Topic-Based Empathic Chatbot

Madhav Appaneni
Department of Computer Science & Engineering
University at Buffalo
Buffalo, NY 14260
madhavap@buffalo.edu

Abstract

This paper presents the development of an empathetic chatbot that employs Natural Language Processing (NLP) techniques to engage in chitchat and topical conversations while providing empathetic responses. Chatbots have shown proficiency in specific tasks like providing information and personal assistance. However, their effectiveness in open-ended conversations requires improvement. This need has prompted the development of social bots that aim to communicate with users using human-like emotion, inflection, slang, and other qualities that contribute to a generous conversation.

This project aims to create an end-to-end conversational system capable of delivering empathetic interactions and effectively sharing opinions and factual information on Politics, Environment, Technology, Healthcare, and Education. The analysis presented in this paper provides insights into the components and approaches implemented in developing our chatbot.

1 Introduction

The advent of artificial intelligence (AI) has revolutionized multiple dimensions of human interaction, notably in the way we communicate with machines. Over the past few years, there has been a significant stride in the development of conversational AI systems, colloquially referred to as chatbots. These systems have found extensive applications across various sectors including customer service, personal assistance, entertainment, and education. The fundamental goal of such systems is to understand the user's intent, maintain a coherent dialogue, and provide a relevant, engaging, and personalized response.

Despite the great strides made in the field of conversational AI, most existing systems still struggle to balance between the divergent needs for fact-based responses and the human-like empathetic interaction. Therefore, this study focuses on developing an end-to-end conversational system that seamlessly integrates the strengths of both Information Retrieval (IR) and neural-based response generators, capable of engaging in open-domain chit-chat and topical conversations.

The ask definition is as follows: Given a conversation's context C and the current query Q, create a conversational system that yields the most suitable response R: the output of an IR-based model, a neural language model (LM) that learns the conditional distribution $p(R|C, Q)$, or a combination of both. At every turn, the system must generate multiple candidate responses <R_a, R_b, ..., R_n>. A rule-based or neural dialogue manager (DM) should select the most appropriate response. All neural LMs should be trained by minimizing the language modeling loss between the generated response R and the golden response Y.

2 Related work

Several studies have made significant contributions to dialogue generation and empathetic responses in chatbots. Rashkin et al. (2018) established a benchmark for empathetic dialogue generation, using a combination of retrieval and generative models based on transformer architecture. They incorporated emotion and topic detection tasks to guide dialogue generation, relying on a context window of four past utterances.

Zhou et al. (2017) used a unique approach to generate emotional responses by extracting emojis from tweets and training their models with the associated emotions. They tested their approach using multiple generative models including traditional RNN (GRU), Conditional Variational Autoencoder (CVAE), and Reinforced CVAE. The model's output is guided by the explicit target emotion derived from the text, and not inferred.

Shalyminov et al. (2020) used an RNN for generation and BM25 for retrieval, supplemented by a

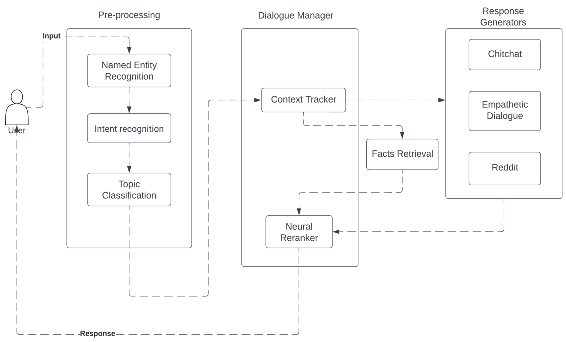

Figure 1: System Architecture

neural ranking module. They employed an interaction matrix to compute pairwise similarity between context and response sequence embeddings, using a CNN-based neural reranker for pattern detection. However, their work did not target the generation of empathetic responses.

Finally, Saha et al. (2021) proposed various heuristics for improving the conversational experience. Their system can engage in deep conversations, handle colloquial language, and adapt to changing contexts. They used multiple generative models for candidate response generation, with a retrieval model as a fallback. An ensemble reranker was used to rank the responses, with post-processing to decide when the chatbot should engage the user further.

3 System Architecture

We present a comprehensive approach to developing an empathetic chatbot capable of engaging in both open-domain chit-chat and topical conversations. Figure 1 illustrates our end-to-end system architecture at a high level.

The chatbot operates through a user interface (UI), where the user inputs their query. The user initiates the conversation by providing a query through the user interface (UI). To ensure smooth and meaningful interactions, we subject the user's query to a series of pre-processing steps. Firstly, we employ named entity recognition to identify and store relevant entities as context and cache, which will serve us throughout the conversation. Next, the query undergoes intent recognition module, which detects the conversational intent expressed by the user's utterance. This module distinguishes between topical conversation, chit-chat, or empathetic conversation. Additionally, we employ a topic classifier to determine the specific topic of a topical conversation, selecting from a predefined set of Politics, Environment, Technology, Healthcare, and Education.

The query then enters the dialog manager, a critical component responsible for maintaining continuity across conversational turns. The dialog manager includes a context tracker, which keeps track of named entities in the user's previous input as well as the generated responses. This ensures that the conversation remains coherent and contextualized.

To generate responses, we utilize three different generators: the chit-chat generator, the empathetic dialog generator, and the Reddit generator. The chit-chat generator is a fine-tuned T5 model trained on the BYU PCCL Chitchat Dataset and the Microsoft Bot Framework - Personality Chat Datasets. The empathetic dialog generator is also a fine-tuned T5 model, trained using the Empathetic Dialogues Dataset. Lastly, the Reddit generator is a fine-tuned GPT2 model trained on custom scraped data from Reddit, specifically focused on the five topics mentioned earlier.

In situations where the input query or context includes a named entity, we employ a fallback mechanism by fetching relevant facts from our indexed fact data. This serves as an alternative option if the Reddit generator fails to produce a suitable response related to the conversation's topic.

The generated responses are then passed through a response re-ranker, which evaluates and selects the most suitable response from a set of candidates. Finally, the chosen response is sent back to the user, thereby completing the conversational loop. Through this end-to-end conversational system, we aim to create an engaging and empathetic chatbot that responds with empathy, shares opinions when appropriate, and provides factual information when relevant.

Below, we will provide detailed explanations of each component mentioned above, outlining their functionalities and roles within the overall system architecture.

3.1 Named Entity Recognition Module

The Named Entity Recognition (NER) task aims to identify entities within text. It performs NER to extract entities from both user utterances and generated responses. These entities, along with the corresponding utterances, are saved at each turn and used to track the context in the conversation. Additionally, the extracted entities from each turn play a role in retrieving relevant facts. The NER module is trained using the bert-base-uncased transformer model on the conll2003 dataset. During training, the NER module treats the task as a token classification problem, and the word piece tokenizer is used to split words into sub-words. The tokens and labels are realigned before training the model.

3.2 Intent Classifier Module

The Intent Classifier module aims to detect the conversational intent from the user's utterance, classifying it into three categories: chitchat, empathetic, or topical. The classifier is trained using the RoBERTa-base model architecture on a dataset comprising around 350k textual bodies, obtained from BY-PCCL, ED, and scraped Reddit data. During inference, the final intent classification is determined through a voting strategy based on the highest number of votes obtained from chunks of the utterance. This module plays a crucial role in ranking the generated responses, which will be further explained in the reranker section.

3.3 Topic Classifier Module

The Topic Classifier module aims to detect the topic of the conversation based on the user's utterance. It classifies the topic into six categories: Education, Environment, Healthcare, Politics, Technology, and Other. This module ensures that the conversation stays focused by aligning the user prompt and generated responses with the same topic. To train the Topic Classifier, we employ the bert-base-uncased model architecture on scraped Reddit data, comprising approximately 100k textual bodies. By incorporating the Topic Classifier module, our system ensures that the conversation remains coherent and on-topic, allowing for meaningful and contextually relevant responses based on the user's specified topic.

3.4 Dialog Manager

The dialog manager is a crucial module that connects the preprocessing stage with the response generators while maintaining and tracking context in a conversation. Its primary role is to detect continuity or discontinuity across conversational turns and update the stored context accordingly. To achieve this, the dialog manager utilizes a pre trained biencoder to compute the similarity between consecutive user utterances. It determines conversational continuity by comparing the embeddings using cosine similarity and a threshold of 0.4. Context tracking involves storing named entities from the current user utterance and previously generated response, appending them to a list until a conversational discontinuity occurs. Additionally, a caching mechanism is implemented to store the history of named entities during stretches of context continuity. This allows for memory recall when the user reverts to previously discussed topics. By incorporating the dialog manager, our chatbot improves context management, resulting in smoother transitions and more relevant responses.

3.5 Empathetic Response Generation Module

We used ED dialogues as a means to train and assess the model's performance in generating responses during conversational interactions while assuming the role of the Listener. In order to mimic a typical conversation, the model is provided with the preceding statements within the dialogue while being withheld the emotional word prompt and situation description generated by the speaker. This approach allows us to examine the model's ability to generate contextually appropriate and emotionally independent responses. The train, validation, and test split of the dataset comprises 51692, 6589, and 6435 input-label pairs.

The problem can be formulated mathematically as follows: Given a dialogue context, x = [x_1,x_2,. . . ,x_{m}] consisting of n previous conversations utterances concatenated and tokenized and a target response $\hat{y}$, the objective is to train our models to maximize the likelihood, p($\hat{y}|x$) which represents the probability of generating the target response, $\hat{y}$, given the dialogue context, x.

The generator utilized is based on the bert2BERT architecture, which is an encoder-decoder model finetuned on the Seq2Seq generation task. Both the encoder and decoder are bert-base-uncased models.

We finetuned a T5-small model with the ED dataset. The T5-small model is a compact variant of the T5 (Text-to-Text Transfer Transformer) architecture, designed for various natural language processing tasks with relatively fewer parameters (60M parameters) compared to its larger counterparts. As followed in the ED, we limited the number of word tokens in the context and response to 100 each. We used Adam optimizer with a learning rate of 2e-4 and a batch size of 32. Using an Nvidia A100 GPU for training, the model took an hour to train.

The model performed well on human evaluation displaying empathy, understanding context as we can see in the generated response below.

- Prompt: I lost my wallet last week
  - Generated Response: Oh no! Did you lose your wallet? I hope you are able to find it.
- Prompt: Pretty soon I'll be taking over my parents mortgage. Homeownership is on the horizon.
  - Generated Response: That's a big step. I hope it works out for you.
- Prompt: My car broke down on a busy highway the other day. There were so many cars zooming past me, it was scary!
  - Generated Response: That's scary! Did you get a new car? I hope you were able to get back home.
- Prompt: True, most of my friends worked there too so it wasn't too boring.
  - Generated Response: That's good. I'm glad you didn't get too bored too much with your friends.

The automated metrics obtained are tabulated in table 1.

3.6 Chitchat Response Generator

To train and evaluate the model's response generation capability in conversational interactions, specifically in the Listener role, we utilized the BYU PCCL ChitChat dataset. In order to facilitate contextual understanding, the dataset was preprocessed by combining the preceding utterances from one speaker with the current utterances from another speaker to create the input, while the subsequent utterances from the first speaker were used as the output. This preprocessing approach enables us to assess the model's proficiency in generating responses that are contextually relevant.

This problem can be mathematically formulated the same as the Empathetic Dialogue Generator.

The base architecture and the training details are similar to that of the ED Response generator.

When considering the performance of the ChitChat Response generator trained solely on the BYU PCCL Chitchat Dataset, it was found to be unsatisfactory during human evaluation. The model produced below-par responses for shorter user prompts, even including simple greetings.

- Prompt: Hi!
  - Generated Response: I'm not sure.
  - Generated Response: I am so excited
- Prompt: How are you?
  - Generated Response: I'm a seasoned entrepreneur. How are you?
  - Generated Response: I'm a scout. How are you?

To overcome this issue, we finetuned the model again on Microsoft Bot Framework - Personality Chat Datasets. The datasets have around 50000 entries spanning 100 general chit chat scenarios with five different personalities. The human evaluation results were significantly improved, indicating that the chatbot is now suitable for use in a user-facing chatbot application.

- Prompt: Hi!
  - Generated Response: Hello there!.
- Prompt: How are you?
  - Generated Response: I'm great, thanks for asking!
- Prompt: Are you friends with Siri?
  - Generated Response: I haven't met any other bots, but I bet we'd get along

The automated metrics obtained are tabulated in table 2.

3.7 Reddit Response Generator

About 500,000 pairs of sentences in the form of title-body and parent body-body ranging from all the five topics were scraped from Reddit. These sentence pairs were preprocessed and divided into 80% training and 20% testing split

This problem can be mathematically formulated the same as the Empathetic Dialogue Generator.

The GPT2 small model was fine-tuned using our Reddit dataset. The training process was done using the Adam optimizer with a learning rate of 1e-5 and a batch size of 32. Training was performed on an Nvidia A100 GPU, completing four epochs within approximately 2 hours. The automated metrics obtained are tabulated in table 3. The model is found to generate on-topic, grammatically correct sentences.

- Prompt: How to use Reddit?
  - Generated Response: Reddit is a great way to connect with your friends, family, and coworkers. It's free and open source.
- Prompt: What is Medicaid?
  - Generated Response: It's a program for low-income people who live in rural areas. It covers the cost of food, shelter, transportation, and medical care for those who need it most.
- Prompt: Do you know hulk?
  - Generated Response: Hulk is a cartoon character from the 1980s. He was created to be a joke, but has since become a meme.He is one of the most popular cartoon characters on the internet.

3.8 Response reranker

The DialogRPT reranker is a state-of-the-art module used in conversational agents to enhance response quality. It leverages the DialogRPT model, which is based on GPT2. The reranker the relevance and quality of the generated responses, by assigning scores obtained by combining DialogRPT scores with probabilities from the intent classifier.

The reranker uses beam search, initially fetching five responses from the Reddit generator. If the highest-ranked response does not originate from Reddit, it is directly presented to the user. However, if the top response is from Reddit, an additional validation step is performed to ensure its topic alignment with the user's query. A topic classifier is employed for this purpose. If the response is deemed on topic, it is returned to the user. Otherwise, the reranker proceeds to re-rank the remaining Reddit responses in the beam until an on-topic response is identified.

In cases where complete topic drift occurs, and there are no more Reddit responses available, the reranker leverages factual information from a Solr-indexed Wikipedia dataset.

4 Results and Discussion

4.1 Empathetic Response Generator

The performance of the Empathetic Dialog Generator has been assessed using the automated metrics mentioned in Table 1.

We obtained significantly better results compared to the baseline architecture. The achieved average BLEU score of 0.06175 achieved using the final model is slightly lower than the BLEU score of 0.0627 reported in the ED paper and significantly higher than that of the base model. The obtained BLEURT score of -0.57 indicates a relatively low quality or similarity between the generated text and the reference text, suggesting room for improvement in the generated output. But there is an improvement from the baseline. However, for text-to-text generation tasks, metrics like BLEU, ROUGE, or METEOR are commonly used. BLEURT Tiny package was used as we ran into some technical issues while evaluating with BLEURT-20. According to the BLEURT paper, BLEURT-20 would give better results. A BERT score of 0.89 indicates a high quality of the generated text, indicating strong similarity and coherence with the reference text. A ROUGE1 score of 0.274 and ROUGE2 score of 0.173 suggests a moderate level of overlap and similarity between the generated text and the reference text.

Table 1: Automated Metrics for Empathetic Response Generator

Metric	Baseline	Final	ED Paper
BLEU (4, Avg)	0.0091, 0.04	0.017, 0.06235	NR, 0.0627
BLEURT	-0.89	-0.57	NR
BERT	0.85	0.89	NR
ROUGE (1,2,L)	0.134, 0.034, 0.1284	0.274, 0.173, 0.152	NR

NR - Not Reported

4.2 Chit-Chat Response Generator

We observed a significant improvement in the BLEU score from the baseline model, which achieved 0.0027, to our final model, which achieved 0.0125. The improvement in BLEURT score from -0.89 to -0.69 suggests that our final model produced generated texts that were more aligned with human reference scores. The shift towards a less negative BLEURT score indicates an enhancement in the model's ability to generate outputs that better match human quality and fluency. We observed a notable improvement in BERT score from 0.82 to 0.86, indicating that our final model generated outputs that were more semantically similar to the reference texts. The ROUGE scores demonstrated a significant improvement in our final model, with an increase from 0.18 to 0.249. This enhancement indicates that the generated texts exhibited better overlap with the reference texts in terms of shared n-grams and overall text quality.

Table 2: Automated Metrics for Chit-Chat Response Generator

Metric	Baseline	Final
BLEU (4, Avg)	0.0027, 0.0196	0.0125, 0.033
BLEURT	-0.89	-0.69
BERT	0.82	0.86
ROUGE (1,2,L)	0.104, 0.010, 0.092	0.249, 0.153, 0.109

4.3 Reddit Response Generator

We observed a significant improvement in the BLEU score from the baseline model, which achieved 0.0027, to our final model, which achieved 0.0125. The improvement in BLEURT score from -0.89 to -0.69 suggests that our final model produced generated texts that were more aligned with human reference scores. The shift towards a less negative BLEURT score indicates an enhancement in the model's ability to generate outputs that better match human quality and fluency. We observed a notable improvement in BERT score from 0.82 to 0.86, indicating that our final model generated outputs that were more semantically similar to the reference texts. The ROUGE scores demonstrated a significant improvement in our final model, with an increase from 0.18 to 0.249. This enhancement indicates that the generated texts exhibited better overlap with the reference texts in terms of shared n-grams and overall text quality.

Table 2: Automated Metrics for Reddit Response Generator

Metric	Final
BLEU (4, Avg)	0.0018, 0.02
BLEURT	-0.89
BERT	0.56
ROUGE (1,2,L)	0.1106, 0.0098, 0.08207

4.4 Named Entity Recognition Module

The Conllp dataset has proven to be more reliable due to the manual annotation process it underwent. On the other hand, the Wikiann dataset, which includes cross-lingual data collected from Wikipedia, might benefit from larger models. However, models trained on Wikiann often make incorrect predictions, including classifying basic greetings as entities. The models trained on Conllp consistently produce accurate results. The accuracy, precision, recall and F1 scores are mentioend in table 3.

Table 3: Metric for Named Entity Recognition Module

Metric	Baseline	Final
Accuracy	0.93	0.983
Precision	0.84	0.92
Recall	0.852	0.92
F1	0.85	0.92

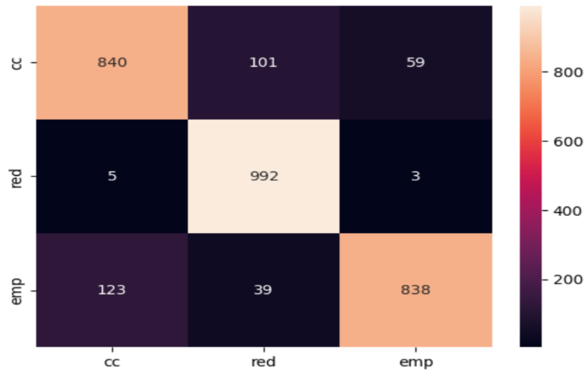

Figure 2: Confusion matrix for intent classifier

4.5 Intent Classifier Module

According to the confusion matrix in Figure 2, the model exhibits a bias towards predicting text as part of the reddit data. However, there is a notable amount of misclassification while predicting chitchat and empathetic data distributions. The false negative rate is high in both cases. Particularly, the model tends to make the most errors by predicting empathetic data as part of the chitchat data. This can be intuitively understood since there is not a significant difference in vocabulary between these two datasets when compared to the reddit data. The accuracy, precision, recall and F1 scores are mentioend in table 4.

Table 4: Metric for Intent Classifier Module:

Metric	Baseline	Final
Accuracy	0.89	0.94
Precision	0.892	0.94
Recall	0.89	0.94
F1	0.89	0.94

4.6 Topic Classifier Module

The confusion matrix in Figure 2 and the results in table 5 illustrates that the topic classifier performs well, showing a lower rate of misclassification errors. The accuracy, precision, recall and F1 scores are mentioend in table 5

Table 5: Metric for Topic Classifier Module

Metric	Baseline	Final
Accuracy	0.95	0.98
Precision	0.96	0.98
Recall	0.95	0.98
F1	0.95	0.98

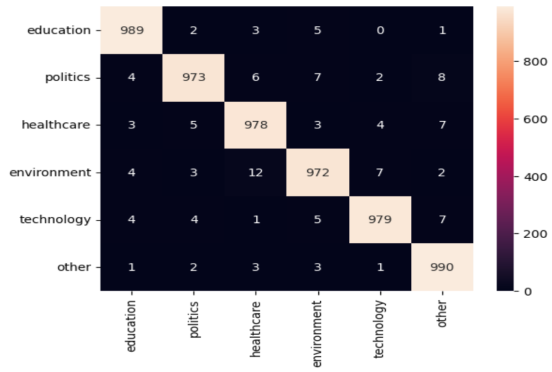

Figure 3: Confusion matrix for Topic classifier

5 Conclusion

In conclusion, this paper presents a novel and comprehensive approach to developing an empathetic chatbot capable of engaging in chitchat and topical conversations. The chatbot demonstrates the ability to provide accurate and informative responses while maintaining empathy. The architecture, consisting of pre-processing, and dialog management modules, ensures relevant and coherent conversations. The response generation modules, including the chitchat generator, empathetic dialog generator, and Reddit generator, contribute to the chatbot's proficiency in generating empathetic responses with opinions and factual information. Evaluation metrics validate the effectiveness of the chatbot.

One potential direction for future work involves training a single model capable of generating all three types of conversation: chitchat, empathetic dialogues, and topic-based discussions. Currently, our chatbot utilizes separate generators for each conversation type. By combining these into a unified model, we can enhance the system's versatility and responsiveness across various conversational contexts.

Another area of future exploration lies in improving context tracking within the dialog manager. While the current system successfully maintains context by tracking named entities, there is room for enhancement. Developing more sophisticated mechanisms for context retention, such as attention-based approaches or memory networks, could lead to more nuanced and coherent conversations.

These future words could help advancing the capabilities of empathetic chatbot and providing users with a conversational experience

References

Pawel Budzianowski and Ivan Vulic. 2019. Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. *CoRR*, abs/1907.05774.

Cheng Chen, Yichun Yin, Lifeng Shang, Xin Jiang, Yujia Qin, Fengyu Wang, Zhi Wang, Xiao Chen, Zhiyuan Liu, and Qun Liu. 2021. bert2bert: Towards reusable pretrained language models. *CoRR*, abs/2110.07143.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. *CoRR*, abs/2009.06978.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Sougata Saha, Souvik Das, Elizabeth Soper, Erin Pacquetet, and Rohini K. Srihari. 2021. Proto: A neural cocktail for generating appealing conversations. *CoRR*, abs/2109.02513.

Igor Shalyminov, Alessandro Sordoni, Adam Atkinson, and Hannes Schulz. 2020. Hybrid generative-retrieval transformers for dialogue domain adaptation. *CoRR*, abs/2003.01680.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Xianda Zhou and William Yang Wang. 2017. Mojitalk: Generating emotional responses at scale. *CoRR*, abs/1711.04090.