Next Utterance Prediction for Mental Health Counseling

Abhay Shakya
IIIT Delhi
New Delhi, India
abhay24108@iiitd.ac.in

Chaitanya Lakhchaura
IIIT Delhi
New Delhi, India
chaitanya24027@iiitd.ac.in

Ayush Kumar Verma
IIIT Delhi
New Delhi, India
ayush24025@iiitd.ac.in

Abstract

This project focuses on developing an NLP-based system designed to predict the next appropriate utterance in mental health counseling conversations. The objective is to generate contextually relevant, empathetic, and emotionally supportive responses to enhance therapeutic dialogues. By utilizing real-world counseling dialogue data, the system is trained to predict the next conversational turn in such sensitive exchanges. The generated responses are evaluated using the BLEU and BERT scores to ensure their fluency and contextual alignment. This system aims to assist mental health professionals in improving the flow and emotional relevance of their conversations, providing a helpful tool for managing complex dialogues. Our results indicate that the model can effectively generate responses with semantic accuracy, though further improvements are required for optimal performance. The project paves the way for further exploration into AI's role in mental health support, highlighting the potential for conversational AI in assisting therapeutic interactions.

1 Introduction

Mental health counseling plays a crucial role in helping individuals manage and improve their emotional well-being. In these sensitive exchanges, therapists (T) and patients (P) engage in meaningful dialogue, where therapists must respond with care, understanding, and empathy. However, one of the main challenges in such conversations is ensuring that responses are not only relevant but also emotionally supportive and contextually appropriate, particularly given the emotional complexity involved.

Predicting the next utterance in these conversations can be challenging, as it requires a deep understanding of the emotional tone, context, and flow. Ensuring that AI-generated responses are natural and aligned with therapeutic goals is a key challenge. Yet, with recent advancements in conversational AI, there is growing potential for AI to assist mental health professionals by suggesting the next appropriate step in the conversation or improving the overall flow of the interaction.

This project aims to develop an NLP-based system capable of predicting the next appropriate utterance in mental health counseling conversations. The goal is to create a model that generates responses therapists could use, ensuring that the responses are contextually relevant, empathetic, and emotionally supportive. By training the system on real counseling dialogues, we hope to develop a model that delivers helpful and coherent responses, evaluated using metrics like BLEU for n-gram overlap and BERTScore for semantic similarity. These evaluations will ensure that the AI-generated responses align with human-like interactions, both in terms of relevance and emotional support.

The motivation behind this project is to explore how AI can improve mental health counseling by helping therapists manage emotionally complex dialogues more efficiently, while maintaining empathy and context awareness. As studies have shown, AI-powered chatbots already play a role in mental health care, particularly by providing emotional support and offering conversational practice. This project builds on this foundation by enabling therapists to navigate challenging topics with greater ease, fostering supportive and contextually appropriate conversations, ultimately enhancing the quality of mental health care provided.

2 Related Work

The task of predicting the next response in a mental health counseling conversation is quite unique and requires specialized approaches. Several studies have focused on improving dialogue systems for this task by using advanced NLP techniques.

One important model in this area is READER,

which combines transformer-based architectures with reinforcement learning to predict the next dialogue-act and generate appropriate responses. READER uses Proximal Policy Optimization (PPO) to guide the model in producing responses that maintain the flow of the conversation. This is crucial in mental health counseling, where responses must be aligned with therapeutic goals. The model performs well on datasets like HOPE, a counseling conversation dataset, and outperforms other baselines across various evaluation metrics such as ROUGE, BLEU, and BERTScore.[3]

Another relevant study is ConSum (Srivastava et al., 2022), which is focused on counseling summarization. This model filters out irrelevant utterances and generates summaries using mental health knowledge, which is helpful in extracting key dialogue components in therapy conversations. The methods used in ConSum for identifying important dialogue elements are very much applicable to the task of predicting the next utterance in counseling conversations.[4]

Additionally, CARE (Zhong et al., 2020) proposes a model that combines commonsense knowledge with emotional context to generate responses that are both contextually relevant and emotionally appropriate. This model improves the quality of responses by balancing rationality and emotion, which is crucial in therapeutic settings where emotional understanding plays a key role in response generation.[3]

These studies highlight the importance of incorporating both contextual understanding and emotional intelligence in generating responses for mental health counseling. By using techniques like dialogue-act prediction, reinforcement learning, and domain-specific knowledge, models like READER, ConSum, and CARE are pushing the boundaries of how virtual assistants can effectively participate in mental health counseling conversations.

3 Methodology

The task aims to predict the next utterance in a mental health counseling dialogue, utilizing transformer-based models, specifically BART and GPT-2, for sequence generation. Both models were fine-tuned with specific modifications to better handle dialogue structure, speaker differentiation, and response generation.

The BART model, a sequence-to-sequence architecture, is used for generating the next utterance. The input and target text are tokenized using the BartTokenizerFast from Hugging Face, where the special separator token [SEP] is removed and the sequences are padded or truncated to a maximum length of 1024 tokens. The custom preprocessor handles this, ensuring the input and target text are formatted correctly.

A crucial modification to the BART model is the inclusion of speaker-specific embeddings. The therapist and patient are differentiated using the utterer labels, T: for therapist and P: for patient. These labels are encoded into token embeddings, which are then added to the model's input embeddings. This allows the model to distinguish between the two speakers in the conversation. The NextUttPreprocessor class creates these embeddings by identifying the speaker's tokens in the input sequence and generating a corresponding mask.

Additionally, the BART model's positional embeddings were expanded from the default 1024 tokens to 2048 tokens. This adjustment was necessary to handle longer dialogues in counseling sessions, ensuring that the model can process and retain information from extended conversations.

Similar to BART, the GPT-2 model is fine-tuned to generate responses based on the preceding conversation. The input and target text are concatenated into a single sequence, and the text is processed similarly to the BART model using the GPT2TokenizerFast. This tokenizer is configured to pad sequences to a maximum length of 1024 tokens, with the [SEP] token removed. The model is trained using the Trainer class from Hugging Face, and the preprocessed text is input into the model for training.

In GPT-2, the speaker-specific embedding mechanism is also implemented, though differently from BART. The text sequences for each speaker are combined into a single input, allowing GPT-2 to generate a response based on the entire conversation history. Like BART, the model is fine-tuned with a batch size of 8, a learning rate of 1e-5, and trained for 10 epochs. The evaluation is done using BLEU and BERT scores, ensuring that the generated responses are semantically relevant.

The models are trained using the Trainer class with specific training arguments that manage batch size, evaluation strategy, and other hyperparameters. For both models, the training is performed on a custom dataset containing therapy dialogues, with input and target sequences. The preprocessed

dataset is split into training and validation sets, and the training procedure is carried out over 10 epochs with a learning rate of 1e-5.

During training, the model computes a loss based on the difference between the predicted and actual target tokens, which is minimized through backpropagation. For evaluation, BLEU and BERT scores are used to assess the quality of the generated responses. BLEU measures the n-gram overlap between the generated and reference responses, while BERT score assesses semantic similarity by comparing BERT embeddings of the generated and reference text.

1. Utterer Identification: Both models integrate a custom speaker embedding mechanism. The NextUttPreprocessor class identifies the therapist and patient in the conversation and assigns them unique utterer IDs, which are then added to the input embeddings. This helps the models to recognize which participant is speaking and generate contextually appropriate responses.

2. Positional Embedding Adjustment: The BART model's positional embeddings were increased from the default 1024 tokens to 2048 tokens using the function. This allows the model to handle longer sequences of dialogue, which is crucial for maintaining the context of extended conversations in therapy sessions.

3. Text Preprocessing: The preprocessing involves removing unnecessary special tokens (like [SEP]) and padding or truncating sequences to the model's maximum token length. For GPT-2, the input and target texts are concatenated, whereas for BART, the input and target texts are kept separate. The preprocessing ensures that the input is consistently formatted for both models.

Both BART and GPT-2 were fine-tuned with the necessary modifications to handle speaker differentiation and longer dialogues in mental health counseling. The inclusion of speaker-specific embeddings and the extension of positional embeddings allows the models to better process and generate appropriate responses in a counseling context. The training and evaluation strategies, including the use of BLEU and BERT scores, ensure that the models are producing high-quality, contextually relevant responses. These approaches contribute to improving the quality of dialogue-based systems in mental health applications.

4 Dataset

The dataset used in this project consists of therapy session conversations between therapists (T) and patients (P). The data is structured for supervised learning tasks in text generation, particularly for therapeutic dialogue systems.

- Dataset: 4008 rows with `input_text` and `target_text` features.
- Dataset: 968 rows with `input_text` and `target_text` features.
- Dataset: 576 rows with `input_text` and `target_text` features.

Each data point contains:

- `input_text`: The conversation history with speaker tags (T/P)
- `target_text`: The appropriate therapeutic response

Conversation turns are separated by the `[SEP]` token. This structure enables the model to learn appropriate therapeutic responses based on the conversation history while maintaining the dialogue context through explicit speaker identification.

5 Experimental Setup

The preprocessing of data is handled by the NextUttPreprocessor class, which is responsible for preparing the input text for the model. First, the input text is tokenized using the BartTokenizerFast from the Hugging Face transformers library. The tokenizer is set up to handle special tokens, and it ensures that the sequences are truncated or padded to a fixed length of 1024 tokens. A key part of the preprocessing is generating the speaker-specific labels. The therapist's utterances are marked with the label T: while the patient's utterances are marked with P:. These labels are crucial for distinguishing between the two speakers in the conversation and are added to the input tokens during tokenization.

The model is based on the BartForConditionalGeneration architecture, which is designed for sequence-to-sequence tasks like text generation. To handle the dual-speaker dialogues, the model includes an additional embedding layer that differentiates the utterances of the therapist and the patient. This is done by adding separate embeddings for the two speakers (T: and P:) to the token

embeddings. This modification allows the model to understand which speaker is producing each part of the dialogue, which is important for generating contextually appropriate responses.

The model is trained with a learning rate of 1e-5, which is quite small to allow fine-grained updates to the model parameters. It uses a batch size of 8, ensuring that the model processes a manageable number of examples at a time. Training occurs over 10 epochs, during which the model is trained on the training dataset and periodically validated on a validation set to monitor performance. Adam optimizer is used to update the model's parameters, ensuring that the loss is minimized efficiently.

During the training process, the model computes and logs the training loss for each epoch. This loss is recorded for both the training and validation sets, allowing us to track the model's progress over time. Additionally, after each epoch, the training and validation losses are saved, and a plot showing these losses is generated to visually observe the model's learning curve.

For evaluating the model, two primary metrics are used. The BLEU score is computed to measure the precision of the n-grams between the predicted and target output, providing an indication of how well the model's generated responses match the ground truth. Additionally, the BERT score is used, which evaluates the quality of the generated responses using BERT embeddings. This includes measuring precision, recall, and F1-score to assess the model's ability to generate relevant and accurate responses.

At the end of the training, the model is evaluated using these metrics, and the final results give an indication of the model's performance in generating the next appropriate utterance in a conversation.

Hyperparameters

- **Max Tokens:** The input and output sequences are truncated or padded to a maximum length of 2048 tokens.
- **Learning Rate (lr):** The learning rate for training the model is set to 1e-5.
- **Batch Size:** The batch size for training is set to 8.
- **Epochs:** The model is trained for 10 epochs.
- **Optimizer:** Adam optimizer is used for updating the model parameters.

6 Results

The evaluation metrics of the model are as follows:

Metric	BaseLine1 Score	BaseLine2 Score	Model Score
BLEU Score	0.0143	0.0026	0.0162
BERT Precision	0.8599	0.7366	0.8583
BERT Recall	0.8493	0.8375	0.8470
BERT F1-Score	0.8543	0.7832	0.8523

Table 1: Results

7 Observations

Our experiments revealed a limitation in the capacity of a relatively small model such as BART-base (139M) to effectively learn the nuances of conversational turn-taking, specifically in the context of therapist-patient interactions. We observed that this model, while capable of generating coherent text, struggled to consistently capture the subtle mannerisms, pacing, and specific phrasing characteristic of therapist utterances. This suggests that accurately predicting the next therapist utterance, a task that demands a deeper understanding of therapeutic dialogue, necessitates a model with greater representational power and a broader knowledge capacity. A larger model, pre-trained on a more extensive and diverse corpus, may be better equipped to internalize the complexities of therapeutic discourse and thus generate more contextually appropriate and stylistically accurate responses.

8 Conclusion

In this project, we developed an NLP-based system to predict the next utterance in mental health counseling conversations. The model, trained on real counseling dialogues, was evaluated using BLEU and BERT scores to ensure its contextual relevance and emotional appropriateness. The results demonstrated that the model can generate responses that align well with the conversation's context, although there is room for improvement in the BLEU score, indicating the need for better n-gram matching. The BERT score results, however, show that the model successfully captures semantic meaning and context, making it a promising tool for assisting therapists in mental health counseling.

For future work, several improvements can be made to enhance the system's performance. These include increasing the size and diversity of the dataset to improve generalization, incorporating more advanced techniques like attention mechanisms and reinforcement learning, and exploring the use of larger models for more robust perfor-

mance. Additionally, fine-tuning the model with domain-specific knowledge and emotional intelligence could improve its effectiveness in providing contextually appropriate and emotionally supportive responses. Further experimentation with different architectures and evaluation methods, such as human evaluations of response quality, would help refine the system for real-world applications in mental health counseling.

9 References

1. **BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension**, Mike Lewis, Yinhan Liu, Naman Goyal, https://arxiv.org/pdf/1910.13461

2. **Language Models are Unsupervised Multitask Learners**, Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

3. **Response-act Guided Reinforced Dialogue Generation for Mental Health Counseling**, Aseem Srivastava, Ishan Pandey, Md. Shad Akhtar, Tanmoy Chakraborty, https://arxiv.org/abs/2301.12729

4. **Counseling Summarization using Mental Health Knowledge Guided Utterance Filtering**, Aseem Srivastava, Tharun Suresh, Sarah Peregrine (Grin) Lord, Md. Shad Akhtar, Tanmoy Chakraborty1, https://arxiv.org/pdf/2206.03886