot

ABSTRACT

Current chatbots have four main issues: a lack of personalisation, contain misunderstandings, and have a lack of empathy[1]. Firstly, there's a big lack of personalization in their responses. Chatbots tend to offer very generic, unhelpful answers, failing to recognize and adapt to the unique circumstances and needs of a customers' concerns. Secondly, a significant challenge with chatbots is their propensity for misunderstandings and miscommunications. Due to common dependence on keyword recognition instead of more intelligent techniques, chatbots can easily misinterpret customer intentions or miss the nuances of human language. Finally, tying all these together is the issue that chatbots have a large lack of empathy. When facing personal issues, many chatbots respond in tones inappropriate for the corresponding situation, which is essential for a satisfactory customer service experience. All of these as a combination leave customers feeling undervalued and more frustrated from when they started, demanding an audience with a customer representative which ends up wasting everyone's time in the end. According to Zendesk's 2022 CX Trends report, around 60% of customers see disappointment when dealing with chatbots[2]. The goal of our project is to create a chatbot that adds empathy in responses specifically for professional use i.e the airline industry, and behaves as close to how a human representative would, providing appropriate responses according to the situation at hand.

[1] https://blog.bcwebwise.com/excessive-use-of-chatbots-can-lead-to-customer-dissatisfaction/
[2] https://www.cxtoday.com/speech-analytics/customers-frustrated-with-chatbots/

Shaamer Kumar
Ryusuke Suehiro

INTRODUCTION

For this project, we are using the dialoGPT[3] model from HuggingFace which is trained on 147M conversation-like text from sources such as Reddit, and is optimized from the existing GPT-2 model to cater more towards conversational texts and generating human responses. To enhance the conversational abilities of the model to provide a more efficient interaction experience for angry passengers seeking assistance from companies like United, Delta, Spirit, etc. GPT-2 is already known for its proficiency in generating human-like responses, and we will fine-tune this using more specialized datasets tailored to customer support in the aviation industry. One potential goal we also want to optimize on is generating text that will anger customers less by being too robotic, which we can measure by similarity to automated response bots which anger customers. The data we have found for this are from customer conversations[4], and general conversations[5]. Our goal is to fine tune the dialoGPT model with this dataset, ensuring that responses are personal, professional, and empathetic to a customer's needs.

FINE-TUNING PROCESS AND RATIONALE

Since our goal is to finetune our model to output responses that most closely resemble a human, and to be as different as possible from automated bots that currently exist in their websites, we **decided to make custom loss functions that penalize responses that are very different from humans and most similar to automated bots**. Although we recognize that our objective is not necessarily fine tuning our model to output the most useful or accurate responses, we believe this better aligns with our goal and the needs of this industry, due to the big prevalence of poor automated bots angering customers away from airlines. Furthermore, with this loss function, we

[3] https://huggingface.co/docs/transformers/model_doc/dialogpt
[4] https://huggingface.co/datasets/Kaludi/Customer-Support-Responses/tree/main
[5] https://www.kaggle.com/datasets/kreeshrajani/3k-conversations-dataset-for-chatbot/

can easily fine-tune our model using text-response similarity metrics like BLEU and METEOR instead of having to manually label the quality of responses. Additionally, our pretrained model, Dialo-GBT, has already been tested on human-labeled data, and ensures that our model passes the "human-eye test," being manually labeled by humans as a better or equivalent response than a human response to conversational prompts when both were shown.

We test on 6K multi-ref dataset from Reddit. The results are summarized in below

Experiment	NIST2	NIST4	BLEU2	BLEU4	METEOR	ENT-4	DIST-1	DIST-2	Avg. Len
Human response	3.41	4.25	17.90%	7.48%	10.64%	11	14.50%	63.00%	13.1
DialoGPT 117M	2.39	2.41	10.54%	1.55%	7.53%	10.78	8.60%	39.90%	12.8
DialoGPT 345M	3	3.06	16.96%	4.56%	9.81%	9.13	6.80%	26.30%	12.2
DialoGPT 762M	2.84	2.9	18.66%	5.25%	9.66%	9.72	7.76%	29.93%	11.2

Relevance: A and B, which one is more relevant to the source prompt.

System A	A Wins (%)	Ties (%)	B Wins (%)	System B
DialoGPT 345M	2671 (45%)	513 (9%)	2816 (47%)	Human responses
DialoGPT 345M	3281 (72%)	394 (9%)	882 (19%)	PersonalityChat
DialoGPT 345M w/ MMI	**2871** (48%)	522 (9%)	2607 (43%)	Human responses

Informativeness: A and B, which one is more contentful and informative.

System A	A Wins (%)	Ties (%)	B Wins (%)	System B
DialoGPT 345M	2722 (45%)	234 (4%)	3044 (51%)	Human responses
DialoGPT 345M	3490 (77%)	206 (5%)	861 (19%)	PersonalityChat
DialoGPT 345M w/ MMI	**3011** (50%)	234 (4%)	2755 (46%)	Human responses

Datasets and APIs used in Fine-tune Process:

- Airline Customer Service Transaction Log Dataset[6]
- Generic Customer Service Question and Answer Dataset[7]

Custom Loss Function:

- Use BLEU, METEOR, and CIDEr score API[8] and Heavy Penalization for responses with high score with the automated bot response and low score with human sample response on dataset. We assign the costs in an exponential way where responses deviating heavily from the human responses are penalized much more than moderately deviating responses.

TODOS:

- Model Functionality : At the moment, we are facing issues with training our initial model on the new dataset. So a big todo here is to find out the bug and why things are not working the way we expect it to with the new training data we procured from online. We

[6] https://huggingface.co/datasets/Kaludi/Customer-Support-Responses/tree/main
[7] https://www.kaggle.com/datasets/kreeshrajani/3k-conversations-dataset-for-chatbot/
[8] https://gist.github.com/kracwarlock/c979b10433fe4ac9fb97

also just acquired access to a server with a lot of GPU that should help our training process with these big datasets we want to fine tune with.

- Enhanced Data : It would be optimal for us to add data with Queries directly relating AIRLINE issues, not just general customer concern questions. This will teach our chatbot how to deal with customer issues, but not necessarily those for that of airlines. We have a dataset with a list of airline queries on standby, however, those have no corresponding responses. We are also thinking of tagging common customer issues such as “Refunds,” “Cancellation,” Information Query,” and putting additional weights to responses relating to those in the fine-tuning process
- Loss Function Enhancement Testing : Our current plan is to use cross-entropy loss during training, seeing that our model outputs something as close to what is expected out of customer response type answers. One thing to add would be to modify this cross entropy to include an additional term which adds more loss for very long responses or very simple responses. Additionally, we want to make sure negative phrases are not included at all and we can incorporate this into our loss function. So as of now, we have the code written for this loss function, but we just need to test it out to see its efficacy and potentially add more words and negative phrases to it.
- Visualizations and Revision: Upon completion of above, we will conduct multiple iterations of retraining, validating, and evaluating our output where the test set results measured by the similarity scores given by the BLEU, METEOR, etc are used comprehensively to see what our model may be overfitting on.

Shaamer Kumar
Ryusuke Suehiro

GitHub Link: https://github.com/ShaamerKumar/cs182FinalProj