MEng. Individual Project

Department of Computing

Imperial College of Science, Technology and Medicine

A virtual psychotherapist that can understand human language

Author:
Hongyuan Yan

Supervisor:
Anandha Gopalan
Second marker:
Lucia Specia

June 22, 2022

Submitted in partial fulfillment of the requirements for the MEng Joint Mathematics Computing of Imperial College London

Abstract

With the increasing demand of psychotherapy all over the world, digital psychotherapy has become a scalable substitute of psychotherapy in may cases. The Self-Attachment Techniques (SAT) is a set of self-administrable protocols suitable for digitalisation. This leads to the birth of the SAT chatbot. The current version of the SAT chatbot, however, cannot accept open-text inputs, which severely limits its ability to understand and communicate with users. Therefore, in this paper, we propose and explore a novel design of the SAT chatbot that extracts useful information from open-text inputs through the participation in empathetic open-domain chitchats with the users. Under this framework, we've focused on intent classification and response generation of the SAT chatbot, receiving promising results in human trials.

Acknowledgments

First and foremost, the completion of this project could not have been accomplished without the continuous support of my supervisor, Dr Anandha Gopalan. I want to thank him for not only being an incredibly helpful advisor, but also being an adamant source of encouragement.

I would also like to extend my thanks to Professor Abbas Edalat, for his precious feedback on my progress and for guiding me through the project.

I am also thankful Lisa Alazraki and Neophytos Polydorou for always being patient explaining to me all my questions.

I would like to thank every member of the SAT project team at the weekly meetings. The weekly discussions have offered me great inspirations throughout the year.

Finally, I would like to thank my parents for always being supportive in this challenging year and Lisa Chai for helping me collect the immensive dataset used in this project.

Contents

Chapter 1

Introduction

1.1 Motivation

Mental health is one of the most challenging problems globally. In a systematic and comprehensive analysis on 354 diseases and injuries for 195 countries and territories from 1990 to 2017 [1], it shows that mental disorder represents the most prevalent threat to human health, responsible for over 1 billion YLDs (one full year of healthy life lost due to disability or ill-health) in 2017. In recent years, the situation deteriorates with the outbreak of the COVID-19 pandemic; depression and anxiety become more prevalent [2].

In the meantime, traditional psychotherapy has a very limited scalability to match this increasing demand of mental support, especially in low-income and middle-income countries [3].In contrast, these countries report much higher statistics of YLDs caused by mental health problems comparing to high-income countries [4]. Under these circumstances, a new class of psychotherapy - digital psychological interference - is being developed as a complementary and scalable mental health service for wider populations [3; 5].

Under this background, the Algorithmic Human Development (AHD) group at Imperial College London has developed a self-administrable psychotherapy: Self Attachment technique (SAT) [6; 7]. It aims at nurturing positive emotions and reprocessing negative experiences by guiding the users through a set of carefully-designed protocols [8]. A digitalised version of SAT - the SAT chatbot - has been designed and implemented to facilitate the administration of SAT in a larger scale [9; 10]. The current version of the chatbot is heavily based on multiple choice questions. To improve the novelty and engagingness of the chatbot, these questions are augmented and randomised. However, the chatbot always follow a predetermined conversation flow [9]. While this design ensures the safety of the chatbot-generated responses, it only has a generic and rigid understanding of each user's unique mental state.

1.2 Objective

The main objective is to enable the current SAT chatbot to communicate with the users through open-text inputs, with the long term goal of understanding and interacting with the users in open-domain conversations. The specific objectives are as follows:

(1) Respond to users' open-text inputs with related and empathetic responses.
(2) Understand users' mental states and the causes of negative emotions.
(3) Recommend suitable SAT protocols at the end of conversations.

1.3 Contributions

Towards fulfilling the objectives listed above, this paper provides several contributions that reveal valuable insights into developing an open-domain SAT chatbot and lay the necessary groundwork for future research.

The main achievement is the improvement over the current version of the extensively rule-based SAT chabot's ability to understand open-text user inputs and to generate relevant and empathetic responses. Specifically, we have designed and explored a novel architecture for the SAT chatbot that enables it to:

(1) Engage with the users in empathetic open-domain chit-chat conversations.
(2) Through chit-chats and questions, extract and analyse useful information from the users, including user intents, user emotions and sentiments, causes of negative emotions, etc.
(3) Recommend appropriate SAT protocols based on information gathered in the chit-chats or questions.

Chapter 2

Background

The foundations of the work presented in this paper rest on existing researches in Self-Attachment Techniques and its automation via the medium of chatbots. This project extends from an existing, partially rule-based version of the SAT chatbot.

2.1 Digital Psychotherapy

Psychotherapy is viable to be conducted in a digital and automated manner, which is realised using mental support chatbots [11]. These chatbots usually have fixed, typically template-based, structures for conversation (including user intents and dialogue states) modelling and response generation; in these structures, the user inputs are modelled by slot filling, while responses are generated using methods such as filling in the templates written by experts [11; 12]. These fixed structures make the conversation flows controllable and safe, which is important for patients seeking mental support. However, this design has the drawback of making conversations more rigid and less engaging, especially when the patients have to interact with the chatbot on a regular basis [13].

2.2 Empathy

In this paper, the concept of empathy in conversations has the same reference as in the previous implementation of SAT chatbot - to the work of ***Godfrey T. Barrett-Lennard*** who identifies three main phases of an empathetic dialogue between two individuals - for consistency [9; 14]. Empathy is an important feature of daily conversations and social interactions [15]; it is also an essential part of the therapist-client interactions in psychotherapy (and hence digital psychotherapy) [16].

2.3 Self-Attachment Technique (SAT)

Self-attachment technique aims at assisting an individual to reprocess negative experiences as well as fostering positive emotions. It bases its therapy on attachment theory [6; 7] and achieves its goal by creating an affectionate bond between each participant and his/her childhood self [8]. Via this secure attachment, it helps the participants to enhance their emotional and mental states [6]. This idea is accomplished through the practice of 20 protocols extracted from the 20 exercises in the paper [8]:

(1) Connecting with the child.
(2) Falling in love with the child.
(3) Vowing to adopt the child as own child.
(4) Maintaining our love with the child.
(5) Protocols for creating zest for life.
(6) Loosening facial and body muscles.
(7) Protocols for love of nature.
(8) Laughing on our own.
(9) Laughing at our two childhood photos.
(10) Continuous laughter.
(11) Processing current negative emotions.
(12) Laughing at our pain.
(13) Perspective change for getting over negative emotions.
(14) Protocols for socializing the child.
(15) Recognizing and containing narcissism and the internal persecutor.
(16) A more optimal internal working model.
(17) Solving personal crisis.
(18) Reprocessing childhood traumas.
(19) Updating our beliefs to enhance creativity.
(20) Practicing Affirmations.

2.4 Language Models

2.4.1 Language Modelling

For a sequence of words $W = w_1 ... w_n$, using the chain rule of probability, language models compute the probability of $P(W)$ as [17]:

$$P(W) = P(w_{1:n}) = \prod_{k=1}^{n} P(w_k | w_{1:k-1}). \tag{2.1}$$

With the Markov assumption, each of the components of the product in equation 2.1 can be approximated for N-grams as [17]:

$$P(w_n | w_{1:n-1}) = P(w_n | w_{n-N+1:k-1}). \tag{2.2}$$

2.4.2 Perplexity (PP)

Perplexity is a metric to evaluate language models. The perplexity of a language model on a test set for a sequence of words $W = w_1...w_n$ is [17]:

$$PP(W) = \sqrt[n]{\frac{1}{P(W)}} = \sqrt[n]{\frac{1}{\prod_{k=1}^{n} P(w_k|w_{1:k-1})}}. \tag{2.3}$$

2.4.3 Attention Networks

In the attention process [18], the inputs have 3 components: query, key and value. Each key vector is associated with a value vector. Denote the query vectors as $\boldsymbol{q_i} \in \mathbb{R}^{d_q}$, the key vectors as $\boldsymbol{k_j} \in \mathbb{R}^{d_q}$ and the value vectors as $\boldsymbol{v_j} \in \mathbb{R}^{d_v}$. For each query vector q_i, the system computes a "similarity" score between the query and all key vectors $\boldsymbol{k_j}$; then it computes the weighted sum of all value vectors $\boldsymbol{v_j}$ according to this "similarity" score. The equation for the single-head attention can be written as equation 2.4, where $a(\cdot)$ is a row-wise activation function. In *soft-attention,* $a(\cdot)$ is the softmax function; in *hard-attention,* $a(\cdot)$ returns a one-hot vector with the argmax element equal to 1. Figure 2.1 illustrates the computation of the output of one single element in a self-attention network.

Self-attention is a the specific attention that sets $K = V$.

$$Attention(Q, K, V; a) = a(\frac{QK^T}{\sqrt{d_q}})V. \tag{2.4}$$

Multi-head Attention

Intuitively, multi-head attention [18] repeats the single-head attention process from different "perspectives". In multi-head attention, the inputs are projected into different sub-spaces. Then they are processed by dot product attention in their corresponding sub-spaces. Finally, these outputs of each attention head are concatenated together and projected back to the original space. This process can be described by equation 2.5 and equation 2.6. Multi-head self-attention is the core of the transformers.

$$Multihead(Q, K, v; a) = concat(head_h 1, ...head_d h)W^O. \tag{2.5}$$

$$head_i(Q, K, v; a) = Attention(QW_i^Q, KW_i^K, VW_i^V; a). \tag{2.6}$$

2.4.4 Transformers

Transformers are powerful deep learning models that adopt the self-attention mechanism [18]. Typically, transformer blocks consist of ***multi-head attention***, ***layer normalisation*** and ***point-wise feed-forward networks***. Figure 2.2 shows an example of a transformer block. The standard inputs for a transformer model are the

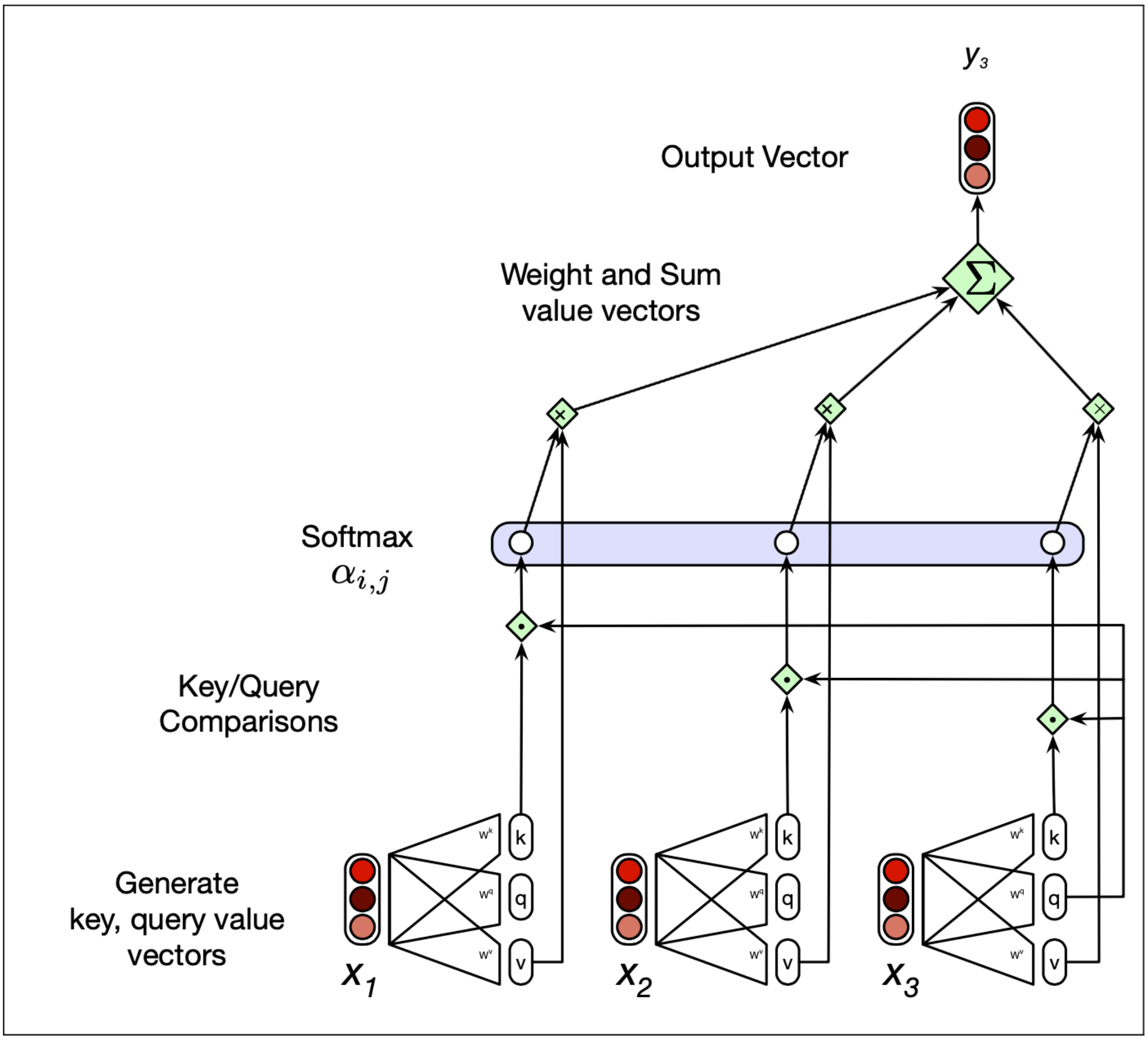

Figure 2.1: Self-Attention Network [17].

word embeddings and the ***position encodings***. For Q and $\boldsymbol{q_i}$ as defined in equation 2.4, the new input queries $\tilde{Q}$ and $\tilde{\boldsymbol{q_n}}$ are formulated as equation 2.5. A common choice of PE function is the sinusoid embedding.

$$\tilde{Q} = (\tilde{\boldsymbol{q_1}}, ..., \tilde{\boldsymbol{q_N}})^T, \tilde{\boldsymbol{q_n}} = f(\boldsymbol{q_n}, PE(n)). \tag{2.7}$$

2.4.5 Pre-trained Language Models

The work presented in this paper is significantly based on pre-trained language models. The models explored and evaluated in this project are listed in this sub-section.

RoBERTa

RoBERTa is introduced in the paper ”RoBERTa: A Robustly Optimized BERT Pretraining Approach” [19]. RoBERTa is one of the most popular variants of BERT [20]

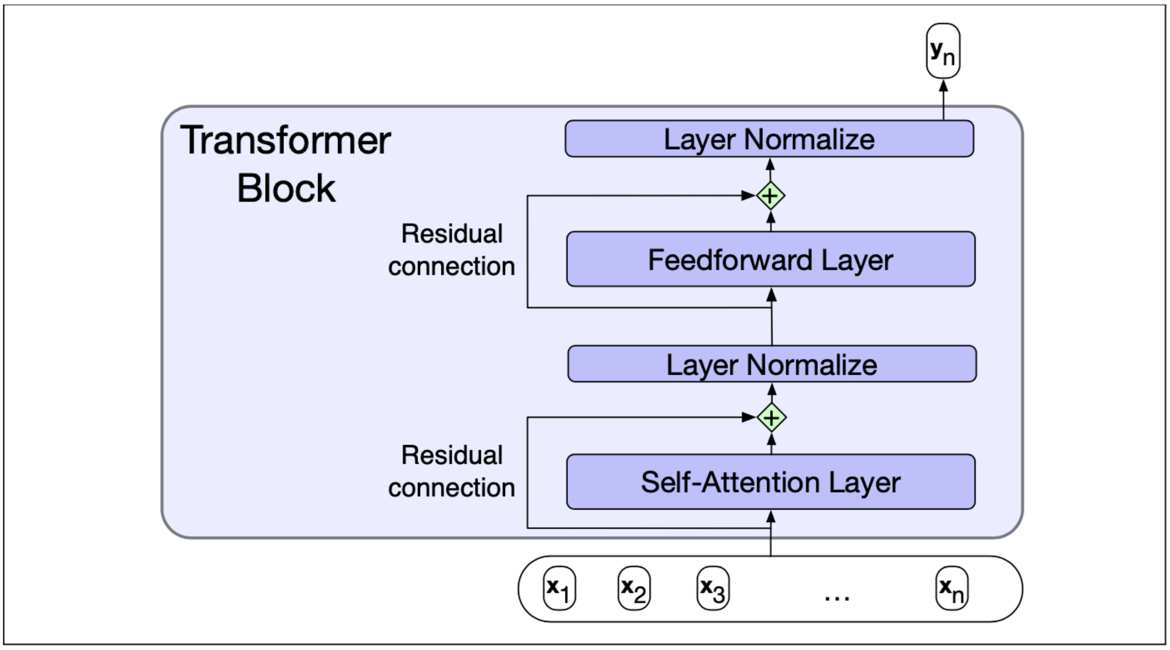

Figure 2.2: Transformer Block [17].

with some modifications. There are several differences between RoBERTa and BERT. Firstly, RoBERTa's tokenizer uses a byte-level BPE while BERT uses a WordPiece tokenizer. Secondly, RoBERTa uses dynamic masking while BERT uses static masking. Thirdly, RoBERTa doesn't have token-type-ids and removes the next-sentence pre-training objective. Lastly, RoBERTa is pre-trained with much larger mini-batches and learning rates on more data.

ALBERT

ALBERT is introduced in the paper "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations" [21]. ALBERT is another popular variant of BERT with parameter reduction techniques. ALBERT splits the embedding matrix into two smaller matrices and uses repeating layers split among groups. These two techniques reduce the memory usage and training time. A key difference between ALBERT and BERT is that ALBERT uses absolute position embeddings. Therefore, it is recommended that inputs are padded on the right rather than the left for training ALBERT models.

MUPPET-RoBERTa

MUPPET is introduced in the paper "Muppet: Massive Multi-task Representations with Pre-Finetuning" [22]. MUPPET proposes a method for pre-finetuning - a stage between a model's pre-training stage and fine-tuning stage. During pre-finetuning, MUPPET uses large-scale multi-task learning to increase the performance of the pre-trained models. In the following sections, the MUPPET pre-finetuned version of RoBERTa model is utilised and evaluated.

XLNet

XLNet is introduced in the paper "XLNet: Generalized Autoregressive Pretraining for Language Understanding" [23]. The models introduced above are autoencoding based (including BERT); in contrast, XLNet is an autoregressive language model. Similar to BERT, XLNet can learn bidirectional contexts. The difference is that BERT uses masked language modelling while XLNet uses permutation language modelling. XLNet also overcomes the pretrain-finetune discrepancy problem in BERT. XLNet can be used for sequence classification, token classification, etc.

GPT2

GPT2 is introduced in the paper "Language Models are Unsupervised Multitask Learners" [24]. GPT stands for generative pre-trained model. Similar to XLNet, GPT2 is an autoregressive model. However, unlike all the models above, GPT2 is a casual language model, meaning that the model focuses only on left-context, thus being uni-directional. This makes the model excel at next token prediction. Therefore, GPT2 is most commonly used for the task of response generation.

2.5 Dialogue Systems

2.5.1 Human Conversations

A dialogue consists of a sequence of ***turns***; each turn is generated by one speaker and can be of arbitrary length [17]. Furthermore, each turn conveys an ***action*** intended by the speaker, which is formally defined as a ***dialogue act*** [25]. There are several characteristics of human conversations that enable dialogues to be conducted smoothly. Firstly, the speakers tend to establish a ***common ground*** of knowledge; they acknowledge each other's meaning regularly in conversations [26; 27]. Secondly, conversations needs to be initiated. Despite special circumstances such as interviews, conversations usually have ***mixed initiatives*** - able to be initiated by every member in a conversation [28]. Thirdly, conversations contain ***implicatures*** [29]. For example, people don't repeat the exact terms and topics of a discussion in every turn, but they know that the conversations are relevant.

2.5.2 Chatbots

Dialogue systems are capable of generating coherent responses given a history of conversations (context). Dialogue systems have been developed for a wide range of purposes. ***Informative chatbots*** provide related information to user queries using their knowledge bases; ***conversational chatbots*** converse with users like another human being; ***task-based dialogues*** provide specialised functionalities such as booking a restaurant [30]. ***This paper is concerned with open-domain conversational chatbots.***

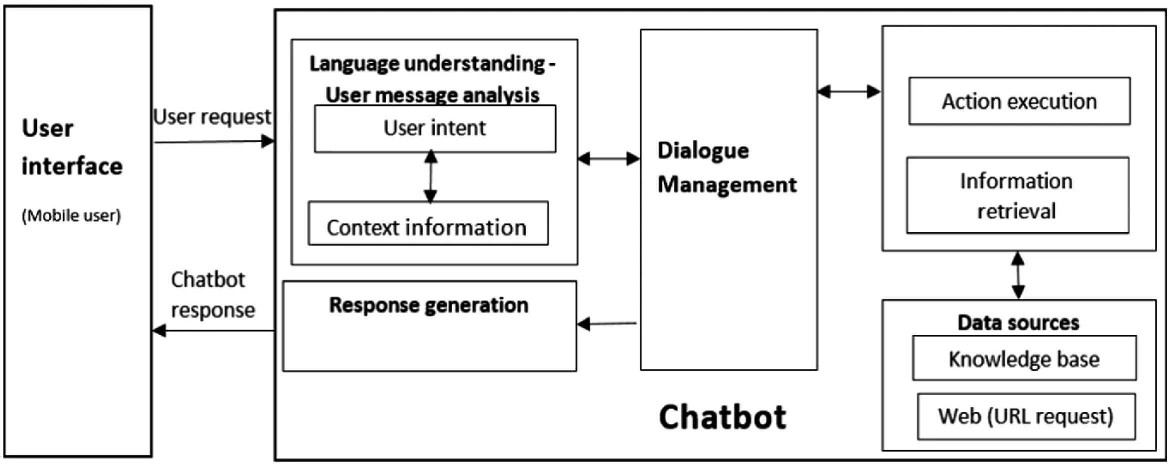

Figure 2.3: Modular Chatbot Structure [30].

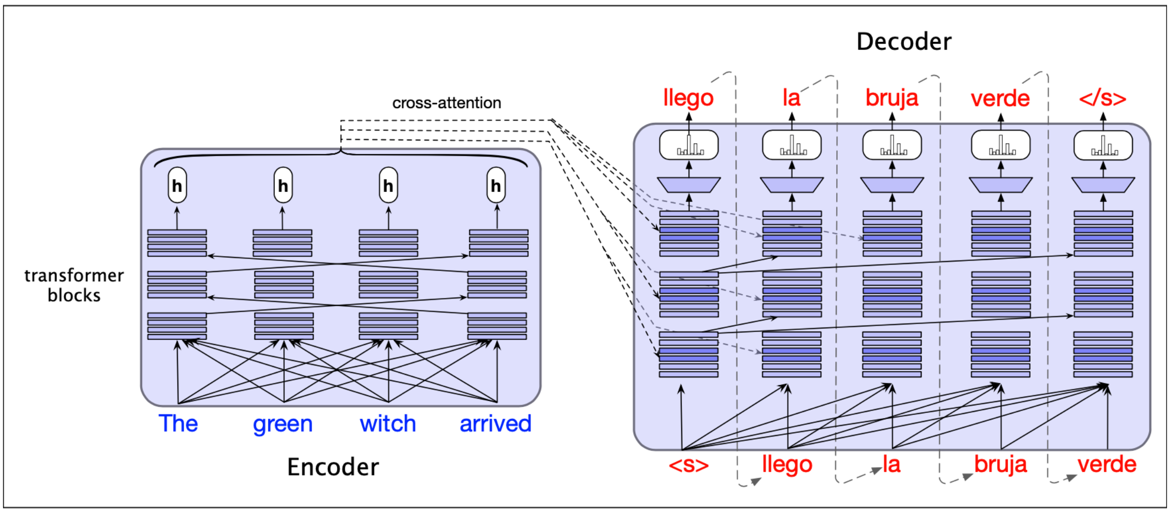

Figure 2.4: End-to-End Chatbot with a Encoder-Decoder Model using Transformers [17].

Based on the ways of generating responses, dialogue systems can be classified as ***generative dialogue systems*** and ***retrieval-based dialogue systems***. Generative dialogue systems usually have an internal representation to facilitate their understanding of current user inputs and history; they produce responses using generative methods based on these representations [31]. Retrieval-based dialogue systems don't necessarily have to understand user inputs; their goal is to pick the most appropriate response from a pool of candidates [17].

Based on the different architectures of the dialogue systems, dialogue systems can be classified into two other classes: ***modular dialogue systems*** and ***end-to-end dialogue systems*** [30; 31]. Modular dialogue systems are composed of several different components, including the ***natural language understanding (NLU)*** component, the ***dialogue state tracker (DST)***, the ***dialogue policy (DP)***, the ***natural language generation (NLG)*** component, etc. [17]. Figure 2.3 shows the structure of a typical modular dialogue system. If the dialogue system supports audio inputs, there can be two additional components - the ***automatic speech recognition (ASR)*** component and the ***text-to-speech (TTS) component*** [17]. Although modular dialogue systems have the ability to understand complicated circumstances and generate sophisticated responses, they are difficult to optimise; the optimisation of a single module doesn't guarantee the optimisation of the whole system [32]. In an end-to-end structure, the dialogue systems generates responses directly from user inputs and history without going through the separate modules [32]. Therefore, they don't need to optimise the modules individually. Furthermore, end-to-end dialogue systems are usually neural-based and don't need manually defined features such as dialogue acts; therefore, they don't require less labour work. A popular model for end-to-end dialogue systems is the ***encoder-decoder model***. Encoder-Decoder models can be constructed using transformers [17], as shown in Figure 2.4.

2.5.3 Natural Language Processing in Modular Chatbots

Natural Language Understanding & Dialogue State Tracking

Although these are two separate modules in Figure 2.3, they are closely related to each other. Natural language understanding is the technique to read the user's utterances and to extract the useful information; then the dialogue state tracking techniques update the current state of the dialogue to the dialogue acts according to the information extracted [17]. In this way, the chatbot maintains a knowledge of the dialogue, including the current topic, the user's intentions and interests, etc.

A simple implementation is a ***slot filling*** model. It uses BIO tagging to locate the spans of interest in the user's utterances that corresponds to the blanks in the dialogue acts; then the changes in slots' values are tracked dynamically [17]. A more advanced implementation is to combine these two steps using a ***reading comprehension*** model. This model reads the conversations up to the current step to generate the newest values for each slot [33].

Dialogue Policy

The Dialogue Policy determines a chatbot's subsequent action (reply) after it updates the newest utterance from the user into its knowledge base. For example, it can ask for further information from a user or offer a user some suggestions using current information. Dialogue policies can be rule-based. They can also be implemented by training neural classifiers using slot fillers and utterances as inputs. More sophisticated methods include using ***reinforcement learning*** in which the agent learns to take actions based on current state [34].

Natural Language Generation

Eventually, a chatbot uses its knowledge about the conversation to generate natural language. A classical implementation for this section uses an Encoder-Decoder

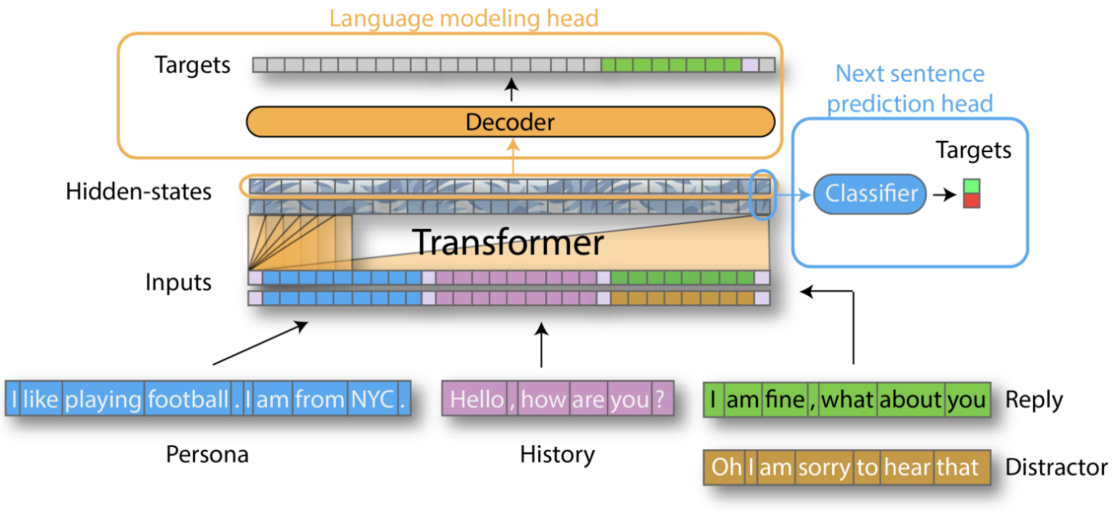

Figure 2.5: TransferTransfo Model Inputs [37].

model to map from dialogue acts into delexicalized sentence templates [35; 36].

2.5.4 Natural Language Processing in End-to-End Chatbots

TransferTransfo

In the paper "TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents", ***TransferTransfo*** [37] proposes a method of fine-tuning GPT models for response generation and shows promising results using the `PERSONA-CHAT` dataset [38]. In this method, the input sequence to a GPT model is constructed by concatenating persona, dialogue history and the target response. Special tokens are added in the tokenizer and the inputs as delimiters and segment indicators. Figure 2.5 shows the structure of the inputs [37]. ***TransferTransfo*** uses a GPT2 model with a language modeling head and a classification head for multi-task training. The model is optimized with respect to a multi-task-loss combining the language modeling objective and the next-sentence classification objective. During training, the target utterance will either be the actual target or a distractor sampled randomly from the dataset.

2.5.5 Decoding

Greedy Decoding

Greedy decoding is the most intuitive decoding strategy. This method uses the trained language model to select the most likely next-token at each step until it reaches the end-of-sequence token. However, greedy decoding doesn't maximise the probability of the decoded output because it ignores the high-probability tokens hiding behind the low-probability tokens.

Beam Search

Beam-search mitigates greedy decoding's drawbacks by maintaining a beam of possible responses along the way. The best response in this beam is selected at the end.

Nucleus Sampling

For high-entropy tasks, there is great variety in the required outputs. Greedy decoding and beam search, however, produce very similar outputs. In these cases, they are replaced by sampling methods at each step of next-token prediction [24]. Nucleus sampling (top-p sampling) is one of these sampling methods. It samples the top tokens and restricts their cumulative probability to be above a threshold p.

Chapter 3

Related Works

3.1 Empathetic Chatbots

While a large variety of conversational agents have been designed, developed and enhanced with great zeal, researchers haven't realised the significance of chatbot's empathetic aspects until recent years [39]. The work presented in this paper is closely related to the recent advances of empathetic chatbots. One of the early approaches towards empathetic response generation is incorporating speakers' emotions [39]. This idea is extended to extracting and including the underlying causes of users' emotional states as well to generate more empathetic and relevant responses [40; 41]. Besides these efforts in empathetic response generation, many researchers have assessed the possibilities and performances of incorporating empathy in the chatbots as a whole product. EMMA is a mental health agent that uses multi-modal knowledge to assess users' mental states to provide assistance through well-being interventions [42]. CAiRE is an empathetic chatbot that generates responses by double fine-tuning generative pre-trained transformer (GPT) models; the first fine-tuning step over general dialogue datasets expands the domain of dialogues it can handle and the second fine-tuning step over empathetic datasets with incorporated sentiments makes the responses empathetic [43]. Finally, XiaoIce is a sophisticated empathetic social chatbot [44]. It encodes and tracks the users' emotional states in every component of the chatbot, including dialogue states, dialogue act decisions and response generation. Despite the complexity of these chatbots and the different approaches undertaken, these researches have inspired many ideas embedded in the work presented in this paper.

3.2 Digitalisation of SAT - the SAT Chatbot

SAT is suitable for digitalisation and automation because of its self-administrable protocols at the core [45]. Delivery of SAT protocols via the medium of chatbots improves the scalability of SAT as a technique for psychotherapy. The chatbot needs to infer a person's mental states and emotional concerns through multiple turns of conversations and suggest the most suitable SAT protocols accordingly at the end. The current version of the SAT chatbot is capable of engaging in conversations through

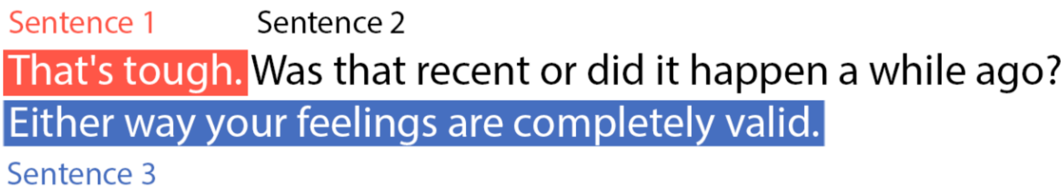

Figure 3.1: Corpus Augmentation in existing implementation of the chatbot [9].

one turn of open text inputs and multiple turns of carefully engineered multiple choice questions [9]. The questions generated by the chatbot are randomised from a pool and augmented with sentences reinforcing their empathy, novelty and engagingness [9]. 3.1 shows an example of this corpus augmentation. The conversations are guided by a predetermined conversation flow.

3.3 Datasets - Learn the Skills

Even though many sophisticated chatbots have very comprehensive dialogue acts, it is still possible for an unknown situation to occur - stemming from unknown user intents, unknown topics, etc. One successful way to handle these situations is to identify isolate these unknown cases. The chatbot uses designed dialogue acts for designed purposes; when that chat enters an unknown space, the chatbot utilises its general chit-chat skills to maintain the conversation. For chatbots to generate high-quality responses in chit-chats, they need to learn different conversational skills from human dialogue datasets. This section lists some high-quality datasets for different chit-chat skills that are complementary to the datasets used in this paper.

3.3.1 PERSONA-CHAT

The `PERSONA-CHAT` dataset is introduced in the paper ”Personalizing Dialogue Agents: I have a dog, do you have pets too?” [38]. The goal of this dataset is to give a chatbot a profile so that they can have a consistent personality in chit-chats. The researcher constructs a personality prompt for the chatbot in every conversations. In total, there are 1155 personalities constructed.

3.3.2 BLENDEDSKILLTALK

The `BLENDEDSKILLTALK` dataset is introduced in the paper ”Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills” [46]. The goal of this dataset is to experiment with the possibility of developing a chatbot with integrated chit-chat skills, including possessing a personality, being knowledgeable, being empathetic, etc. The researchers blend these skills and inject them into the conversations together.

3.3.3 Anno-Mi

The `Anno-Mi` dataset is introduced in the paper ”Anno-Mi: A dataset of expert- annotated counselling dialogues” [47]. It is a recent dataset constructed from counselling sessions. It is one of the first datasets that focus on the counselling skills of chatbots.

Chapter 4

Ethical Issues

We discuss the ethical issues listed in the ethical checklist that are relevant to this project in this section.

The primary concern is the safety of the users. The chatbot engages with users in conversations and suggests its users appropriate protocols to practice from the pool of the protocols [9; 10]. At the same time, a natural-language-generating agent is capable of creating biases [48] or generating toxic sentences [49]. Therefore, the training data of the agent is carefully examined to exclude toxic utterances. The human trial evaluations of this chatbot are also limited to non-clinical trials only.

Another ethical issue related to this project is the protection of personal data. Users are more vulnerable to providing personal information while talking to human-like chatbots [50; 51]. In this research, personal data is protected via anonymity in the trials. The non-clinical trials for the SAT chatbot's evaluation have the ethical approval from Imperial College's Research Ethics Committee. This is extended from the approval for the trials of the previous implementation of the SAT chatbot. Based on the current regulations, the information collected during the non-clinical trials is limited to the minimum required by the study. Furthermore, every participant gives consent after being advised on how their information from the trials will be processed. They can also access the corresponding information, including deleting it, at any time.

Chapter 5

Implementation

5.1 Overview

All the code in this section is available at *https://github.com/HYCJX/SAT_Chatbot*.

The SAT chatbot implemented in this paper is an extension to Lisa's implementation [9]. The following components are constructed and evaluated in this extension:

(1) User input understanding via a user intent classifier that categorises user inputs to different purposes and topics including denying, confirming, stating relationship issues, etc.
(2) Mental state tracking via the combination of an emotion tracker and a sentiment tracker a dialogue.
(3) Empathetic response generation via a chit-chat language model that reacts to open-domain user inputs.
(4) A rule-based dialogue policy that consists of two parts:
    (a) A policy for open-domain chit-chat that extracts relative information from users.
    (b) A policy for asking a designed pool of questions based on the information extracted in chitchats.
(5) SAT Protocol recommendation at the end of a dialogue.

5.2 Intent Classification

5.2.1 Introduction

This functionality is designed to enable the SAT chatbot to extract the intents from user utterances in a dialogue. This is an experimental setup to evaluate the impact of intent classification on understanding open-text inputs. Therefore, the intents recognisable in this implementation don't cover every possible topics. However, this method is easily scalable through the extension or modification on the list of recognisable intents and the addition of training data. In this implementation, intents that facilitate the understanding of users' mental states for the recommendation of SAT protocols are prioritised. User intents of the following categories are designed:

(1) Intents for a user's dialogue acts:
   (a) Greeting: Greetings as in daily conversations.
   (b) Yes: Positive responses to a yes/no question.
   (c) No: Negative responses to a yes/no question.
   (d) Maybe: Indecisive responses to a yes/no question.
   (e) Goodbye: Farewell protocols the end of a conversation.

(2) Intents for recognising the causes responsible for a user's emotions:
   (a) Abuse: Descriptions of the event in which the speaker is abused by someone.
   (b) Assessment: Descriptions of an assessment event, such as exams and interviews.
   (c) Death: Descriptions of an event of death.
   (d) Health: Descriptions about health issues.
   (e) Injustice: Descriptions of events in which the speaker is treated unfairly.
   (f) Jealousy: Descriptions of the speaker's jealousy towards someone.
   (g) Loneliness: Statements that reflects the loneliness of the speaker.
   (h) Missing: Statements that involves "missing someone".
   (i) Partner: Statements related to the speaker's partners.
   (j) Self-blame: Statements in which the speaker blames himself/herself for some reasons.
   (k) Trauma: Descriptions of the speaker's trauma in memory.
   (l) Work: Statements related to the speaker's work/job.

5.2.2 Dataset

To implement the functionality of user intent classification for the SAT bot, a novel dataset - the `Intent` dataset - is constructed. The `Intent` dataset is labelled manually using utterances from two other crowd-sourced dataset.

The data for user intents related to dialogue acts (1st category in section 5.3.1) are extracted from an existing intent-classification dataset - the `CLINIC-OOS` dataset. This dataset is introduced in the paper "An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction" [52]. The dataset contains 23,700 utterances, including 22,500 in-scope utterances labelled with 150 intents and 1,200 out-of-scope utterances. The data labelled with the 5 user intents in the first category in section 5.2.1 is added to the `Intent` dataset.

The data for user intents related to concerns of users (2nd category in section 5.3.1) are constructed using utterances from the `EmpatheticDialogues` dataset. This dataset is introduced in the paper "Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset" [39]. This dataset contains 25 open-domain one-on-one dialogues with 32 available emotion labels. Each dialogue in this dataset is grounded on a given emotional scenario. Therefore, the dialogues in this dataset are closely related to both emotional concerns and empathetic responses. To construct the second part of the `Intent` dataset, a subset of utterances is sampled from the `EmpatheticDialoagues` dataset. Each utterance from this subset is labelled with one of the user intents beloning to the second category in section 5.2.1. If an utterance cannot be labelled with the available user intents, it is discarded.

The label distribution of the constructed `Intent` dataset is summarised in Table 5.1.

`Intent` Dataset			
User Intents	Training	Validation	Test
Greeting	120	15	15
Yes	129	15	15
No	120	15	15
Maybe	120	15	15
Goodbye	120	15	15
Abuse	34	4	5
Assessment	25	3	4
Death	66	7	7
Health	52	7	7
Injustice	20	3	3
Jealousy	29	4	4
Loneliness	16	2	3
Missing	15	2	2
Partner	33	4	5
Self-blame	38	5	2
Trauma	13	2	2
Work	20	3	3
Total	971	121	122

Table 5.1: `Intent` Dataset Label Distribution Summary.

5.2.3 Method

We define user intent classification as a sequence classification task. Currently, this area is dominated by pre-trained language models (PLMs) [20; 19]. Therefore, the intent classification model in the SAT chatbot is implemented and evaluated by fine-tuning PLMs over the dataset introduced in section 5.2.2. Model performance is evaluated on 3 different PLMs: XLNet, RoBERTa and MUPPET-RoBERTa. The Results are evaluated in chapter 6.

5.3 Emotion/Sentiment Classification

5.3.1 Introduction

Although emotion and sentiment are often used interchangeably, they are different concepts [53; 54]. Emotion is a complex but raw psychological state; in contrast, sentiment is an organised mental attitude. Sentiment involves expressing thoughts and feelings about a particular thing. Examples of emotions include ”happy”, ”sad”, ”angry”, etc. Examples of sentiments include ”positive”, ”negative”, etc.

We hypothesise that emotion classification and sentiment analysis have complementary effects to better understand the mental states of users. Therefore, the SAT chatbot is implemented with both an emotion tracker and a sentiment tracker. The two implementations are independent and discussed separately in the subsections 5.3.2 and 5.3.3. The usefulness of the combined tracking mechanism is discussed in chapter 6.

5.3.2 Emotion Classification

Dataset

We select the `MELD` dataset for the training of an emotion classifier. This dataset is introduced in the paper ”MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations” [55]. The authors use utterances from the TV series *Friends* and annotate them using Ekman’s 6 universal emotions and the emotion ”neutral”. The label distribution of the `MELD` dataset is summarised in Table 5.2.

`MELD` Dataset			
Emotions	Training	Validation	Test
Anger	1109	153	345
Disgusts	271	22	68
Fear	268	40	50
Joy	1743	163	402
Neutral	4710	470	1256
Sadness	683	111	208
Surprise	1205	150	281
Total	9989	1109	2610

Table 5.2: `MELD` Dataset Label Distribution Summary.

This dataset is chosen for the following reasons:

(1) The dataset contains training data labelled one of the 7 emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear. This is an appropriate extension on the existing emotion classification capacity (4 emotions) of the previous SAT chatbot.

(2) The emotion labels contain the ”Neutral” emotion. This is the most common emotion in chit-chats - which compose the major dialgogues in the SAT chatbot.

Algorithm 1: Building an input sequence[1]

Given the current utterance $\boldsymbol{x_t} \in \{\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_M}\}$;
$max\_tokens = 512 - 2$;
$sequence = [SEP] + tokenize(\boldsymbol{x_t}) + [SEP]$;
$i = 1$;
while $len(sequence) <= max\_tokens$ **do**
  Prepend $speaker(\boldsymbol{x_{t-i}}) +$ " : " $+ \boldsymbol{x_{t-i}}$ to $sequence$;
  Append $speaker(\boldsymbol{x_{t+i}}) +$ " : " $+ \boldsymbol{x_{t+i}}$ to $sequence$;
  $i = i + 1$;
end
Remove the last appended / prepended utterances;
$sequence = [CLS] + sequence + [EOS]$;

Figure 5.1: Emotiona Classification Data Augmentation Algorithm [56].

Method

We define emotion classification as a sequence classification task. Similar to user intent classification described in section 5.2, the emotion classification model in the SAT chatbot is implemented and evaluated by fine-tuning PLMs over the MELD dataset.

Furthermore, since data distribution is imbalanced for the emotion labels, this paper experiments with a data augmentation technique introduced in the paper "EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa"; the algorithm is illustrated in Figure 5.1 [56]. Model performance is evaluated on 3 different PLMs: ALEBRT, RoBERTa and XLNET, with each of the models trained with original data and augmented data as a comparison. The Results are evaluated in chapter 6.

5.3.3 Sentiment Classification

Dataset

Sentiment classification is trained on the Stanford Sentiment Treebank v2 (SST2) dataset [57]. It contains 11855 data each labelled as either "positive" or "negative. The SST family dataset is selected because it's a standard dataset with promising performances on PLMs. The SST2 binary classification dataset is selected instead of the SST regression dataset because the SAT chatbot is more concerned with distinguishing users' positive sentiments from negative sentiments than knowing an exact sentiment value.

Method

We define sentiment classification as a sequence classification task. Similarly, the sentiment classification model in the SAT chatbot is implmented and evaluated by fine-tuning PLMs over the SST2 dataset. Because training on this dataset is very fast, there are enough training resources to evaluate the model performance on 5 PLMs: ALBERT, RoBERTa base, RoBERTa large, MUPPET Roberta and XLNET. The results are evaluated in chapter 6.

5.4 Response Generation

5.4.1 Dataset

Two datasets are used for response generation training primarily in this project: the DailyDialog dataset [39] and the EmpatheticDialogues dataset [58]. Their basic statistics are listed in Table 5.3.

EmpatheticDialogues & DailyDialog Dataset		
Statistics	Total Dialogues	Average Speaker Turns
DailyDialog	13,118	7.9
EmpatheticDialogues	24850	4

Table 5.3: Statistics of the EmpatheticDialogues Dataset and the DailyDialog dataset.

The EmpatheticDialogues dataset is introduced in section 5.2.2. This dataset contains empathetic responses from crowd-source workers in response to different emotional concerns under different circumstances. The DailyDialog dataset is introduced in the paper ”DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset” [58]. This dataset is constructed using raw data crawled from English learning websites on which the learners practice daily English conversational skills.

In the latter stage of the project, we experimented with the addition of the AnnoMI dataset (introduced section 3.3.3) to the fine-tuning for one of the models.

5.4.2 Method

GPT2 is a pre-trained language model suitable for language generation tasks. In this implementation of the SAT chatbot, the language generation model uses GPT2 as the base model. It is then fine-tuned with the language modelling objective on different datasets in several ways.

To generate high-quality responses, the GPT2 model needs to have a good understanding of the structure of dialogues. We draw inspirations from the finetuning method described in TransferTransfo [37] to design the model inputs with several modifications:

(1) We don't have the persona clause because it's absent in the training datasets we use.
(2) We omit the next-sentence prediction loss because we're not sure of the impacts on empathy if random sentences from the dataset are drawn as distractors.
(3) We construct our own history from the datasets because they don't provide this field. We ensure that the replying speaker is always ⟨speaker2⟩ to maintain a fixed perspective.

Model Inputs

To illustrate the approach of model input construction, use the following data as an example:

```
history = [
    ['how', 'are', 'you', '?'],
    ['i', 'am', 'fine', 'thanks', '.']
]
reply = ['that', 'is', 'great']
```

Firstly, we add the following special tokens with their corresponding meanings to the model tokenizer:

- ⟨bos⟩: Beginning of speech.
- ⟨eos⟩: End of speeh.
- ⟨speaker1⟩: 1st speaker.
- ⟨speaker2⟩: 2nd speaker.

These special tokens mark the model input embeddings for a better understanding of the conversational structure of one-on-one dialogues as well as its relationship to the latest reply. The processed input then becomes:

```
['<bos>', '<speaker2>', 'how', 'are', 'you', '?',
'<speaker1>', 'i', 'am', 'fine', 'thanks', '.',
'<speaker2>', 'that', 'is', 'great', '<eos>']
```

Next, we construct the input ids and the token type ids. The input ids are generated from the tokenizer with all special tokens added to both the tokenizer vocabulary and the model's embeddings.

The token type id for each token is either ⟨speaker1⟩ or ⟨speaker2⟩, depending on the speaker of the sentence. ⟨bos⟩ has the same token type id as the first token; textlangleeos⟩ has the same token type id as the last token, which is ⟨speaker2⟩ in our case.

The language modelling target keeps only the reply. It assigns the value -1 to the rest, including the speaker token in the reply.

This concludes the construction of model inputs.

Finetuning

The model is finetuned on the datasets introduced in section 5.4.1. We implemented and evaluated five fine-tuned GPT2 models. The first two models are fine-tuned on the `DailyDialog` dataset and the `EmpatheticDialogues` separately. However, these two models have their limitations. Fine-tuning on the `DailyDialog` dataset improves the chit-chatting ability of the model; fine-tuning on `EmpatheticDialogues` dataset improves the empathy of the model. On the other hand, however, the `DailyDialog` dataset provides little knowledge on empathy, while the `EmpatheticDialogues` dataset is based on emotional scenarios which are very different from everyday chit-chats. To combine the advantages of the two models, we experimented with the following two ways:

(1) Mix the datasets into one dataset and use this dataset to fine-tune GPT2.

(2) Fine-tune GPT2 on one dataset first and then fine-tune this model on the other dataset.

For the second method, we fine-tune GPT2 on `DailyDialog` first because it has a more general set of conversations. We use the `EmpatheticDialogues` dataset as the second dataset in one setting and the mixed dataset of `EmpatheticDialogues` and `AnnoMI` in an latter setting. This concludes all five models fine-tuned in this paper for response generation. The results of all models are evaluated in chapter 6.

Decoding

We implement the decoder of the fine-tuned GPT2 model using nucleus sampling. As introduced in section 2.5.5, nucleus sampling is capable of generating more variant responses than beam search, which is essential for open-domain chit-chats

5.5 Dialogue Policy

5.5.1 Introduction

As illustrated in section 5.1, the current dialogue policy is hierarchical and consists of two secondary policies: an open domain chit-chat policy and a rule-based question-asking policy. Because of the primary policy's limited ability to guide conversations, the transitions between the two secondary policies is rigid. In the current primary policy, the chatbot chit-chats with a user under the chit-chat policy for a predetermined number of turns. Then the chatbot switches to the question-asking policy. Finally, the chatbot recommends a set of SAT protocols at the end. The policy is illustrated in Figure 5.2.

5.5.2 SAT Protocol Recommendation & Question-Asking Policy

The SAT chatbot adopts a similar logic of recommending SAT protocols as in the previous version. However, as the current chatbot cannot guide the users to practice SAT protocols during chit-chats, the protocol recommendation strategy in this paper

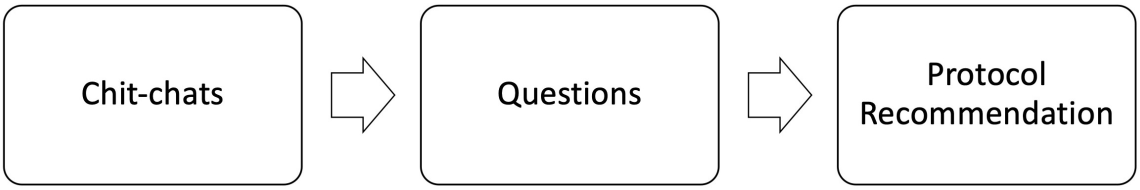

Figure 5.2: Primary Dialogue Policy.

involves only the question-asking branch in the previous implementation [9]. The questions and corresponding SAT protocols are listed in Table 5.4. There is a modification to the question-asking policy. The chatbot doesn't stop asking questions when it gets a positive answer. This setting facilitates the testing of the chit-chat policy.

Base Question	Positive Answer	Negative Answer
Have you strongly felt or expressed any of the following emotions towards someone: envy, jealousy, greed, hatred, mistrust, malevolence, or revengefulness?	Add protocols 13, 14.	Add protocol 13.
Do you believe that you should be the saviour of someone else?	Add protocols 8, 15, 16, 19.	Add protocol 13.
Do you see yourself as the victim, blaming someone else for how negative you feel?	Add protocols 8, 15, 16, 19.	Add protocol 13.
Do you feel that you are trying to control someone?	Add protocols 8, 15, 16, 19.	Add protocol 13.
Are you always blaming and accusing yourself for when something goes wrong?	Add protocols 8, 15, 16, 19.	Add protocol 13.
Is it possible that in previous conversations you may not have always considered other viewpoints presented?	Add protocols 13, 19.	Add protocol 13.
Are you undergoing a personal crisis (experiencing dif- ficulties with loved ones e.g. falling out with friends)?	Add protocols 13, 17.	Add protocol 13.

Table 5.4: Protocol Recommendation.

Question-Asking Policy

The Question-Asking Policy of the SAT chatbot uses the SAT protocol recommendation strategy illustrated above as the backbone. It will not ask a base question if the answer is already obtained (during the execution of the Chit-Chat policy).

5.5.3 Chit-Chat Policy

Under the chit-chat policy, the chatbot converses with the users through open-domain chit-chats. In these conversations, the chatbot utilises its abilities of intent classification, emotion classification and sentiment analysis to extract answers to the base questions in Table 5.4. The logic flow of this process is depicted in the flow chart Figure 5.3.

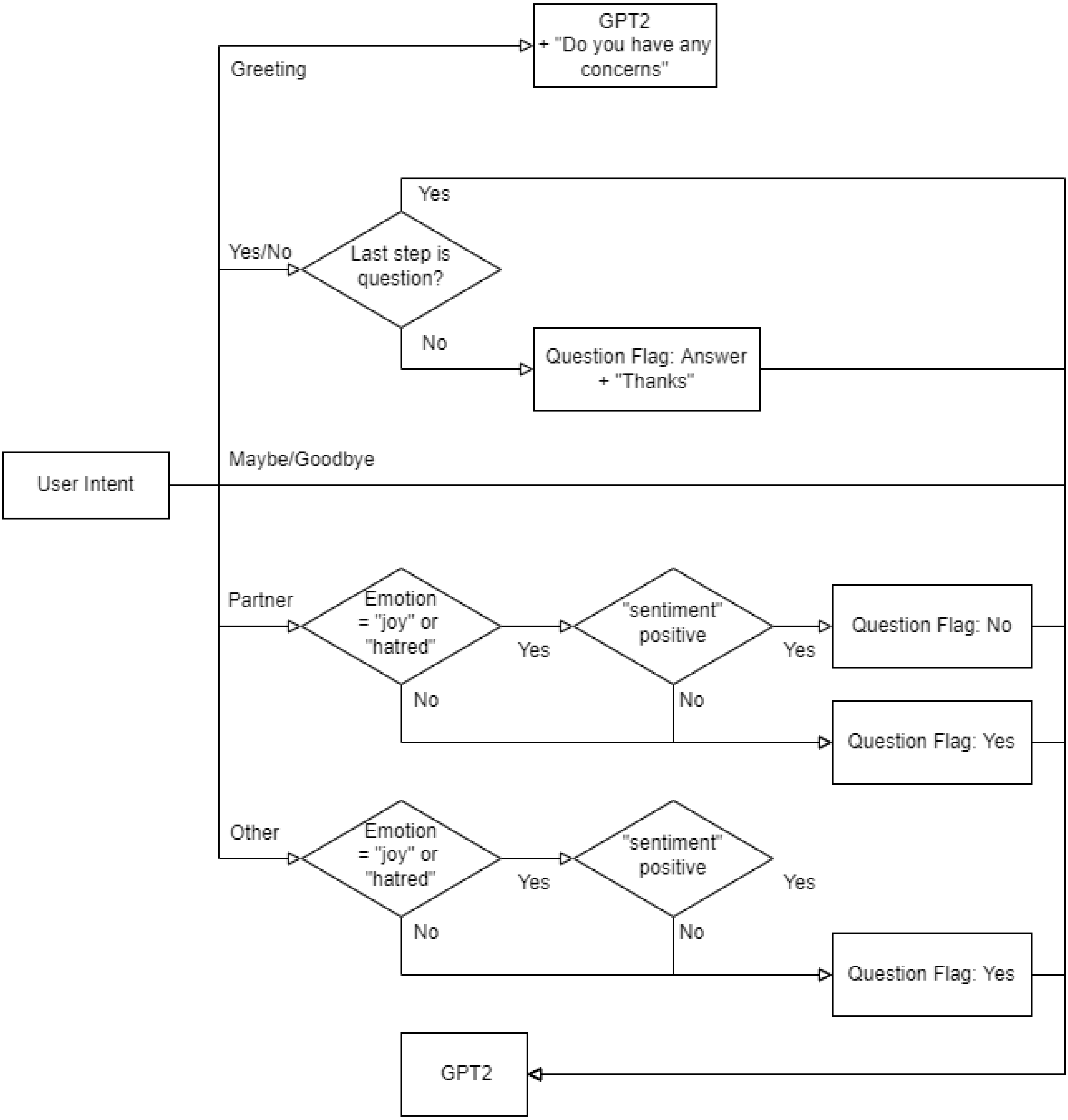

Figure 5.3: Chit Chat Policy.

Chapter 6

Evaluation

6.1 Intent Classification

The hyperparameters of intent classifiers are searched within the following ranges:

- Training epochs: Constant - 10.
- Training batch size: Constant - 16.
- Evaluation batch size: Constant - 32.
- Optimizer: AdamW.
- Learning rate: Log domain - (5e-7 to 5e-3).
- Warm-up ratio: Linear domain - (0.1 to 0.3).
- Weight decay: Log domain - (0.001 to 0.1).
- Maximum gradient: Constant - 1.0.
- Early stopping: True.
- Metric: Weighted f1 on validation set.

The hyperparemters are searched in 25 trials. The search algorithm is the Tree-structured Parzen Estimator) algorithm [59] using the Optuna framework. Using this algorithm, model performance starts to make steady progress after 10 trials.

The intent classifiers are trained using the ***Huggingface Trainers***. Table 6.1 shows the test set performances of the fine-tuned models trained with tuned hyperparameters.

Model	Weighted F1	Micro F1	Macro F1
muppet-roberta-base	**0.800**	0.832	0.581
roberta-base	0.574	0.624	0.290
xlnet-base-cased	0.742	0.752	0.565

Table 6.1: Intent Classification Results.

The fine-tuned MUPPET RoBERTa base model has achieved the best performance in the trials. Weighted f1 is in an acceptable range, indicating that the design and construction of the intent classification data is successful. However, the macro f1 score for all 3 models is quite low, indicating that the intent classification for minor class data is inaccurate comparing to majority classes.

6.2 Emotion/Sentiment Classification

6.2.1 Emotion Classification

The MELD dataset is too large to be fine-tuned properly with existing computational resources. Therefore, all emotion classifiers are trained using the ***Huggingface Trainer*** under the following settings:

- Training epochs: 15.
- Training batch size: 16.
- Evaluation batch size: 32.
- Optimizer: AdamW.
- Learning rate: 1e-5.
- Warm-up ratio: 0.2.
- Weight decay: 0.01.
- Maximum gradient: 1.0.
- Early stopping: True.

The test results of the fine-tuned models are shown in Table 6.2. The ***SOTA*** performance on this dataset achieves the weighted f1 score of 66.71. The model with the best performance in this trial the RoBERTa base model trained without data augmentation. It achieves a weighted f1 score of 62.4. In general, the data augmentation technique doesn't work well. We hypothesise that the cause is lack of hyperparameter tuning because the addition of training data may benefit from a different learning rate. The trained model achieves an above-average performance comparing to other models on the benchmark. The low macro f1 score means that the model is inaccurate at predicting labels for minor class data.

Model	Data Augmentation	Weight F1	Micro F1	Macro F1
albert-base-v2	False	0.587	0.603	0.388
	True	0.393	0.514	0.188
roberta-base	False	**0.624**	0.648	0.428
	True	0.607	0.628	0.425
xlnet-base-cased	False	0.621	0.635	0.437

Table 6.2: Emotion Classification Results.

6.2.2 Sentiment Classification

The hyperparameters of sentiment classifiers are searched within the following ranges:

- Training epochs: Constant - 10.
- Training batch size: Constant - 16.

- Evaluation batch size: Constant - 32.
- Optimizer: AdamW.
- Learning rate: Linear domain - (1e-5 to 5e-5).
- Warm-up ratio: Linear domain - (0.1 to 0.3).
- Weight decay: Log domain - (0.001 to 0.1).
- Maximum gradient: Constant - 1.0.
- Early stopping: True.
- Metric: Weighted f1 on validation set.

The hyperparemters are searched in the same way as the intent classifiers.

Table 6.3 shows the test set performance of the fine-tuned models trained with tuned hyperparameters.

Model	Weighted F1
albert-base-v2	0.923
muppet-roberta-base	**0.960**
roberta-base	0.948
roberta-large	0.953
xlnet-base-cased	0.933

Table 6.3: Sentiment Classification Results.

The MUPPET pre-finetuned RoBERTa base model achieves the best performance in this trial.

In contrast to the emotion classifiers, the sentiment classifiers achieve very high performance. This is due to the reduction in the task difficulty caused by both the smaller number of classes and the more objective features.

6.3 Response Generation

The models are trained on different datasets. Therefore, it is difficult to compare their performances using automatic metrics such as perplexity. Furthermore, there has been researches showing that automatic metrics can have low correlations with model performance under some circumstances [60]. Therefore, the evaluation of SAT chatbot's response generation functionality is done by human trials. The exact set-up of the tries are discussed in the next section.

6.4 Study Set-up

The SAT chatbot is evaluated in a human trial formally. There are 10 volunteers from the non-clinical population participating in the trials. All participants are familiar with the SAT protocols.

The trial consists of two parts. In the first part, the participants score the responses generated by the five response generation models for three prompts. Each response is scored on three aspects: relatedness, fluency and empathy. The models and their respective fine-tuning strategies are listed in Table 6.4. These models will be referred to using the assigned code throughout the rest of this chapter. The prompts and the generated responses are displayed in Table 6.5.

The prompts are selected deliberately to convey different contextual messages. The 1st prompt is daily conversational English. It's most similar to the data in the `DailyDialog` dataset among all datasets we have used for training. The 2nd prompt has a vague meaning because the context of this prompt influences its meanings. In a daily conversational context, it might be the simple and straightforward meaning of feeling tired after a day of work. In an emotional context, it can represent a negative attitude towards living, even fostering mental health problems. The 3rd prompt is obviously embedded within an emotional scenario, which is most similar to the data in the `EmpatheticDialogues` among all the datasets that we've used for training.

Using these three prompts, we want to evaluate how well the five fine-tuned response generation models perform in these dialogues that usually happen in very different situations. We also want to compare how these models interpret the vague contexts such as the one provided in prompt two.

Fine-tune Strategy	Model Code
Fine-tuned on the `DailyDialog` dataset.	Model 1
Fine-tuned on the `EmpatheticDialogues` dataset.	Model 2
Fine-tuned on the mixed `DailyDialog` and `EmpatheticDialogues` dataset.	Model 3
Fine-tuned on the `DailyDialog` dataset first and then finetuned on the mixed `EmpatheticDialogues` and `AnnoMi` dataset.	Model 4
Finetuned on the `DailyDialog` dataset first and then finetuned on the `EmpatheticDialogues` dataset.	Model 5

Table 6.4: Response Generation Models.

In the second part of the trial, the participants interact with the SAT chatbot as an overall evaluation. During the trial, the chit-chat number of turns is set to be 7. This means that the participants will converse with the SAT chatbot in open-domain chit-chats for 7 turns. Then the SAT chatbot will switch to its question-asking dialogue policy before providing a recommendation of protocols at the end of a conversation. After the trial, each participant is given a feedback. This feedback contains the whole chatbot-human dialogue in the trial, including both the chit-chatting turns and the question-asking turns. In the feedback, this dialogue is aligned with the SAT chatbot's interpretation of every line of a participant's input, including the participant's intents, emotions and sentiments. Using this information, the participants will complete the second part of a survey related to the overall qualities of the SAT chatbot.

Prompt	Response	Model Code
Good morning. How are you doing?	Morning. I'm not feeling very well.	Model 1
	I'm pretty good. I've always wanted to see you out!	Model 2
	Morning. I'm really glad to see you again.	Model 3
	I'm pretty good.	Model 4
	I'm great!	Model 5
I'm so tired.	What time is it?	Model 1
	Oh no, Is something wrong with you?	Model 2
	I'm tired too.	Model 3
	Oh no, Is that okay?	Model 4
	Oh no, Is something wrong?	Model 5
My girlfriend broke up with me yesterday.	What happened?	Model 1
	Oh no, Did you try calling and saying goodbye?	Model 2
	Oh no, Is she okay?	Model 3
	Oh no, My apologies. I know it's terrible, But– but, Uh, I would be really angry if we spoke about this with each other.	Model 4
	Wow. Was it not her fault?	Model 5

Table 6.5: Response Generation Prompts and Responses.

There are 10 volunteers in total who have participated in the human trials and have completed the feedback.

6.5 Human Trial Evaluation

6.5.1 Response Generation

The scores of responses generated are summarised in Plot 6.1. These results have produced many insights on the fine-tuning strategies for response generation.

Firstly, the charts show that model 5 has the most stable performance across all 3 prompts. It's the only model that maintains a good relatedness score and fluency score of generated responses across 3 different prompts. Comparing to the scores of model 3, this shows that ***models fine-tuned sequentially on a grounding and broad dataset and then a dataset for specific skills have a more stable performance in response to the change of dialogue contexts comparing to models trained using mixed data***. We hypothesize that if the first fine-tuning corpus covers a large set of dialogue contexts, the second fine-tuning on datasets for specific situations and properties of dialogues doesn't reset the learnt conversational skills out of its scope.

On the other hand, if we mix the datasets together, the data will become noises of each other that have negative impacts on pre-trained language models learning of conversational skills.

Our second observation is that the fine-tuned GPT2 models performs better in dialogues with similar contexts to their training data. For example, model 1 performs better in prompt 1's situations while model 2 performs better in the other two situations.

Model 5 is also used for response generation in the second part of the survey.

As an overall evaluation of the response generator, model five has demonstrated a good performance under different contexts. It's average score of relatedness in all three prompts is higher than 4, which is about 80 percent of the full score. It's also capable of generating fluent sentences under different contexts. Its lowest average score in fluency is 3.5 in prompt 3, which is much higher that the lowest fluency average score of any other four models. The empathy scores of the response generators, however, are low on average. This means that the responses generated by the response generators are very different from humans in empathetic settings. The response generators have unpredictable performance with large fluctuations under different contexts.

6.5.2 The SAT Chatbot

In this part of the trial, the participants are asked to answer 6 questions after they read the feedback generated by the SAT chatbot in the trial. The questions are listed in Table 6.6. Plot 6.2 shows the participant feedback.

Number	Statement
1	The chatbot was good at guessing my emotion.
2	Does the sentiment label or the emotion label capture your emotional state better?
3	The chatbot is good at understanding my answers to yes/no questions.
4	The chatbot is good at understanding my general intents.
5	When I interact with the chatbot, I found that it displayed empathy in the responses throughout the conversation.
6	When I interact with the chatbot, I found the conversation to be engaging.

Table 6.6: SAT Chatbot Overall Trial Statements.

Feedback of question 1 shows that the SAT chatbot is good at capturing users' emotions. Every participant provide an answer of ”agree” or ”strongly agree”, which

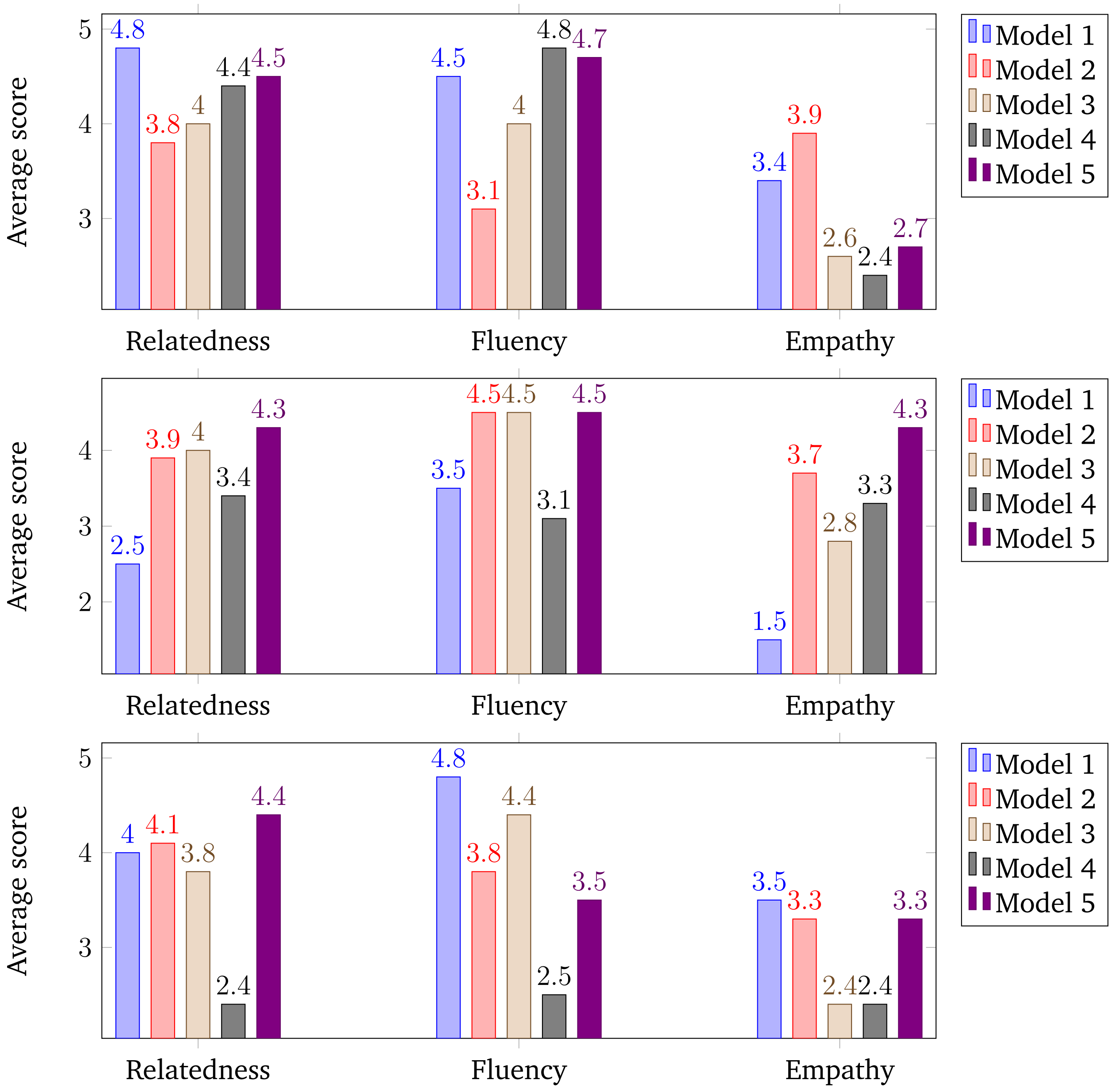

Figure 6.1: Scores for Response Generation on 3 Prompts.

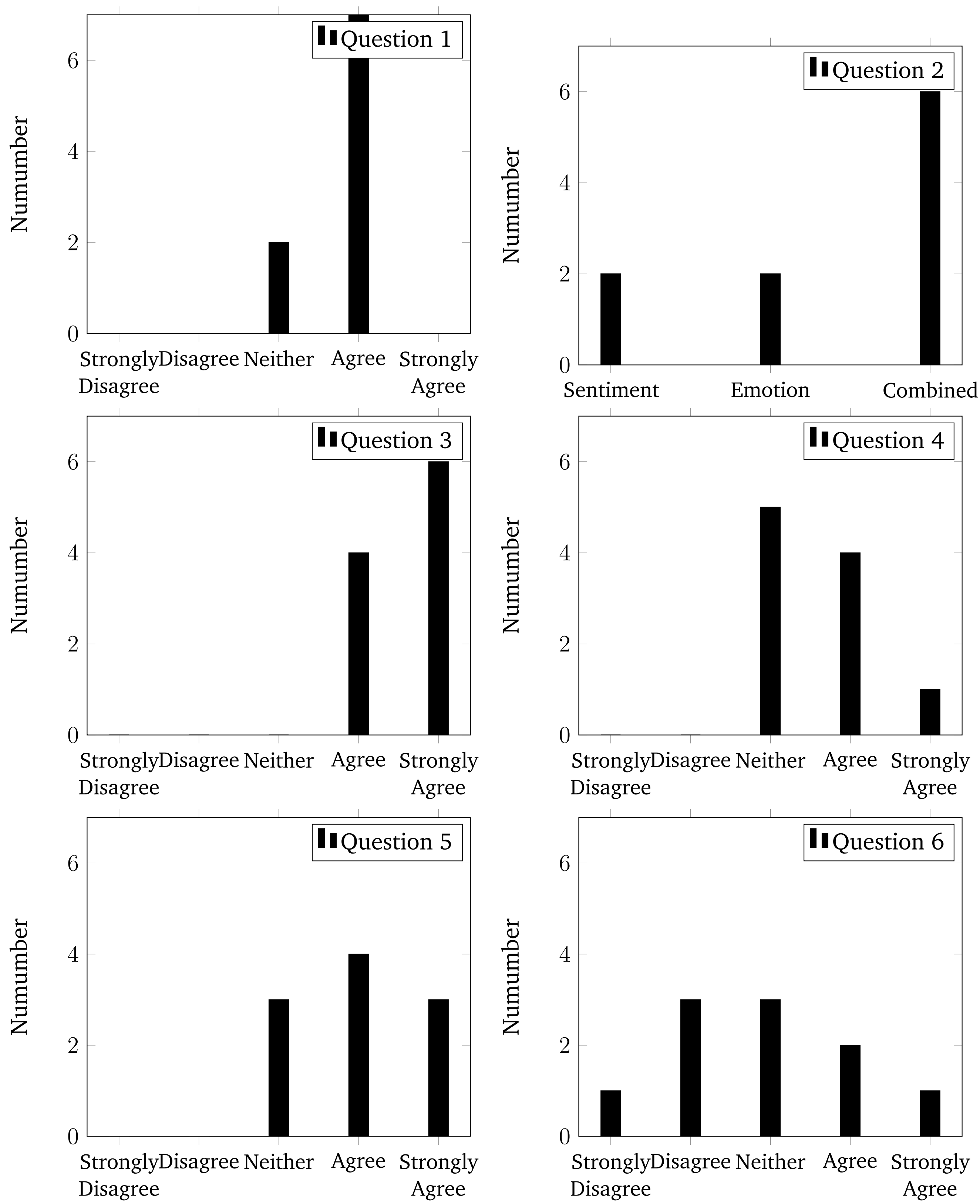

Figure 6.2: Scores for Overall Response in Human Trials.

makes an improvement from the trials for the previous implementation [9]. We hypothesise that this improvement is due to two factors. Firstly, the emotion classifier implemented in this version can recognise 7 emotions, more that the 4 emotions recognisable in the previous implementation. Secondly, the implementation of the sentiment classifier complements the situations when the emotion classifier don't work well. When the emotion classifier gives a "neutral" output, the sentiment classifier can usually produce a positive/negative classification on the user's attitude. The conclusion that the complementary work of the sentiment classifier is useful can also be drawn from the dominant feedback in question 2 that the combined emotion recognition works better.

Question 3 shows that the chatbot is good at understanding user's answers to yes/no questions. However, question 4 shows that the SAT chatbot sometimes struggles to understand the other intents of the users. By examining the logs in human trials, we discover that this problem is often cause by ***out-of-scop (oos) user intents*** [61]. For example, in the current implementation of user intent classification, the chatbot will struggle if the user starts to ask knowledge quiz questions. The intent classifier will try to categorise this intent into one of known categories even if it doesn't fit.

The responses collect in question 5 shows that there is a drop in the SAT chatbot's empathy performace. This result matches the responses obtained in the first part of the trial. The cause is that generated responses are less empathetic than responses written by human retrieved in the previous implementation.
Lastly, the SAT chatbot's engagingness performance is quite bad as suggested by the feedback to question 6. This issue is observed during the trials. Although the chatbot can provide fluent and empathetic responses to user inputs, it doesn't have the ability to initiate conversations. Therefore, the users usually get into the circumstances that they don't know what to say to the chatbot and they have to think very hard to make the conversation continue.

Chapter 7

Conclusions & Future Work

7.1 Our Work

In this paper, we have theorised and experimented with a novel design of the SAT chatbot that understands open-text inputs, engages in open-domain empathetic conversations and recommends appropriate SAT protocols. Human trials have shown promising results on several aspects of this design, including emotion and sentiment tracking, user intent classification and empathetic response generation. Besides these achievements, this chatbot framework is highly scalable with a great potential for improvements.

Under the introduced chatbot framework, we've focused on user intent understanding and chit-chat generation. To enable the SAT chatbot to understand user intent, we've constructed a new user intent dataset which includes dialogue acts such as "yes" and "greeting" as well as emotional causes such as "abuse" and "self-blame". We've trained an intent classification model on this dataset with proper f1 score and performance in human trials. In response generation, we've experimented with several fine-tuning techniques to enable the model to learn different conversational skills from different datasets. In our evaluation of different fine-tuned GPT2 models, we theorise that sequential fine-tuning from datasets with general conversations to datasets with specific intentions yield better learning results comparing to fine-tuning on mixed datasets.

7.2 Future Work

Firstly, the natural language understanding component can be improved. In human trials, there are lots of user intents that cannot be recognised correctly by the chatbot. The most straightforward improvement is to increase the training data and user intents. However, drawing inspirations from recent researches, there are also many other effective ways of enhancing this component that we are not able to cover in this paper. For example, Zhan et al have proposed an effective method to detect out-of-scope user intents [62]. This method can be explored on the chatbot along with an innovative design of dialogue acts of Alquist [63], in which the chatbot maintains hierarchical dialogue acts and with one of them specific to deal with out-of-scope intents using GPT2 generation.

Another component that can be improved immensely is the dialogue policy. Currently, the SAT chatbot uses a set of rule-based dialogue policies, which limit the extension of its functionalities. There are many advanced ways to build sophisticated dialogue policies using neural networks to facilitate smooth and effective transitions between topics. Furthermore, the training of these neural models can incorporate the emotion and sentiment component develop to make the dialogue policies more empathetic and human-like [44].

Furthermore, the response generation component can be enhanced to improve the engagingness of conversations with the SAT chatbot. Currently, the chatbot can respond to users emotional concerns with great empathy. However, the chatbot doesn't have the ability to either guide the conversation or to start a new topic. This makes the conversations tedious in the long run and limits the chatbot's opportunities to learn about the users. In counselling sessions, therapists usually take the initiative of conversations, as presented in the session dialogues collected by the`AnnoMI` dataset [47]. Therefore, learning to ask relevant questions and to initiate conversational topics are essential in the development of a more sophisticated and more effective mental support chatbot.

Appendix A

Sample Conversations

This is an example of a dialogue between the SAT chatbot and a human.

Speaker	Utterance
Chatbot	Hi!
Human	Hi there! How are you?
Chatbot	I'm fine, Have a great life, And I hope you have fun and enjoy life. Do you have any concerns?
Human	I am a little tired and stressed lately.
Chatbot	Well just start to relax and have fun!
Human	I will try, thank you. Do you have any more advice for me?
Chatbot	Just relax and have fun! Just do it!
Human	Haha ok, then I'll just try that.
Chatbot	You must be so happy.
Human	Now that I am talking to you, dear chatbot, I am admittedly happier.
Chatbot	Yea that's for sure. Thank you for your conversations. May I ask you a few more questions
Human	Yes, sure.
Chatbot	Do you believe that you should be the saviour of someone else?
Human	No, not really.
Chatbot	Do you see yourself as the victim, blaming someone else for how negative you feel?
Human	No I do not.
Chatbot	Do you feel that you are trying to control someone?
Human	Nope.
Chatbot	Are you always blaming and accusing yourself for when something goes wrong?
Human	Yes, I definitely do that sometimes
Chatbot	Is it possible that in previous conversations you may not have always considered other viewpoints presented?
Human	It is possible, I mean nobody is perfect
Chatbot	Are you undergoing a personal crisis(experiencing difficulties with loved ones e.g. falling out with friends)?
Human	No, I am fine.
Chatbot	We recommend protocols 13, 14

Table A.1

Bibliography

[1] James SL, Abate D, Abate KH, Abay SM, Abbafati C, Abbasi N, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 Diseases and Injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. The Lancet. 2018;392(10159). Available from: https://doi.org/10.1016/S0140-6736(18)32279-7. pages 1

[2] Shah SMA, Mohammad D, Qureshi MFH, Abbas MZ, Aleem S. Prevalence, Psychological Responses and Associated Correlates of Depression, Anxiety and Stress in a Global Population, During the Coronavirus Disease (COVID-19) Pandemic. Community Mental Health Journal. 2021;57(1). Available from: https://doi.org/10.1007/s10597-020-00728-y. pages 1

[3] Fu Z, Burger H, Arjadi R, Bockting CLH. Effectiveness of digital psychological interventions for mental health problems in low-income and middle-income countries: a systematic review and meta-analysis. The Lancet Psychiatry. 2020;7(10). Available from: https://doi.org/10.1016/S2215-0366(20)30256-X. pages 1

[4] Organisation WH. Depression and other common mental disorders: global health estimates. World Health Organization. 2017. pages 1

[5] Renn BN, Hoeft TJ, Lee HS, Bauer AM, Areán PA. Preference for in-person psychotherapy versus digital psychotherapy options for depression: survey of adults in the U.S. Npj Digital Medicine. 2019;2(1). Available from: https://doi.org/10.1038/s41746-019-0077-1. pages 1

[6] Edalat A. Introduction to self-attachment and its neural basis. In: 2015 International Joint Conference on Neural Networks (IJCNN). vol. 2015-September; 2015. p. 1-8. pages 1, 4

[7] Edalat A. In: Érdi P, Sen Bhattacharya B, Cochran AL, editors. Self-attachment: A Holistic Approach to Computational Psychiatry. Cham: Springer International Publishing; 2017. p. 273-314. Available from: https://doi.org/10.1007/978-3-319-49959-8_10. pages 1, 4

[8] Edalat A. Self-attachment VR Intervention: Detailed Protocols; 2021. pages 1, 4

[9] Alazraki L, Ghachem A, Polydorou N, Khosmood F, Edalat A. An Empathetic AI Coach for Self-Attachment Therapy. IEEE; 2021. p. 78-87. Available from: `https://ieeexplore.ieee.org/document/9750315/`. pages 1, 3, 14, 16, 17, 25, 36

[10] Ghachem A. Evaluation of a virtual agent in guiding users from the non-clinical population in self-Attachment intervention; 2021. pages 1, 16

[11] Cameron G, Cameron D, Megaw G, Bond R, Mulvenna M, O'Neill S, et al. Towards a chatbot for digital counselling. vol. 2017-July; 2017. p. 1-7. pages 3

[12] Denecke K, Vaaheesan S, Arulnathan A. A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test. IEEE Transactions on Emerging Topics in Computing. 2021;9(3):1170-82. pages 3

[13] Følstad A, Brandtzaeg P. Users' experiences with chatbots: findings from a questionnaire study. Quality and User Experience. 2020 04;5. pages 3

[14] Barrett-Lennard GT. The empathy cycle: Refinement of a nuclear concept. Journal of Counseling Psychology. 1981;28. pages 3

[15] Decety J. The neurodevelopment of empathy in humans. Developmental Neuroscience. 2010;32. pages 3

[16] Elliott R, Bohart AC, Watson JC, Murphy D. Therapist Empathy and Client Outcome: An Updated Meta-Analysis. Psychotherapy. 2018;55. pages 3

[17] Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (second edition). USA: Prentice-Hall, Inc.; 2009. pages 4, 5, 6, 7, 8, 9, 10

[18] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc.; 2017. Available from: `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`. pages 5

[19] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al.. RoBERTa: A Robustly Optimized BERT Pretraining Approach; 2019. pages 6, 19

[20] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171-86. Available from: `https://aclanthology.org/N19-1423`. pages 6, 19

[21] Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv; 2019. Available from: https://arxiv.org/abs/1909.11942. pages 7

[22] Aghajanyan A, Gupta A, Shrivastava A, Chen X, Zettlemoyer L, Gupta S. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics; 2021. p. 5799-811. Available from: https://arxiv.org/abs/2101.11038. pages 7

[23] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc.; 2019. Available from: https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf. pages 8

[24] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners; 2019. . pages 8, 12

[25] White AR, Austin JL, Urmson JO. How to Do Things with Words. Analysis. 1963;23. pages 8

[26] Clark HH. Using language. Cambridge University Press; 1996. pages 8

[27] Stalnaker RC. Pragmatics. Synthese. 1970;22(1/2):272-89. Available from: http://www.jstor.org/stable/20114754. pages 8

[28] Walker M, Whittaker S. Mixed Initiative in Dialogue: An Investigation into Discourse Segmentation. In: Proceedings of the 28th Annual Meeting on Association for Computational Linguistics. ACL ’90. USA: Association for Computational Linguistics; 1990. p. 70–78. Available from: https://doi.org/10.3115/981823.981833. pages 8

[29] Grice HP. In: Logic and Conversation. Brill; 1975. p. 41 58. Available from: https://brill.com/view/book/edcoll/9789004368811/BP000003.xml. pages 8

[30] Adamopoulou E, Moussiades L. An Overview of Chatbot Technology”. vol. 584 IFIP; 2020. p. 373-83. pages 8, 9

[31] Zhang Z, Takanobu R, Zhu Q, Huang M, Zhu X. Recent Advances and Challenges in Task-oriented Dialog System. arXiv; 2020. Available from: https://arxiv.org/abs/2003.07490. pages 9

[32] Gao J, Galley M, Li L. Neural approaches to conversational AI; 2018. . pages 10

[33] Gao S, Sethi A, Agarwal S, Chung T, Hakkani-Tur D. Dialog state tracking: A neural reading comprehension approach. In: SIGDIAL 2019 - 20th Annual Meeting of the Special Interest Group Discourse Dialogue - Proceedings of the Conference; 2019. Available from: https://doi.org/10.18653/v1/w19-5932. pages 10

[34] Fazel-Zarandi M, Li SW, Cao J, Casale J, Henderson P, Whitney D, et al. Learning Robust Dialog Policies in Noisy Environments. 2017 12. pages 10

[35] Wen TH, Gašić M, Mrkšić N, Su PH, Vandyke D, Young S. Semantically conditioned lstm-based Natural language generation for spoken dialogue systems; 2015. . pages 11

[36] Wen TH, Gašić M, Kim D, Mrkšić N, Su PH, Vandyke D, et al. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking; 2015. . pages 11

[37] Wolf T, Sanh V, Chaumond J, Delangue C. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents; 2019. pages 11, 22

[38] Zhang S, Dinan E, Urbanek J, Szlam A, Kiela D, Weston J. Personalizing Dialogue Agents: I have a dog, do you have pets too? In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 2204-13. Available from: `https://aclanthology.org/P18-1205`. pages 11, 14

[39] Rashkin H, Smith EM, Li M, Boureau YL. Towards empathetic open-domain conversation models: A new benchmark and dataset; 2020. . pages 13, 19, 22

[40] Kim H, Kim B, Kim G. Perspective-taking and Pragmatics for Generating Empathetic Responses Focused on Emotion Causes; 2021. Available from: `http://arxiv.org/abs/2109.08828`. pages 13

[41] Li Y, Li K, Ning H, Xia X, Guo Y, Wei C, et al. Towards an Online Empathetic Chatbot with Emotion Causes. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM; 2021. . pages 13

[42] Ghandeharioun A, McDuff D, Czerwinski M, Rowan K. EMMA: An Emotion-Aware Wellbeing Chatbot. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII); 2019. p. 1-7. pages 13

[43] Lin Z, Xu P, Winata G, Siddique F, Liu Z, Shin J, et al. CAiRE: An end-to-end empathetic chatbot. vol. 34; 2020. p. 13622-3. pages 13

[44] Zhou L, Gao J, Li D, Shum HY. The design and implementation of xiaoice, an empathetic social chatbot. Computational Linguistics. 2020;46:1-62. pages 13, 38

[45] Ghaznavi I, Jehanzeb U, Gillies D. Usability evaluation of an immersive virtual reality platform for self-attachment psychotherapy; 2019. . pages 13

[46] Smith EM, Williamson M, Shuster K, Weston J, Boureau YL. Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills. arXiv; 2020. Available from: `https://arxiv.org/abs/2004.08449`. pages 14

[47] Wu Z, Balloccu S, Kumar V, Helaoui R, Reiter E, Reforgiato Recupero D, et al. Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022. p. 6177-81. pages 15, 38

[48] Blodgett SL, Barocas S, Daumé I, Wallach H. Language (Technology) is Power: A Critical Survey of "Bias" in NLP; 2020. pages 16

[49] Gehman S, Gururangan S, Sap M, Yejin C, Smith N. REALTOXICITYPROMPTS: Evaluating neural toxic degeneration in language models. Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2020. 2020 09. pages 16

[50] Henderson P, Sinha K, Angelard-Gontier N, Ke NR, Fried G, Lowe R, et al. Ethical Challenges in Data-Driven Dialogue Systems. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. AIES '18. New York, NY, USA: Association for Computing Machinery; 2018. p. 123–129. Available from: `https://doi.org/10.1145/3278721.3278777`. pages 16

[51] Ischen C, Araujo TB, Voorveld HAM, van Noort G, Smit EG. Privacy Concerns in Chatbot Interactions; 2019. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 11970 LNCS. Available from: https://doi.org/10.1007/978-3-030-39540-7_3. pages 16

[52] Larson S, Mahendran A, Peper JJ, Clarke C, Lee A, Hill P, et al.. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. arXiv; 2019. Available from: `https://arxiv.org/abs/1909.02027`. pages 18

[53] Kim E, Klinger R. A Survey on Sentiment and Emotion Analysis for Computational Literary Studies; 2018. pages 20

[54] Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text. Social Network Analysis and Mining;11(1):81. Available from: `https://doi.org/10.1007/s13278-021-00776-6`. pages 20

[55] Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations; 2019. p. 527-36. pages 20

[56] Kim T, Vossen P. EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa; 2021. pages 21

[57] Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, et al. Recursive deep models for semantic compositionality over a sentiment treebank. EMNLP. 2013 01;1631:1631-42. pages 21

[58] Li Y, Su H, Shen X, Li W, Cao Z, Niu S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. arXiv; 2017. Available from: `https://arxiv.org/abs/1710.03957`. pages 22

[59] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for Hyper-Parameter Optimization. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ, editors. Advances in Neural Information Processing Systems. vol. 24. Curran Associates, Inc.; 2011. Available from: `https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf`. pages 28

[60] Liu CW, Lowe R, Serban I, Noseworthy M, Charlin L, Pineau J. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation; 2016. p. 13. pages 30

[61] Cavalin P, Alves Ribeiro VH, Appel A, Pinhanez C. Improving Out-of-Scope Detection in Intent Classification by Using Embeddings of the Word Graph Space of the Classes. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 3952-61. Available from: `https://aclanthology.org/2020.emnlp-main.324`. pages 36

[62] Zhan LM, Liang H, Liu B, Fan L, Wu XM, Lam AYS. Out-of-Scope Intent Detection with Self-Supervision and Discriminative Training. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics; 2021. p. 3521-32. Available from: `https://aclanthology.org/2021.acl-long.273`. pages 37

[63] Pichl J, Marek P, Konrád J, Matulík M, Šedivý J. Alquist 2.0: Alexa Prize Socialbot Based on Sub-Dialogue Models. 2020 04. Available from: `https://arxiv.org/abs/2011.03259`. pages 37