École Polytechnique Fédérale de Lausanne

Developing an Empathetic Chatbot for Distress Relief

by Yeh Chun-Hung

Master Thesis

Approved by the Examining Committee:

Dr. Pearl Pu Faltings
Thesis Supervisor

Dr. Igor Kulev
External Expert

Kalpani Anuradha Welivita
Thesis Assistant

Electrical and Electronic Section
EPFL IC SCI-IC-PFP Group
INR 214 (Bâtiment IN)
Year 2022

January 17, 2022

Acknowledgments

I would first like to thank my thesis supervisor Dr. Pearl Pu of the Human Computer Interaction Group. She consistently allowed this thesis to be my own work, but steered me in the right direction whenever she thought I needed it.

In addition, I have to thank my second supervisor Kalpani Anuradha Welivita. Without her assistance and dedicated involvement in every step throughout the process, this work would have never been accomplished. I would like to thank you very much for your support and understanding over these months.

I would also like to acknowledge Dr. Igor Kulev at Roche for being the external expert of my thesis, and I am gratefully indebted to him for his valuable review on this thesis.

Finally, I must express my very profound gratitude to my parents for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Lausanne, January 17, 2022 Yeh Chun-Hung

Abstract

In modern era, people more often suffer from severe emotional distress, which may result from the risks such as relationship difficulties, financial strain, or chronic medical illness. Being unable to recover from it may potentially lead to self-destructive behaviors or even suicide. Although therapeutic consultations are available to assist people in distress, most of them are required to be synchronous and face-to-face with therapists. Moreover, in the recent years, more and more people may choose to use text-based platforms for their mental health support. Therefore, the task of empathetic conversation agents has been a popular research topic aiming at generating syntactically and emotionally appropriate responses following dialog contexts.

In this work, we develop multi-turn empathetic dialog models which can not only recognize human emotions but also rely on the well developed cognitive framework to assess expressed empathy in texts. The experiments reveal that our models can evidently generate empathetic responses in accordance with the context of speaker utterances. Lastly, we benchmark them with the baselines of two previous works based on automatic evaluation and human assessment respectively.

Contents

Chapter 1

Introduction

Around the globe, mental health disorders have affected 13% of the worldwide population [1]. The number could increase as people shelter in place and adjust to a new normal amid the coronavirus pandemic. According to the US national survey [2], the COVID-19 pandemic has seriously affected the mental health of 59% of people. Despite the availability of mental health services, people may hesitate to reach for the support because of the public stigma. Even worse, as the consequence of the pandemic, there is a severe shortage of mental health workers [3]. Thus, it is challenging to offer interventions using traditional one-on-one therapeutic methods. To mitigate this, insufficient services have facilitated the utilization of technology to meet their need. One technological solution is the chatbot, a dialog system capable of conversing with human users naturally .

Dialog systems have become popular in the artificial intelligence community. The recent advance shows that deep neural networks can be effectively applied to their development [4]. Nowadays, most cutting-edge systems are able to generate the responses based on both syntactic and contextual points of view. For instance, open-domain chatbots are the systems designed for extended conversations, set up to mimic the unstructured conversational characteristic of the human-human interaction. By contrast, task-oriented dialog systems use conversations with users to help complete tasks. They usually gives a short answer to satisfy users' inquiry. The most common example is the digital assistant (Siri, Alexa, or Google Home) to give directions, control appliances, or make calls. To build a tool aiming at assisting people in distress, we adopt the methodology of the chatbots to understand users' feelings and talk to them comfortingly.

However, a challenge in empathy-focused chatbots is to identify emotions and make a suitable reply accordingly. One of the major reasons is the inadequacy of a large-scale empathetic dialog dataset. To make chatbots communicate more empathetically, it is essential to train them by substantial number of dialogs to grasp general understanding of emotional interaction embedded in human conversations. However, in the data aspect, some past studies [5, 6, 7, 8, 9] propose their datasets with thousands of conversations, which is obviously limited. Furthermore, these also have their dialogs labeled by the limited number of emotion categories. In the modeling aspect, various dialog models of the existing works [10, 11, 12, 13] often require manually crafted rules to determine the emotion state for a response to be generated. Yet, these rules lack the verification from psychology literatures as well as practicability when deploying a chatbot into real world.

To resolve the aforementioned data issue, here we make use of a larger curated dataset, RED (Reddit Emotional Distress), that contains emotional support for people in acute distress. Generally, RED is a text-based dialog corpus carefully scraped from the empathy-related channels in Reddit. As Figure 1.1 indicates, the whole curation process ranges from extracting conversations from 8 subreddits (e.g. r/depression, r/suicidewatch, or r/offmychest) to automatically annotating all dialog turns by the BERT based emotion classifier. In the extraction stage, the collected dialogs can be either dyadic or multiparty. If a dialog is dyadic, it is communication between a speaker and a listener. Otherwise, a multiparty dialog is inclusive of a speaker and multiple listeners. Based on the post structure in Reddit, the post author would play the role of speakers while the commenters are listeners. Up next, we conduct pre-processing to remove the irrelevant parts, such as URLs, tags, or even toxic words. For more specific details about the cleaning procedure, one can refer to the appendix. Lastly, we categorize the emotions by the emotion classifier to attach 32 emotion labels and 8 additional response intents to any utterance.

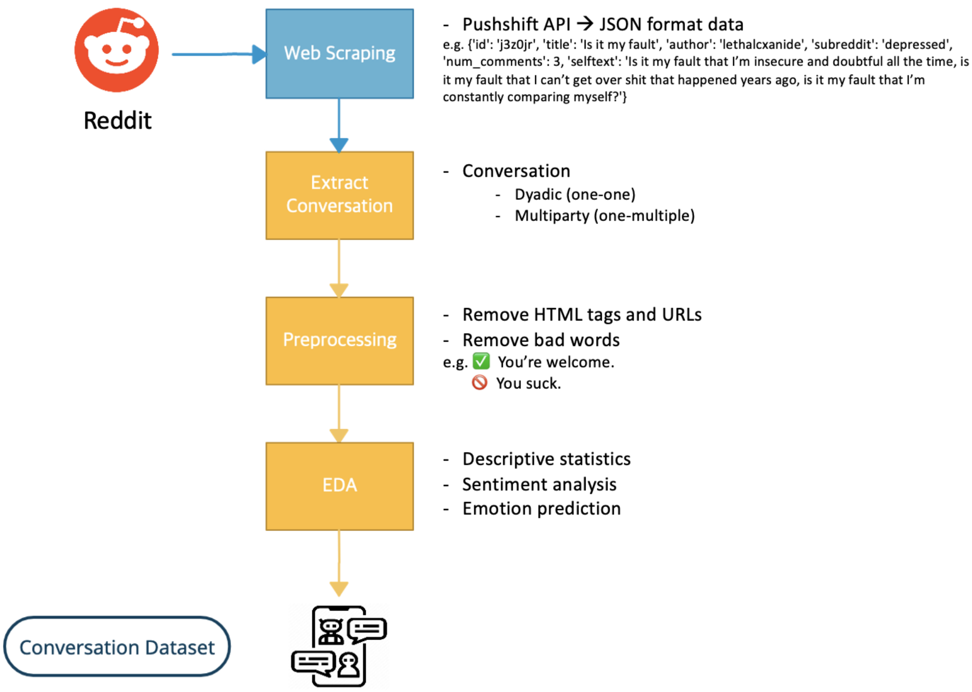

Figure 1.1: RED Data Curation Pipeline

With the dataset ready, we then implement an empathetic conversation agent targeting at learning the emotion interchange within conversations and correspondingly generate empathetic reactions. Specifically, the model encompasses three fundamental components: (1) a response emotion/intent predictor to determine the emotion/intent for the next utterance based on the context (2) an encoder responsible for encoding the dialog context to its vector representation (3) a decoder that generates the response based on the predicted response emotion/intent while consistently taking the encoder outputs into consideration. Besides, we incorporate a novel computational approach, *EPITOME* [14], to effectively identify empathetic conversations with underlying rationales. This framework has been proven by psychologists and demonstrate that the identified components are helpful in improving peer-to-peer support through model-based feedback.

Finally, we deliberately evaluate our dialog models via a set of automated objective metrics as well as the well designed human evaluation experiment on the crowdsourcing platform [15, 16]. The automated metrics we show in this work specify that our models outperform the baselines [17, 15] by the majority of them. On the other hand, through testing on 200 generated dialogs, the human judgement reveals that our models are partially better than the baselines. As a result, we point out some disadvantages of our models and further improvements to do in the end for future work.

Chapter 2

Background

Over the last decade, machine learning has been the fast growing field in the modern technology. With abundant data, people can utilize this methodology to facilitate their work without explicitly designing a precise algorithm. In particular, deep learning have been widely adopted to deal with unstructured data such as texts, audios, and images. For example, combined with natural language processing, we can apply neural networks to decide whether an email is spam or not [18]. Apart from the classification task, we are able to exploit neural nets to summarize a given paragraph [19].

One objective in deep learning is to develop a conversation agent which can recognize human feelings and reply accordingly. In the following sections, we cover some background knowledge to enhance the understanding of the neural based approach to build a chatbot eventually.

2.1 Chatbot

A chatbot is defined as the system that can carry on extended conversations with the goal of mimicking the conversations in human-human interaction. Based on its use, it is typically for entertainment, but also possibly for practical purposes to focus on answering well on a set of particular topics. Like everything else in language processing, chatbot architectures fall into two classes: rule-based systems and corpus-based systems. Rule-based systems mainly utilize the pre-defined rules to identify the utterance patterns to make responses. The famous examples include the early influential ELIZA and PARRY systems [20, 21]. On the other hand, corpus-based systems mine large datasets of human-human conversations, which mainly applies the sequence-to-sequence (Seq2Seq) framework to generate a response from a user utterance. Since it learns the pattern automatically through dialogs, the latter is now the mainstream for the chatbot research and development.

Last but not least, chatbots can also be developed by the hybrid architecture of rule-based and corpus-based systems. For instance, this technique is commonly used in the Amazon Alexa Prize challenge [22], where the teams build chatbots to converse with volunteers on the Amazon Alexa platform, and are scored based on the length and user ratings of their conversations. Regarding our models, we mostly leverage the corpus-based method since we have large empathetic dialog dataset. Ultimately, we expect that the models are trained using the collected conversations to learn how to generate empathy-oriented utterances.

2.2 Sequence-to-Sequence (Seq2Seq)

Seq2Seq is a type of standard modeling paradigm typically used in the sequence-to-sequence tasks [4]. The applications include language translation, text summarization, and conversation models. Particularly, Seq2Seq is a model composed of a set of encoders and decoders (Figure. 2.1). The encoders convert each item in a sequence to a corresponding hidden vector considering the current item and its context. The decoders reverse the process, turning the vector into an output sequence of items, using the previous output as the input context.

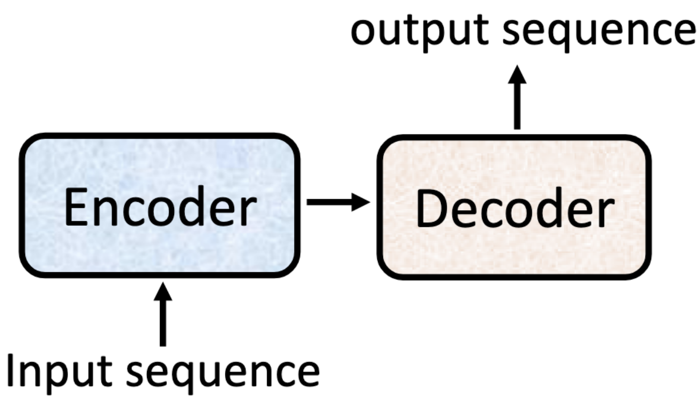

Figure 2.1: Seq2Seq Model Structure

To build this end-to-end model, both encoder and decoder can be done by the use of a recurrent neural network (RNN), or LSTM [23] and GRU [24] to avoid the vanishing gradient [25]. Though these models are firmly established, they still face the problem when handling long sequences. Expressly, these models require very large computation since each token in a sequence has to be processed sequentially. Therefore, this is where the Transformer [26] comes in to address the issue.

2.3 Transformer

In 2017, Vaswani et al. proposed the Transformer [26], a novel encoder-decoder architecture making a major breakthrough in natural language generation (Figure. 2.2). Different from RNN and LSTM, the Transformer extensively utilizes the attention mechanism to achieve its significant effectiveness. Additionally, the Transformer allows an input sequence to be passed in parallel so that the training speed can be sharply increased. In this section, we would like to explain the main components inside the Transformer to better understand its full mechanism.

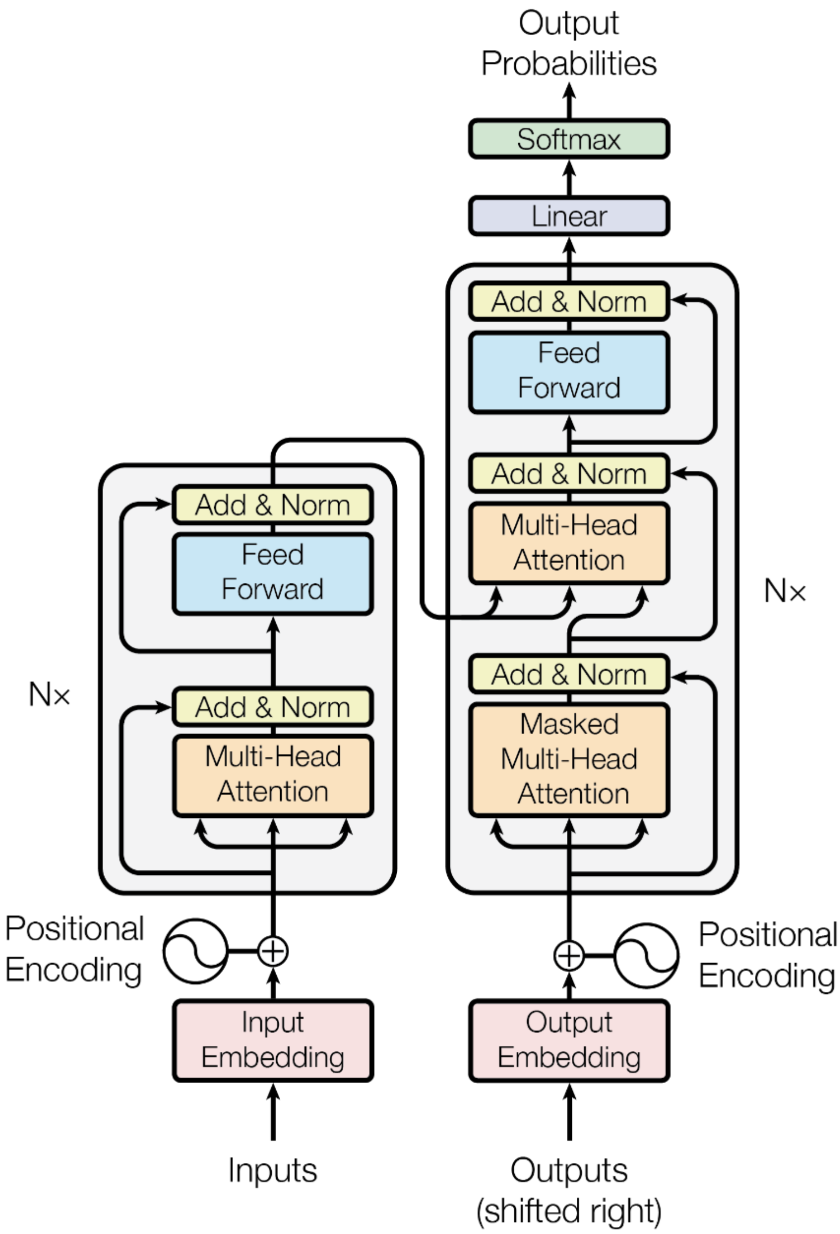

Figure 2.2: Model Architecture of the Transformer [26]

2.3.1 Attention

In general, the attention mechanism looks at an input sequence and decides which parts are important at each step. To achieve this, it requires a query and a set of key-value pairs, where the query, keys and values are all vectors. Then, the operation is to generate the output vector as a weighted sum of the values. The weights are determined by some compatibility function taking the query and its corresponding keys into consideration.

Scaled Dot-Product Attention

In the Transformer, the attention function is done by the scaled dot product between the query and all its keys. If the dimension of keys is d_k, the scaled dot-product attention is computed as the dot product of the query with all keys, followed by dividing each by $\sqrt{d_k}$. Lastly, apply a softmax function to obtain the weights. Note that the reason for adding the scaling term $\frac{1}{\sqrt{d_k}}$ is to prevent the dot product from getting the very large value and pushing the softmax function to the flat region with extremely small gradients.

Instead of running sequentially, the attention weights with respect to each query can be simultaneously computed. If we pack all the queries, keys, and values as the matrices Q, K, V respectively, the weights on the values can be:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

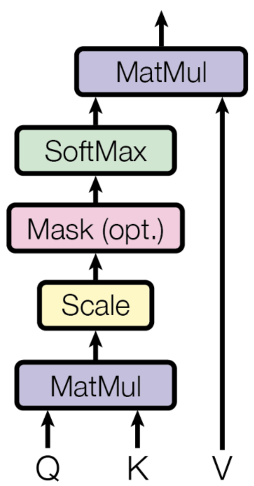

Figure 2.3: Workflow of Scaled Dot-Product Attention [26]

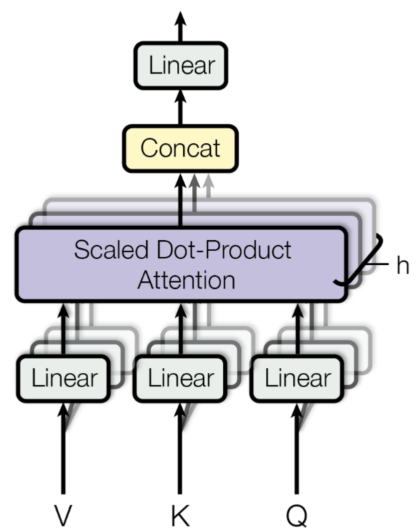

Figure 2.4: Multi-Head Attention with h Attention Layers Running in Parallel [26]

Multi-Head Attention

To improve the performance, the Transformer also exploits the multi-head attention to capture the diverse representations from an input sequence. In other words, it is beneficial to linearly project the queries, keys, and values multiple times with various linear projections. Because of these projected versions of queries, keys, and values, we are able to perform the attention functions in parallel. The output is subsequently concatenated and fed to the final linear layer to generate the attention weights. Lastly, the set of weights considers the information from different representations at different positions, where the single attention head cannot accomplish.

The Transformer employs 8 attention heads, so we end up with 8 sets existed in each encoder and decoder. Each head is used to project the input embeddings into different representation subspaces. As a result, the model can take the different information from these 8 heads into account.

Self-Attention

The self-attention is a key building block in the Transformer. Compared to the original attention mechanism, the self-attention allows inputs to interact with each other and find out who they should pay more attention to. Finally, the outputs are aggregates of these interactions and attention weights. To resolve the tasks involving very long sequence, the self-attention could be restricted to handling only the neighborhood with the smaller size of the input sequence around the specific output position.

Positional Encoding

Another innovation in the Transformer is to add positional encoding to the embedding of an input sequence. Without it, the input embeddings lack the ordinal information about how tokens locate in a sentence. To compensate this, positional encoding explicitly encodes the relative positions of the inputs as vectors and are then added to the input embeddings.

In the Transformer, positional encoding is built upon sine and cosine functions with different frequencies:

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{\frac{2i}{d_{model}}}})$$
$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{\frac{2i}{d_{model}}}})$$

where pos is the position and i is the dimension. Basically, each dimension of the positional encoding is a wave with a different frequency. This allows the model to easily learn to attend to relative positions, since $PE_{(pos+k)}$, where k is some integer, can be represented as a linear function of $PE_{(pos)}$.

2.3.2 Encoder/Decoder

As a Seq2Seq model, the Transformer is constructed by a stack of encoders and decoders. Here, the encoder intends to map an input sequence to another embedded sequence considering the information from the neighboring tokens. On the other hand, given the embedded sequence from the encoder, the decoder tends to generate an output sequence of tokens one at a time. In this section, we go through the structure of the encoder and decoder in the Transformer.

Encoder

The encoder segment of the Transformer consists of a stack of 6 encoding layers. Each encoding layer contains 2 major components, multi-head attention and fully connected feed-forward neural network (See the left side of Figure. 2.2). Multi-head attention will generate the weighted version of the value vectors considering the attention scores. However, before feeding it to the feed-forward neural net, the output of the attention layer has to be passed into a sub-layer (i.e. the "Add & Norm" layer in Figure. 2.2) for the residual connection and layer normalization. The residual connection eases the gradient flow through a network and allows stacking multiple layers [27]. The layer normalization does normalize vector representation of each example in batch to improve convergence stability and sometimes even quality [28]. Once we have the output of the sub-layer, feed it to the feed-forward neural net followed by another sub-layer. After going through the first encoding layer, we can then use the current output to run the other 5 encoding layers.

Decoder

Similar to the encoder structure, the decoder segment is also composed of 6 identical layers. (See the right side of Figure. 2.2) In addition to the ordinary multi-head attention, the attention component in the decoder would be auto-regressive. That is, the component prevents the current position from referring to the following positions. The masking ensures that the prediction of the current position is certainly based on the known outputs from the previous positions.

2.4 BERT

2.4.1 BERT

BERT [29], which stands for Bidirectional Encoder Representations from Transformer, is a language model designed to generate contextualized representations from texts. Unlike any other models looking at a text sequence from either left-to-right or combined left-to-right and right-to-left, BERT is bidirectionally trained. In other words, the model has a deeper sense of contexts and the flow compared to the single-direction language models. BERT's architecture highly relies on the Transformer. Specifically, BERT contains only the encoder part of the Transformer. (Figure 2.5) The input to the encoder is tokens, which are firstly converted into vectors and then processed in the neural network. However, before processing, BERT needs the input to be massaged and decorated with some extra metadata (Figure 2.6):

- **Token Embedding** A [CLS] token is added to the input word tokens at the beginning of the first sentence and a [SEP] token is inserted at the end of each sentence.
- **Segment Embedding** A marker indicating Sentence A or Sentence B is added to each token. This allows the encoder to distinguish between sentences.
- **Positional Embedding** A positional embedding is added to each token to indicate its position in the sentence.

Instead of sequentially predicting the next token, BERT applies the following techniques during training:

- **Masked LM** Randomly mask out 15% of the words in the input and replace them with [MASK] token. Afterwards, run the entire sequence through the BERT attention based encoder and then predict the masked words based on the context by the non-masked words in the sequence.
- **Next Sentence Prediction** Separate sentences using the [SEP] token. During training, the model is fed with two input sentences at a time such that: (1) 50% of the time the second sentence comes after the first one. (2) 50% of the time it is a random sentence from the full corpus. Then, BERT predicts if the second sentence is random, with the assumption the random sentence will be disconnected from the first sentence. To predict if the second sentence is connected to the first one, the input sequence goes through the Transformer based model, the output of [CLS] token is converted into a 2×1 shaped vector using a simple classification layer, and IsNext-Label is assigned by the softmax function.

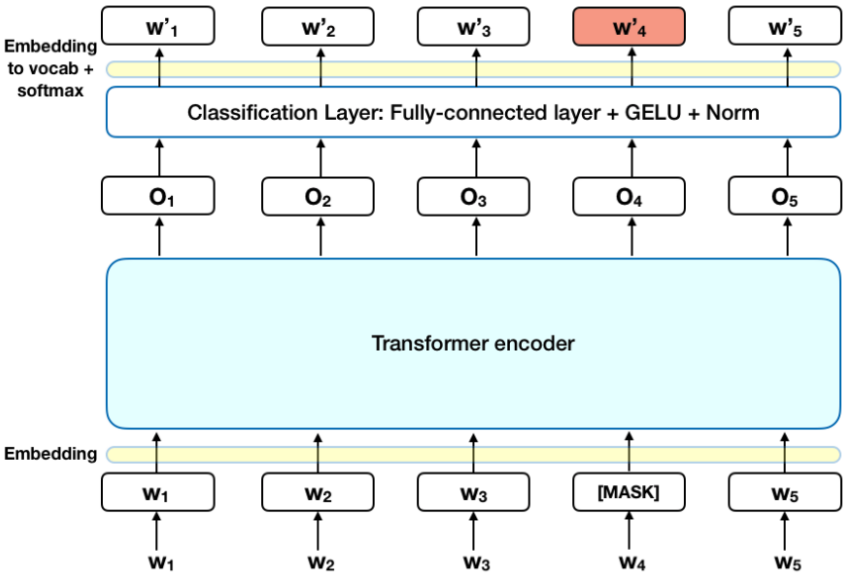

Figure 2.5: High-level depiction showing BERT's whole operating process involving only the Transformer encoders to compute contextualized representations of input texts.

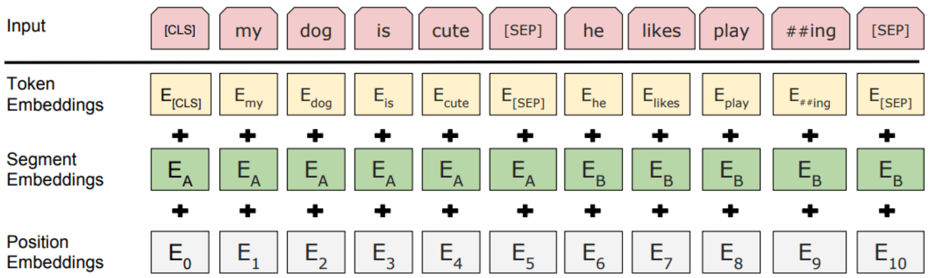

Figure 2.6: BERT's input representation is the sum of the token embeddings, the segmentation embeddings and the position embeddings

2.4.2 RoBERTa

RoBERTa [30], Robustly Optimized BERT Pre-training Approach, is to optimize the training of the BERT to take less time during pre-training. It has the similar structure as the BERT. However, because the BERT is highly undertrained, there are some changes in its structure for further improvement:

- **Removing the Next Sentence Prediction objective** In the next sentence prediction, the model is trained to predict whether the observed document segments come from the same or distinct documents via its loss. The paper [30] concludes that removing the loss improves downstream task performance.
- **Training with larger batches and longer sequences** Originally, BERT is trained for 1M

steps with a batch of 256 sequences. On the other hand, during training, RoBERTa has 125 steps of 2K sequences and 31K steps with 8K sequences in the batches. This brings two benefits. Firstly, training with the larger batches improves the perplexity on the masked language modeling objective and end-task accuracy. Secondly, large batches are easier for distributed parallel training.

- **Dynamically changing the masking pattern** In BERT, the masking is done during data pre-processing, resulting in a single static mask. To avoid using the static mask, the training data is duplicated and masked 10 times, each time with a different mask over 40 epochs of training. As a result, the same mask can be used four times on each sequence during training.

- **Larger byte-level Byte-Pair Encoding** Byte-Pair Encoding (BPE) [31] is a hybrid of character- and word-level representations which allows dealiing with the large number of common vocabularies in corpora. Instead of using character-level BPE vocabulary of size 30K, RoBERTa is trained with a larger byte-level BPE [32] vocabulary with 50K sub-word units, without any pre-processing or tokenization on an input. Though the experiments reveal that this method achieves slightly worse performance on some tasks, the authors believe that the advantages of a universal encoding can eventually outweigh the encoding using the smaller size.

When it comes to the performance, RoBERTa achieves the highest score across all the GLUE single-task development sets [33]. Compared with the models in the setting of ensembles on the test set, RoBERTa also reaches the state-of-the-art results on MNLI, QNLI, RTE, and STS of the GLUE. Furthermore, it outperforms the BERT-Large model and XLNet [34] on RACE [35] as well as the SQuAD [36] in the single model on the development set.

	MNLI	**QNLI**	**QQP**	**RTE**	**SST**	**MRPC**	**CoLA**	**STS**	**WNLI**	**Avg**
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-
Ensembles on test (from leaderboard as of July 25, 2019)										
ALICE	88.2/87.9	95.7	**90.7**	83.5	95.2	92.6	**68.6**	91.1	80.8	86.3
MT-DNN	87.9/87.4	96.0	89.9	86.3	96.5	92.7	68.4	91.1	89.0	87.6
XLNet	90.2/89.8	98.6	90.3	86.3	**96.8**	**93.0**	67.8	91.6	**90.4**	88.4
RoBERTa	**90.8/90.2**	**98.9**	90.2	**88.2**	96.7	92.3	67.8	**92.2**	89.0	**88.5**

Figure 2.7: Results on the GLUE tasks comparing RoBERTa with some cutting-edge models at that time

Model	SQuAD 1.1		SQuAD 2.0	
	EM	F1	EM	F1
Single models on dev, w/o data augmentation				
$\text{BERT}_{\text{LARGE}}$	84.1	90.9	79.0	81.8
$\text{XLNet}_{\text{LARGE}}$	**89.0**	94.5	86.1	88.8
RoBERTa	88.9	**94.6**	**86.5**	**89.4**
Single models on test (as of July 25, 2019)				
$\text{XLNet}_{\text{LARGE}}$			$86.3^{\dagger}$	$89.1^{\dagger}$
RoBERTa			86.8	89.8
XLNet + SG-Net Verifier			$\mathbf{87.0}^{\dagger}$	$\mathbf{89.9}^{\dagger}$

Figure 2.8: Results on SQuAD comparing BERT, XLNet, and RoBERTa.

Model	Accuracy	Middle	High
Single models on test (as of July 25, 2019)			
$\text{BERT}_{\text{LARGE}}$	72.0	76.6	70.1
$\text{XLNet}_{\text{LARGE}}$	81.7	85.4	80.2
RoBERTa	**83.2**	**86.5**	**81.3**

Figure 2.9: Results on RACE.

Chapter 3

Methodology

To endow a chatbot with empathy, our architecture is composed of three main modules: the response emotion/intent predictor, the encoder and the decoder of the response generator. Specifically, the predicted emotion/intent is the input to the encoder and the decoder to condition the response to be generated in the end. In the following sections, we discuss the dataset used in our experiments, neural response emotion/intent predictor based on dialog contexts, and the Transformer based response generation model. Moreover, we introduce the theoretically-grounded framework, *EPITOME*, [14] applied in our models for characterizing communication of expressed empathy within conversations.

3.1 Dataset

Many existing dialog models rely on small datasets, usually with the size of thousands dialogs. Consequently, training models solely on them may not give us a chatbot with desirable performance. To address the issue, we create the new dataset, RED, whose dialogs are completely scraped from Reddit via Pushshift APIs [37]. In particular, we focus on the threads posted on 8 major mental health subreddits to collect conversations involving strong emotion exchanges. Based on the curation process in the figure 1.1, RED contains approximately 2M dialogs and 7M emotional interactions. (See Table 3.1, 3.2 for more details about the descriptive statistics.) Note that the emotion annotation is done by EmoBERT [38], a RoBERTa based classifier to predict 32 emotion labels plus 8 response intents given a dialog sentence. This classifier is trained on 25K conversations from the EmpatheticDialogue corpus [17] labeled with 41 emotion/intent labels. Overall, EmoBERT has annotation accuracy of 65.88%, which is significant compared to the other state-of-the-art dialog emotion predictors. Hence, we take advantage of the labels suggested by the classifier as ground-truth labels.

Due to the hardware constraint, the dialog corpus would not be fully utilized for building the

Subreddit	No. of Dialogs	No. of Turns	No. of Tokens	Avg No. of Turns per Dialog	Avg No. of Tokens per Dialog	Avg No. of Tokens per Turn
All	1,275,486	3,396,476	288,336,762	2.66	226.06	84.89
r/depression	510,035	1,396,044	106,967,833	2.74	209.73	76.62
r/depressed	10,892	23,804	1,940,000	2.19	178.11	81.50
r/offmychest	437,737	1,064,467	109,459,738	2.43	250.06	102.83
r/sad	18,827	42,293	3,088,562	2.25	164.05	73.03
r/SuicideWatch	262,469	791,737	59,267,000	3.02	225.81	74.86
r/depression_help	23,678	51,849	5,412,390	2.19	228.58	104.39
r/Anxietyhelp	8,297	18,351	1,428,287	2.21	172.14	77.83
r/MentalHealthSupport	3,551	7,931	772,952	2.23	217.67	97.46

Table 3.1: Descriptive statistics of dyadic conversations in the entire dataset as well as in each subreddit.

Subreddit	No. of Dialogs	No. of Turns	No. of Tokens	Avg No. of Turns per Dialog	Avg No. of Tokens per Dialog	Avg No. of Tokens per Turn
All	584,427	3,863,841	195,187,799	6.61	333.98	50.52
r/depression	246,268	1,609,795	76,789,493	6.54	311.81	47.70
r/depressed	3,434	18,658	923,429	5.43	268.91	49.49
r/offmychest	196,566	1,232,645	69,483,975	6.27	353.49	56.37
r/sad	6,756	35,085	1,577,803	5.19	233.54	44.97
r/SuicideWatch	119,577	899,460	42,468,629	7.52	355.16	47.22
r/depression_help	7,758	45,446	2,791,784	5.86	359.86	61.43
r/Anxietyhelp	2,990	16,959	825,710	5.67	276.16	48.69
r/MentalHealthSupport	1,078	5,793	326,976	5.37	303.32	56.44

Table 3.2: Descriptive statistics of multi-party conversations in the entire dataset as well as in each subreddit.

Subreddit	No. of Dialogs	No. of Turns	No. of Tokens	Avg No. of Turns per Dialog	Avg No. of Tokens per Dialog	Avg No. of Tokens per Turn
All	558,316	1,333,100	28,943,621	2.77	64.30	22.73
r/depression	209,927	480,079	10,237,964	2.29	48.79	21.33
r/depressed	5,618	12,867	286,092	2.29	50.92	22.23
r/offmychest	200,998	476,712	10,590,386	2.37	52.67	22.22
r/sad	11,356	31,917	743,386	2.81	65.46	23.29
r/SuicideWatch	116,304	269,094	5,506,485	2.31	47.35	23.29
r/depression_help	9,184	51,312	1,321,585	5.59	143.9	25.76
r/Anxietyhelp	3,505	7,870	181,400	2.25	51.75	23.05
r/MentalHealthSupport	1,424	3,249	76,323	2.28	53.59	23.49

Table 3.3: Descriptive statistics of dyadic conversations after preprocessing.

chatbot. We only allow the conversations within 100 tokens to be used for training our dialog model. In addition, we exclude the multiparty conversations because we expect the conversation between users and our chatbot is mutual. Therefore, the available corpus narrows to 560K dialogs. In the table 3.3, 3 main subreddits (r/depression, r/offmychest, and r/suicidewatch) contribute almost all the conversations in the dataset. Regarding the emotion distribution (Figure 3.1 and 3.2), the speakers tend to convey negative feelings, such as *ashamed*, *sad*, and *lonely*, while the listeners are inclined to express positive intents like *questioning*, *agreeing* and *wishing* to support the speakers. All in all, our dataset shows that the listener responses are indeed being helpful for the speakers.

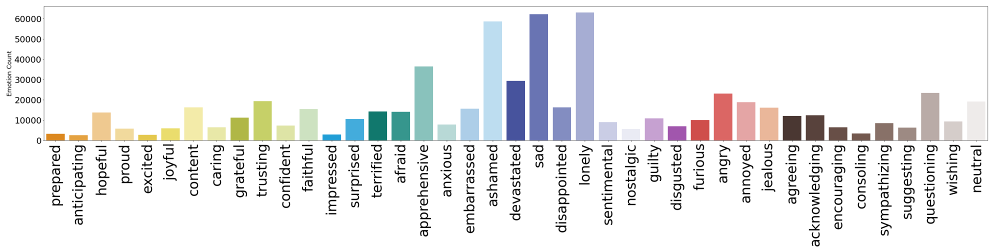

Figure 3.1: Speaker emotion prediction in RED after cleaning

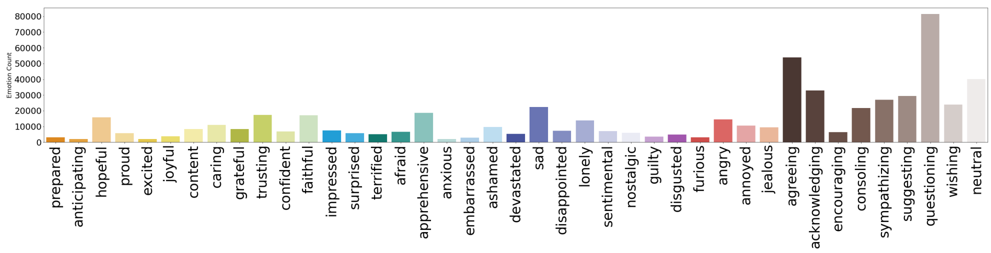

Figure 3.2: Listener emotion prediction in RED after cleaning

3.2 Framework of Expressed Empathy

To understand empathy conveyed in asynchronous text based conversations, we turn to *EPITOME* [14], a conceptual framework incorporating both emotional and cognitive aspects of empathy. Notably, it is approved by clinical psychologists. *EPITOME* comprises of three communication mechanisms exhibiting a comprehensive outlook about empathy — *Emotional Reaction*, *Interpretation*, and *Exploration*. Each mechanism can be in the level of no communication, weak communication, and strong communication.

- **Emotion Reaction** It is particularly aimed at expressing emotions experienced by the listeners after understanding speakers' words. For example, after reading the speaker text, a listener could evoke the emotions such as encouraging, acknowledging, or concern. According to the study by Robert et al. [39], showing these emotions can truly help establish a strong bond to support the people in distress. A weak level of emotion reactions refers to the situation when the emotions are not explicitly revealed. (e.g. Things will be better.) On the other hand, the strong level tends to specially stress the emotions from the listeners. (e.g. I'm sorry to hear that.)
- **Interpretation** This focuses on the realization of emotions and experiences inferred from speakers. In order to develop empathetic rapport, communicating with a clear understanding on the speaker situation is exceptionally critical. A weak interpretation literally mentions the word 'understanding' (e.g. I could understand how you've been through) while a strong one is inclined to have listeners share their real-life experiences. (e.g. I also suffer from anxiety from time to time, which makes me extremely terrified.)
- **Exploration** To improve the understanding regarding their experiences, it is also essential to explore the feelings not stated by the speakers. Exploration makes the speakers perceive that the listeners are actively interested in themselves. Its weak communication is often generic (e.g. What's going on?); however, the strong communication is to the point and labels the speaker feelings which the listeners want to probe. (e.g. Do you feel frightened now?)

3.2.1 Modeling Approach

To capture the communication levels of the RED dialog utterances, we use the multi-task bi-encoder model built upon RoBERTa. Concretely, the model multi-task over the two tasks of empathy identification and rationale extraction. If we have a speaker post and the corresponding response post as $S_i = s_{i1}, ...s_{im}$ and $R_i = r_{i1}, ...s_{in}$, for each pair (S_i, R_i), empathy identification is to identify how empathetic R_i is in the context of S_i. According to *EPITOME*, the levels in R_i can be no communication (0), weak communication (1), or strong communication (2). Once the communication levels are measured, we then apply rationale extraction to extract rationales underlying the identified levels. The goal is to find an extracted rationale as a subsequence of words x_i in R_i. To do so, the bi-encoder model learns to predict a mask $m_i = (m_{i1}, ..., min)$ over the words in R_i. Note that $m_{ij} \in 0, 1$ is a boolean variable with 0 as a non-rationale token and 1 as rationale token. Thus, the extracted rationale is $x_i = m_i \odot R_i$.

As for modeling, the multi-task model trains three identical but independent architectures for the empathy communication levels in *EPITOME*. The bi-encoder structure facilitates the joint modeling of (S_i, R_i) pairs. Furthermore, the attention mechanism boosts the model by providing the context from the speaker post S_i, which is more effective than concatenating S_i and R_i to form a single sequence. As shown in Figure 3.3, for each communication level we use two independently pretrained RoBERTa based encoders, S-Encoder and R-Encoder, to encode both speaker post and response post. The S-Encoder encodes the context of the speaker post and R-Encoder encodes the response post. Then, we leverage attention between them to generate speaker-context aware representation of the response post, which is subsequently used to perform the two tasks of empathy identification and rationale extraction. Note that any speaker post and response post should be concatenated by the special tokens `[CLS]` and `[SEP]` at the beginning and at the end of the posts respectively.

$$
\begin{aligned}
e_i^{(S)} &= \text{S-Encoder}\,(\texttt{[CLS]}, S_i, \texttt{[SEP]}) \\
e_i^{(R)} &= \text{R-Encoder}\,(\texttt{[CLS]}, R_i, \texttt{[SEP]})
\end{aligned}
$$

With $e_i^{(S)}$ and $e_i^{(R)}$ encoded from S_i and R_i, the single-head attention is applied on the two encodings to generate the representation of the response post based on the context of the speaker post. In the Transformer terminology, the query vector is the response post encoding $e_i^{(R)}$. The keys and the values are the speaker post encoding $e_i^{(S)}$. Then, the attention score between the query and the keys is:

$$
a_i(e_i^{(\mathbf{R})}, e_i^{(\mathbf{S})}) = \text{softmax}\,(\frac{e_i^{(\mathbf{R})} e_i^{(\mathbf{S})}}{\sqrt{d}})\, e_i^{(\mathbf{S})}
$$

where d is the hidden size in the RoBERTa-base model ($d = 768$). Lastly, we sum up the encoded response $e_i^{(\mathbf{R})}$ with its representation transformed via attention $_i(e_i^{(\mathbf{R})}, e_i^{(\mathbf{S})})$ to get a residual mapping $h_i^{\mathbf{R}}$ [27]. This forms the final speaker-context aware representation for the response post.

When identifying empathy, we use the final representation of the `[CLS]` token in the response post and pass it through a linear layer to obtain the prediction of the empathy levels. Next, to extract rationales based on the predicted level, we utilize the final representations of the individual tokens in $\mathbf{R}_i$, and feed them to the linear layer to make boolean predictions over each token.

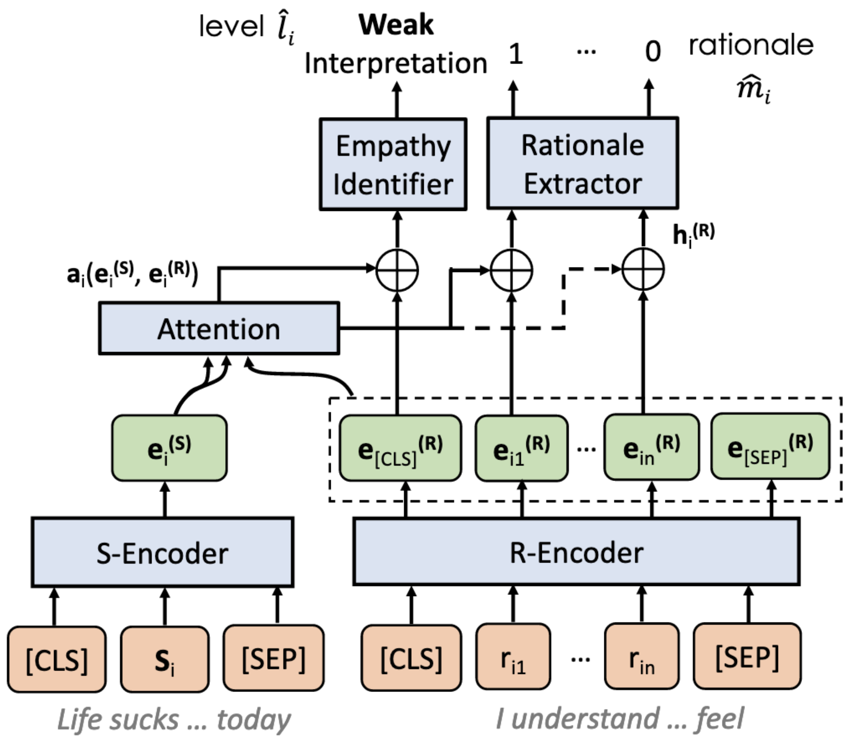

Figure 3.3: The bi-encoder model structure to evaluate the three communication mechanisms in *EPITOME* — emotion reaction, interpretation, and exploration. For each mechanism, the model can identify its empathy level as well as the rationale mask to highlight the words that make the response more empathetic. [14]

3.2.2 Implementation

Regarding the modeling approach to measure empathy, *EPITOME* leverages TalkLife and Reddit as the two major sources to collect the dialogs. However, because of the privacy protection, the TalkLife dataset is restricted to access without the consent of the platform. As a result, we could only use the Reddit portion of the collected dataset, which contains roughly 10k conversational interactions.

Both the S-Encoder and R-Encoder are initialized by the weights from the RoBERTa base model. Including the parameters in the linear layers, the total number of parameters in the bi-encoder model is 251M. The loss function used over the two tasks is cross entropy, so the overall loss in the model is $L = \lambda_{\mathbf{EI}} * L_{\mathbf{EI}} + \lambda_{\mathbf{RE}} * L_{\mathbf{RE}}$ where **EI** and **RE** represent empathy identification and rationale extraction separately. In addition, some initialization performs consistently well [40], so the seed value for randomization is set to be 12. As for the hyperparameters, the following space are also tested to fine-tune the model:

- Learning rate: 10^{-5}, $2 \times 10^{-5}, 5 \times 10^{-5}, 10^{-4}, 5 \times 10^{-4}$

- Loss weights: $\lambda_{\textbf{EI}} = 1;\ \lambda_{\textbf{RE}} = 0.1,\ 0.2,\ 0.5,\ 1$

3.3 Empathetic Dialog Model

To generate empathetic responses given one or more utterances, a dialog model should consider contexts and more importantly respond with an appropriate emotion. Thus, extended from the model proposed by Xie et al. [15], our models contain three essential building blocks (Figure. 3.4):

- A response emotion/intent predictor taking not only an input utterance X but also the total communication level of X into account. The predictor then decides a response emotion/intent based on the dialog context.
- A Transformer encoder in charge of encoding X into a vector representation. Note that during encoding we underscore the word empathy level using the rationale mask generated from the *EPITOME* bi-encoder model.
- A Transformer decoder responsible for decoding the encoded representation from the encoder while conditioning on the predicted response emotion/intent.

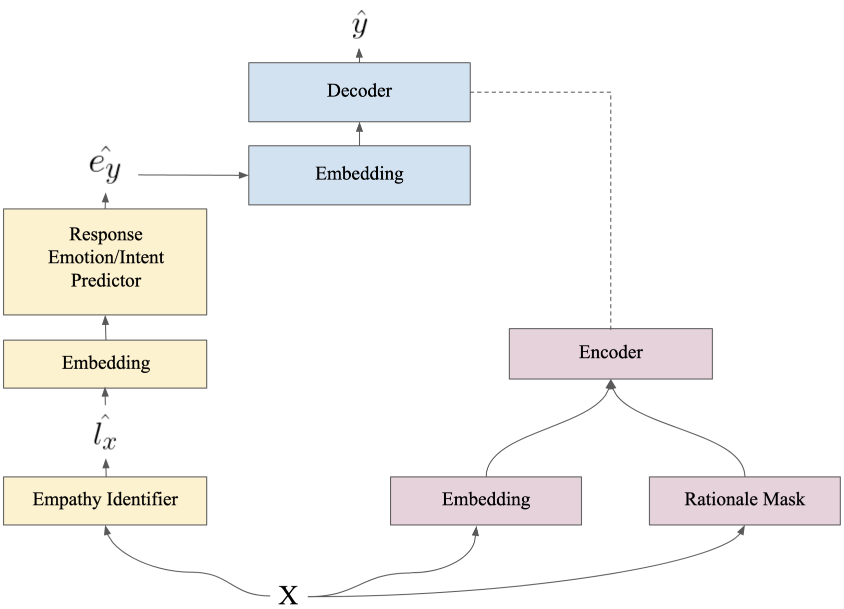

Figure 3.4: Complete model architecture working in the inference mode. The dashed line represents the multi-head attention.

3.3.1 Response Emotion/Intent Predictor

As illustrated in Figure 3.6, the response emotion/intent predictor is constructed on top of a Transformer encoder to obtain the context-inclusive representation from input embedding. Before the encoder, we apply the RoBERTa tokenizer to tokenize context utterances. The input representation is formed by concatenating the utterances by two special tokens: <s> and </s>. Apart from the original embeddings in [15], we add one more communication embedding for a better understanding of the context. (See Figure 3.5 for the overall depiction of the input embeddings.) The communication embedding is computed by the total communication level of emotion reaction, interpretation, and exploration. Given an utterance with the three communication levels $\hat{l}_{x,ER}$, $\hat{l}_{x,IP}$, and $\hat{l}_{x,EX}$, the total level is their sum as $\hat{l}_x = \hat{l}_{x,ER} + \hat{l}_{x,ER} + \hat{l}_{x,ER}$. Similar to the other embeddings, we transform this total level to a vector representation with the same dimension.

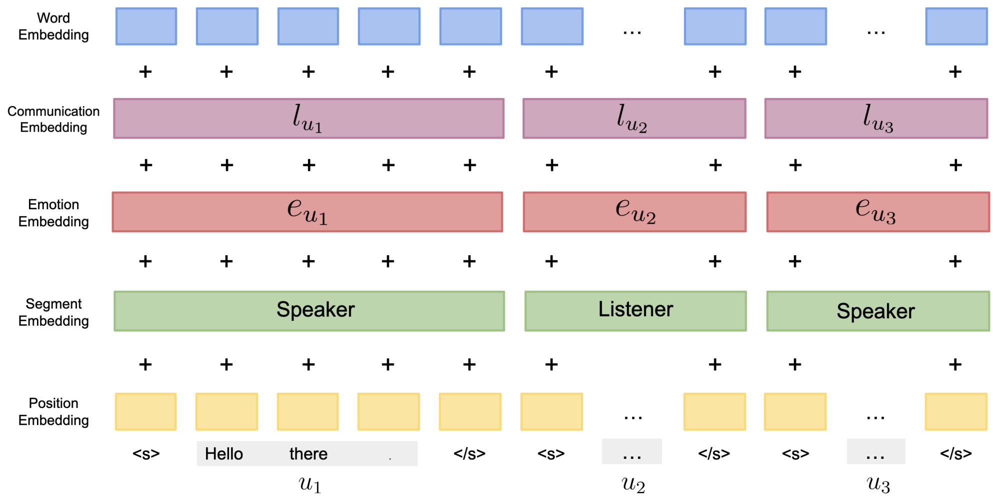

Figure 3.5: Illustration of 5 input embeddings.

Next, the encoder receives these embeddings as the input and outputs high-level vector representations. To pool them into a single vector, we apply the attention mechanism using a trainable vector υ to get the attention weight α_i for each r_i.

$$\alpha_i = \frac{exp(\upsilon^T r_i)}{\sum_{j=1}^{N} exp(\upsilon^T r_j)}$$

Finally, the pooled representation r becomes the weighted sum of r_i and α_i. Then r is treated as the input to the hidden layer followed by the softmax layer to determine $\hat{e}_y$, the predicted response emotion/intent.

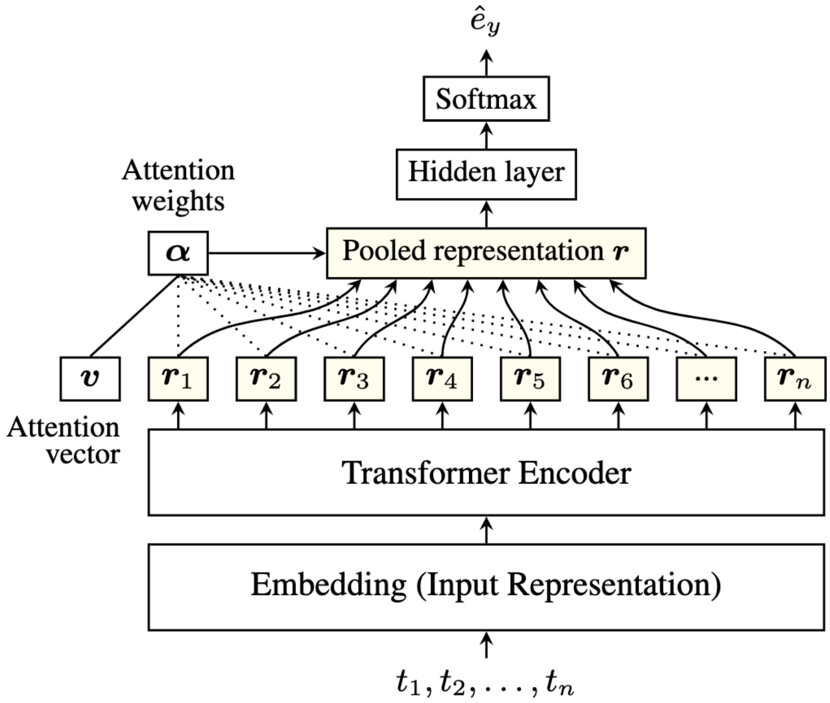

Figure 3.6: An illustration of the response emotion/intent predictor. Note that the dotted lines denote the computation for the attention weights between r_i and v. [15]

$$r = \sum_{i=1}^{N} \alpha_i r_i$$

3.3.2 Response Generation

Principally, the response generation model is based on the plain Transformer's encoder and decoder. As previously discussed, both encoder and decoder consist of multiple identical encoding and decoding layers. The encoder starts by processing the input sequence concatenated by multiple turns in a dialog. The output of the top encoder is then transformed into a set of attention vectors as a key matrix and a value matrix. These are to be used by each decoder in its encoder-decoder attention layer (See Figure 3.7) which helps the decoder focus on appropriate places in the input sequence. Afterward, the decoder stack passes the output vector into the final Linear layer followed by a softmax layer. The softmax layer turns the logit scores gained from the linear layer into probabilities. The highest probability is chosen as the word output for this time step.

Different from the original Transformer, our model incorporates the rationale mask produced by the *EPITOME*'s bi-encoder model. Expressly, we further manipulate the loss function as the

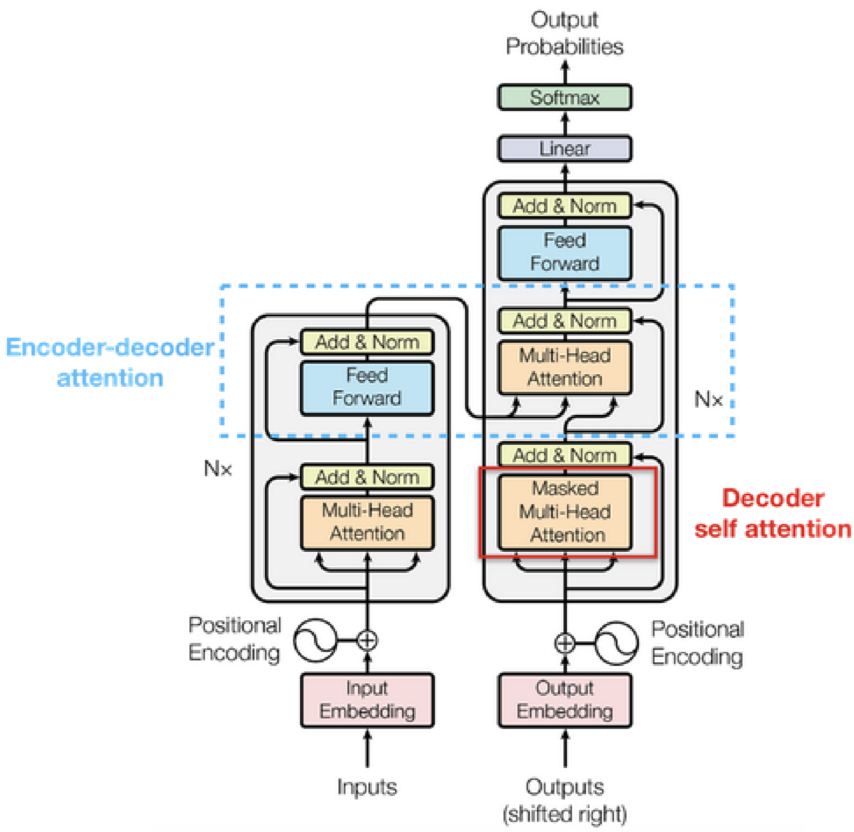

Figure 3.7: The encoder-decoder attention highlighted in the Transformer. Unlike multi-head attention, the encoder-decoder attention creates its queries matrix from the decoder self attention, and takes the keys and values matrix from the output of the encoder stack.

dot product between the cross entropy loss and the rationale mask. The reason doing so is to emphasize the tokens annotated with larger empathy levels, facilitating the model to have higher tendency to generate these words. After combining the rationale masks of the *EPITOME* communication mechanisms, we also observe that it makes no difference if we include the actual empathy levels to the masks or not. That is, only the boolean mask highlighting the rationale words matters. On the whole, with the help of rationale masks, our response generation model tends to frequently output the response containing higher emotion reaction, interpretation, and exploration.

3.3.3 Implementation

Separate Training

Following the procedure suggested in [15], we independently train our models. The whole training process starts with the *EPITOME* bi-encoder model, the response emotion/intent predictor, and ends with the response generation model. Specifically, after training the bi-encoder model, each response utterance is labeled with the predicted empathy levels and rationale masks. Then, we train the predictor by incorporating these levels. Finally, we train the Transformer by feeding both the predicted communication levels and emotion/intent to the embedding layers of the encoder and decoder. During the training, the model intends to minimize our new loss function considering the rationale masks. In this way, separate training endows the response generation model with more controllability since the model will refer to the indicated *EPITOME* communication levels and the predicted response emotion/intent.

Implementation Parameters

In general, our model implementation is fundamentally built upon this work [15]. Accordingly, we apply the RoBERTa tokenizer to tokenize the input utterances, with the vocabulary size set to be 50,265. Also, we restrict our model to have the input size within the 100 tokens. About the Transformer model, we use 4 layers in the encoder and decoder stack, with 6 attention heads used in their multi-head attention separately. The dimension of the hidden units is 300, and the linear layer accepts 1200 as its input dimension. Inside each hidden layer, we choose the GELU [41] activation function combining with the dropout rate to be 0.1. Lastly, we choose the Adam optimizer [42] with the initial learning rate as 5×10^{-5} to optimize the loss function during training.

For inference, we utilize the beam search algorithm with the beam size of 32. To prevent the models from generating repetitive words, we correct the algorithm so that at each step, if any branch contains repetitive 4-grams, the log probability of the branch is set to be infinitely negative, intending to stop it from being further expanded.

Chapter 4

Evaluation & Results

For benchmarking, we not only trained our dialog models but also the baselines using the RED dataset. The baselines include the EmoPrepend-1 and MEED2 originated from the two existing works [17, 15]. To evaluate them in the held-out setting [15], we leverage automatic metrics and crowdsourcing human evaluation.

As described in Table 3.3, we have approximately 560K conversations available in our corpus. Hence, we split 80% for training, 10% as the validation set, and the rest 10% as the testing set. To further accelerate our evaluation process, we sample 2K responses generated by each model given the same contexts. In addition, to ensure consistency, all the models have a hidden size of 300 and are trained until the minimum validation loss.

4.1 Automatic Evaluation

Automatic evaluation is carried out by calculating a score that indicates the similarity between machine generated texts and human written texts. However, human conversations can be expressed in many ways, so it is still critical to select one golden metric to estimate whether generated texts are consistent with the design goal of the system. For example, when assessing the open-domain chatbot, the earlier work [43] has shown that the metrics designed for machine translation or automated summarization have weak correlation with human judgements. Specifically, both the word overlap similarity metrics (*BLEU* [44], *METEOR* [45], and *ROUGE* [46]) and word embedding metrics derived from word embedding models (Word2Vec [47]) correlate poorly with human evaluation. By contrast, for goal-oriented dialog systems, Sharma et al. [48] indicates that *METEOR*, *ROUGE-L*, and some similarity metrics reveal positively stronger correlation with human assessment. As suggested by the recent study [49], it is desirable to use multiple metrics to gauge different aspects of dialog systems' capability. To this end, we adopt the metrics used in this paper [15] as well as those specified in this work [48].

	w/o EPITOME	w/ EPITOME
Accuracy	0.1254	**0.1541**
Precision	0.0865	**0.1407**
Recall	0.1254	**0.1541**
F1-Score	0.0887	**0.0891**

Table 4.1: Accuracy, weighted precision, recall and F-1 score of the response emotion/intent predictor used in our models. Note that the predictor without *EPITOME* is used in the baseline *MEED2* while the one with *EPITOME* is used in our newly created models *EPIMEED* and *EPIMEED+*.

	Emoprepend	MEED2	EPIMEED	EPIMEED+
Perplexity	2707.56	1455.69	**575.93**	1917.37
D1	0.0317	**0.0618**	0.0487	0.0039
D2	0.1178	**0.2889**	0.1912	0.0181
Sentence Embedding Similarity	**0.2683**	0.2115	0.2242	0.2315
METEOR	0.0434	0.0331	0.0365	**0.0637**
ROUGE-L	0.0662	0.069	**0.0746**	0.0559
CIDEr	0.107	0.1352	**0.1683**	0.0112
SkipThought Cosine Similarity	0.4842	0.4874	**0.4911**	0.4268
Embedding Average Cosine Similarity	0.7346	0.7408	0.7285	**0.765**
Vector Extrema Cosine Similarity	0.7346	0.7408	0.7285	**0.765**
Greedy Matching Score	0.3507	0.4113	**0.4194**	0.3681

Table 4.2: Automatic evaluation metrics. Here D1 and D2 are denoted as Distinct-1 and Distinct-2 respectively.

The results of response emotion/intent predictor performance and automatic evaluation are presented in Table 4.1 and 4.2 separately. In general, our new models, *EPIMEED* and *EPIMEED+*, outperform the baselines originated from the previous studies. Even if either of them doesn't achieve the best, they are slightly worse by the small gap. However, the outstanding scores here don't fully reflect if our models are capable of generating empathetic responses. To evaluate them more accurately, we must compare these 4 models via the perspective of human evaluators.

4.2 Human Evaluation

To a dialog system, the ultimate goal is always generating texts valuable to people. For this reason, human evaluation is typically treated as the most important measure. Besides, it is also held as the gold standard when developing new automatic metrics. With the current automatic metrics still falling short of replicating human decisions, many studies [50] tend to include some form of human evaluation to assess chatbots. Thus, human evaluation has become the mainstream in that it gives us the best insight into how well a model performs in a task.

To conduct human evaluation, we carefully follow the framework developed by Welivita and Pu [16]. Firstly, we randomly select 200 dialogues from our RED testing set as the context for our chatbots to reply. Once they make responses, we recruit workers in Amazon Mechanical Turks (AMT) to gauge the empathetic fitness of the generated responses by our 4 models. During the experiment, the AMT workers are taught to drag and drop their judgements into three levels: *Good*, *Okay*, and *Bad*, based on how empathetic a generated response is to the given context. With each set assessed by three workers, the final rating of predicted responses is computed via the majority vote. More details about the experiment setup can be found in the appendices of Welivita and Pu's work [16].

According to Table 4.3, the outcome indicates that *MEED2* still achieves the best with almost 82% of the generated responses viewed as good replies. On the other hand, *EPIMEED+* has the least good rating because only 43% predicted responses are regarded to be good. Nevertheless, more than half of the workers do not agree to the rating as indicated by the low agreement percentage.

To understand the cause, we further investigate our testing conversations evaluated by the workers. In particular, we compare the good and the bad generated responses to clarify why the judgements on *EPITOME+* vary drastically among the workers. Table 4.4 and 4.5 separately list several examples that *EPITOME+* performs well and badly. As we observed all the bad samples, the main drawback is that the response *EPIMEED+* makes does not clearly follow the context. That is to say, *EPIMEED+* poorly answers right in accordance with the speaker utterances, even though it generates the ostensibly empathetic replies. What's more, some predicted samples of *EPIMEED+* incline to making the human raters confused about what rating to give. In Table 4.6, *EPIMEED+* may possibly produce ambiguous responses so that human evaluators have difficulty to determine the empathetic response quality. Of our shown examples, the responses

	Emoprepend	MEED2	EPIMEED	EPIMEED+
Good	51.69%	81.67%	73.88%	43.18%
Okay	15.73%	18.33%	26.92%	22.73%
Bad	32.58%	0%	0%	34.88%
Agreed	89	120	104	88
Total	200	200	200	200
Agreed Percentage	44.5%	60%	52%	44%

Table 4.3: Human evaluation results for empathetic appropriateness of the responses produced by our four models.

from *EPIMEED+* are simply generic questions or statements without rigorously taking speaker utterances into account. As a result, our workers may feel undetermined to make their decisions, leading to large variance on the *EPIMEED+* response ratings. More instances of the good, bad and ambiguous generated responses are included in the appendix for readers' reference.

Overall, our human evaluation experiment points out that *EPIMEED* and *EPIMEED+* don't fully outperform the baselines. For further investigation, we find that the main limitation results from the fact that they sometimes unsuccessfully follow the context to make responses, although the resulting responses seem to be empathetically appropriate in words. As a consequence, the human raters hesitate to confirm good ratings on the generated responses of *EPIMEED* and *EPIMEED+*.

Context:
Just realized school is coming up Jan14 I still have no friends coming in my third semester at a small university as a junior F**K MY LIFE! But ima enjoy 2019 being alone.
Emoprepend (wishing): Congrats!
MEED2 (wishing): Good luck!
EPIMEED (wishing): Congratulations!
EPIMEED+ (wishing): I'm in the same boat as you. I wish you all the best in the world.
Ground-truth (lonely): We can both enjoy 2019 being alone.

Context:
I recently started writing some short stories for the nosleep subreddit and have really enjoyed it. my stories have gotten hundreds of upvotes and lots of positive attention, but i am appalled each time i post a story because the commenters can be so rude. some commenters are so supportive and tell me that they enjoy my stories, but others call my writing lazy and insignificant. i welcome constructive criticism!
Emoprepend (grateful): Thank you.
MEED2 (questioning): Do you want to talk about it?
EPIMEED (acknowledging): Your submission was automatically removed because your post had no body. Title-only posts don't give users enough information or context to provide support.
EPIMEED+ (acknowledging): Thank you for sharing your story, I'm happy to hear it. I'm glad you're feeling better now. I hope it gets better for you. I'm here if you want to talk about it. Let me know what's going on. What's going on?
Ground-truth (furious): It blows my mind how some people do out of their way to be an asshole. Have an internet hug and don't stop writing, especially if you enjoy it.

Context:
"I love you" is the most sacred thing you can ever tell someone and like everybody, I imagined falling in love and saying these words to each other. I guess after hearing it from him, I kinda fell in love with the idea of being in love and the next thing I know... I was in love. Do people fall out of love like that? I hope I get over this one day.
Emoprepend (sympathizing): I'm so sorry you're going through this.
MEED2 (acknowledging): "I love you" and I hope you get over this one day.
EPIMEED (agreeing): I feel the exact same way.
EPIMEED+ (agreeing): I know exactly how you feel. I'm going through the same thing right now. It's so hard. You're not alone. I know what you're going through. If you need someone to talk to, feel free to message me. PM me, I'll be here for you. PM me if you'd like. PM me. I know how you'll pm me if you need to pm me.
Ground-truth (lonely): They do. My ex just blocked from everywhere one morning without saying anything, until she returned after 3 months seeking sympathy.

Table 4.4: Some samples of the predicted responses showing *EPIMEED+* performs well.

Context: I have terrible stress, anxiety and depression. I constantly wish I would just die. I have a therapy appointment on Tuesday and I want to cancel because I can't even talk about it. I have so much stress and anxiety that if I talk about it I think I'll just explode. Anyone else feel this way? **Emoprepend (sympathizing)**: I'm sorry you're feeling this way. **MEED2 (agreeing)**: Yes, I feel this way too. **EPIMEED (agreeing)**: I'm right there with you. **EPIMEED+ (agreeing)**: I'm in the same boat as you. I don't know what to do about it, but it's not worth it. **Ground-truth (confident)**: That's what they're there for. Let yourself explode. Tell them exactly what you just told us. You got this! You'll feel better after the appointment.
Context: I'm 21 years old and I cried because of a mistake I made. I accidentally mixed an acid based cleaner and bleach based cleaner. I was still anxious after hearing that. After the poison control call I was still upset, she tried to tell me people make mistakes but I responded I'm upset I made a stupid mistake, and she just groaned and didn't try to comfort me in any way. **Emoprepend (sympathizing)**: I'm so sorry you had to go through that. I hope things get better for you soon. **MEED2 (questioning)**: Why did she try to comfort you in any way? **EPIMEED (questioning)**: Do you want to talk about it? **EPIMEED+ (questioning)**: Please don't do it. Please don't hurt yourself. What is making you feel this way? Have you told anyone any of the stuff you're going through or looked into getting help? **Ground-truth (annoyed)**: You're being a crybaby because you almost died. Seriously, don't worry about being shook up.
Context: I turn 19 today. I'm not here to brag, but I think today counts as something special. Today, I woke up and was happy. Today ain't very special, but to me, this might be a sign of chance, a glowing light of happiness that I can reach and finally feel what it is. **Emoprepend (wishing)**: Happy birthday! **MEED2 (wishing)**: Happy birthday! **EPIMEED (wishing)**: Happy birthday! **EPIMEED+ (wishing)**: Great job! Keep up the good work! Keep your chin up! Keep it up! **Ground-truth (wishing)**: Happy birthday!

Table 4.5: Some samples of the predicted responses showing *EPIMEED+* performs poorly.

Context: I'm a worthless pathetic loser. Someone who's in a far worse off country deserves my life. **Emoprepend (questioning)**: What country do you live in? **MEED2 (questioning)**: What country do you live in? **EPIMEED (questioning)**: Do you want to talk about it? **EPIMEED+ (questioning)**: What's going on in your life right now? What is making you feel this way? **Ground-truth (sad)**: Just because there is someone worse off in another country, doesn't mean that they feel worse. You cannot measure sadness, or compare it - your pain is your own and it is completely justified, no matter what the reason is.
Context: I would spend hours afraid I would kill myself. It's now 3 years later. Every day is hard. Every day is a struggle but I'm trying. Constant headaches burning and nothing helped. **Emoprepend (sympathizing)**: I'm so sorry you're going through this. I hope things get better for you soon. **MEED2 (wishing)**: I wish you the best of luck. **EPIMEED (questioning)**: What kind of stuff is making you feel this way? You talked to anyone in your life about any of the stuff you're going through or looked into getting help? Made any changes in your life lately? I hope things get better for you. **EPIMEED+ (questioning)**: What's making you feel this way? Have you looked into getting help? Made any changes in your life? **Ground-truth (hopeful)**: As someone whose been really down because of Acid Reflux, it really made me feel better that your magically went away. I hope that happens to me too. I hope things are going better! Life lasts a long time, and the present feels longer when you have to deal with it. But you never know what good turn is around the bend!
Context: I feel distanced. I just don't feel like doing it. I don't feel like doing anything. I just feel like sleeping or something. Depressed people do. **Emoprepend (lonely)**: I just don't feel like doing it. I don't feel like doing anything. I just feel like sleeping or something. **MEED2 (wishing)**: What kind of stuff is making you feel this way? **EPIMEED (questioning)**: I'm in the same boat. You're not alone. **EPIMEED+ (agreeing)**: I know exactly how you feel. I'm in the same boat. It's the only thing I've ever felt. I don't know what you're going through, but it's definitely worth it. You're not alone. I hope you feel better. **Ground-truth (agreeing)**: Yeah, sometimes exercise can help. Sometimes exercise works great, sometimes it doesn't, and sometimes you need a whole combo of that and diet and therapy and meds. Little did I know that having whacked brain chemicals render things physically impossible for me to do sometimes. Keep holding on for the little things. I know everything feels pretty hopeless right now, and you know what?

Table 4.6: Some samples of the predicted responses showing *EPIMEED+* makes vague responses difficult for human raters to confirm ratings.

Chapter 5

Related Work

Dialog systems are always an interesting topic to both machine learning researchers and practitioners. The very first attempt to build such systems can be dated back to the 1960s [20]. Since then, dialogue systems are either designed to perform specific tasks such as flight booking [51], healthcare [52], political debate [53], hence termed task-oriented dialogue systems, or to chitchat, so called chatbots. A task-oriented dialogue system often consists of multiple modules including language understanding, dialogue state tracking, dialogue policy, and dialogue generation. On the other hand, chatbots learn to generate dialogs from offline data without any state, action, or intent involved [54]. The recent progress in deep learning [55] also facilitates the use of end-to-end solutions to dialogue systems which can be more easily trained to simulate the behavior of human communication via an access to a large amount of training data. The process of generating responses conditioned on the existing contexts of a dialogue can be naturally modeled as a translation process where end-to-end solutions such as the sequence-to-sequence (Seq2Seq) model [4] have already been proven effective. In particular, what makes the Seq2Seq model shine is the attention mechanism, revolutionizing the field especially in language translation and generation. Ever since, more and more studies focus on incorporating the attention mechanism to their dialog models. For instance, based on the hierarchical encoder-decoder structure from Serban et al. [56], Xing et al. [57] devised a hierarchical attention mechanism so that the model could pay attention at both token level and utterance level when generating texts.

With the advent of advanced language models, it is possible to learn the nuanced emotion exchanges existing in natural language conversations. Specifically, to make dialog systems more human-like, endowing chatbots with empathy is indispensable. Empathy is the capability of projecting feelings and ideas of the other party to someone's knowledge [58]. It plays an important part in the communication of human beings as it has the potential for enhancing their emotional bond. As stated in the previous study [59], incorporating empathy into a dialogue system is vital for improving user experience in human-computer interaction. More importantly, being empathetic is a necessary step for the dialogue agent to be perceived as a social character by users [60]. On the whole, building an empathetic dialogue system is then premised on the idea that it will result in improved user engagement and more effective communication.

To capture emotion information for guiding response generation, past studies focus on either pre-defined emotion labels or hand-crafted rules. For example, some existing works [61, 62, 63, 64, 65, 66] propose their neural based models by requiring a manually defined emotion label as an input to lead the following response. Usually, the number of defined labels is limited, which may potentially neglect the subtle emotional change. Other studies [67, 68, 69, 13, 70] put forth a computing strategy to decide the response emotion state. The strategy may be determined through reversing the speakers' emotion or maximizing the emotion content based on the chatting history. Despite helping achieve good performances, such deterministic rules still lack the confirmation of psychology literature.

Apart from the common approaches, there are some intriguing works applying extraordinary methods to grasp emotions. For instance, Lubis et al. [71] designed a hierarchical encoder-decoder architecture by including an emotion encoder. For each utterance, context emotion state is obtained through the encoded representation of an input dialog sequence. Then, the generation of the next response is conditioned on the concatenation of the dialog and the emotion contexts. Shin et al. [72] applied the reinforcement learning method whose reward function provides a higher reward to the generative model if the generated utterance improves the user's sentiment. Li et al. [73] adopted an adversarial learning approach by proposing two discriminators to evaluate if a generated response is empathetic and elicits more positive emotions via considering the emotion words in the gold response and the subsequent reply.

Referring to real-world application, XiaoIce [74] would be the most well-known example. It is designed as a social chatbot with an emotional connection to satisfy the human need for communication and affection. Unlike traditional bots built to mimic human conversations but to interact with users' environment, XiaoIce has to interact with the user's environment and access real-world knowledge since it integrates IQ and EQ skills required to help address specific tasks. As a result, it applies the modular structure similar to task-oriented dialogue systems, with different modules dealing with different tasks. Depending on the availability of training data and knowledge for each task, either a rule-based method or a data-driven method, or a hybrid of both is adopted for the task. Since the release in 2014, XiaoIce has been a world-class application and has communicated with over 660 million users with succeeding in establishing long-term relationship.

Most previous studies aim at finding and incorporating an appropriate emotion to a dialog model to make responses conditioned on it. However, empathy is a multi-dimensional idea related not only emotion but cognition. [75]. The emotion aspect is about the emotional stimulation to the feeling expressed by a user. By contrast, the cognition aspect more intentionally involves interpreting the experiences of a user and communicating that understanding to them [76]. In this work, we target at fulfilling the gap by augmenting the computational expressed empathy framework [14] to the work of Xie et al. [15]. In the end, our dialog models are able to addressing both emotional and congnitive aspects of empathy.

Chapter 6

Conclusion

In this work, we developed *EPIMEED* and *EPIMEED*+, brand new empathetic dialog models able to learn not only emotion exchanges but cognitive understanding of expressed empathy within conversations at the more fine-grain level. To achieve this, we take advantage of the state-of-the-art computational framework *EPITOME* to augment our models and train them on our empathetic dialog corpus, RED. Manual inspection of the generated responses unveils that our models are capable of effectively communicating to speakers with empathy. Additionally, we further benchmark the utility of *EPIMEED* and *EPIMEED*+ using the baselines from the works of Xie et al. [15] and Rashkin et al. [17]. According to the automatic evaluation, our models indeed better the baselines in numerous metrics. However, the outcome in the human evaluation turns out to be slightly opposite. In our finding, we speculate that our models occasionally generate responses without attending to the context of conversations carefully. Moreover, they may likely respond with generic statements that are hard for human raters to make decisions. Lastly, we also observe that short utterances are preferable among human raters in our experiment. As a result, our generative models still suffer from fully grasping the contexts to accordingly make responses, even though they have higher tendency to express themselves based on empathy.

As for future work, it is definitely needed to improve the accuracy of the response emotion/intent predictor, as we see that the predicted emotion/intent may deeply affect the subsequently generated responses. We can also increase the number of tokens for the input sequence to our models, since the current models only accept the dialogs within 100 tokens in total. In this way, we will include more diverse conversations for our models to learn and enhance its capability. Lastly, it is worth trying out a more large-scale human evaluation. Specifically, our present experiment merely assigns 200 dialog samples for three AMT workers to judge. As a result, the total number of human ratings is at most 600. If we have more generated responses for more human raters to make assessment, the model performance could be more accurate because of the less biased evaluation result.

Bibliography

[1] Walsh M. *Mental health statistics 2021.* URL: https://www.singlecare.com/blog/news/mental-health-statistics.

[2] SingleCare Team. *Mental health survey 2020.* URL: https://www.singlecare.com/blog/news/mental-health-survey.

[3] J.H. De Kock, H.A. Latham, and S.J. et al. Leslie. *A rapid review of the impact of COVID-19 on the mental health of healthcare workers: implications for supporting psychological well-being.* 2021. DOI: https://doi.org/10.1186/s12889-020-10070-3.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. "Sequence to Sequence Learning with Neural Networks". In: 2014. arXiv: 1409.3215 [cs.CL].

[5] C. Busso, M. Bulut, and CC. et al. Lee. *IEMOCAP: interactive emotional dyadic motion capture database.* 2008. DOI: https://doi.org/10.1007/s10579-008-9076-6.

[6] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. *DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset.* 2017. arXiv: 1710.03957 [cs.CL].

[7] Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Ting-Hao, Huang, and Lun-Wei Ku. *EmotionLines: An Emotion Corpus of Multi-Party Conversations.* 2018. arXiv: 1802.08379 [cs.CL].

[8] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. "SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text". In: *Proceedings of the 13th International Workshop on Semantic Evaluation.* Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019, pp. 39–48. DOI: 10.18653/v1/S19-2005. URL: https://aclanthology.org/S19-2005.

[9] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations.* 2019. arXiv: 1810.02508 [cs.CL].

[10] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. "Affective Neural Response Generation". In: *Advances in Information Retrieval.* Ed. by Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury. Cham: Springer International Publishing, 2018, pp. 154–166. ISBN: 978-3-319-76941-7.

[11] Peixiang Zhong, Di Wang, and Chunyan Miao. "An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 7492–7500. DOI: 10.1609/aaai.v33i01.33017492. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4740.

[12] Qintong Li, Piji Li, Zhumin Chen, and Zhaochun Ren. *Towards Empathetic Dialogue Generation over Multi-type Knowledge*. 2020. arXiv: 2009.09708 [cs.CL].

[13] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. "MoEL: Mixture of Empathetic Listeners". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 121–132. DOI: 10.18653/v1/D19-1012. URL: https://aclanthology.org/D19-1012.

[14] Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. *A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support*. 2020. arXiv: 2009.08441 [cs.CL].

[15] Yubo Xie and Pearl Pu. "Empathetic Dialog Generation with Fine-Grained Intents". In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, Nov. 2021, pp. 133–147. URL: https://aclanthology.org/2021.conll-1.10.

[16] Anuradha Welitivita and Pearl Pu. *HEAL: A Knowledge Graph for Distress Management Conversations*. 2022.

[17] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. *Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset*. 2019. arXiv: 1811.00207 [cs.CL].

[18] Bryl A. Blanzieri E. "A Survey of Learning-based Techniques of Email Spam Filtering". In: *Artif Intell Rev 29*. 2008.

[19] Ani Nenkova and Kathleen McKeown. "A Survey of Text Summarization Techniques". In: Jan. 2012.

[20] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine". In: *Commun. ACM* 9 (1966), pp. 36–45.

[21] Kenneth Mark Colby. *Artificial Paranoia: A Computer Simulation of Paranoid Processes*. USA: Elsevier Science Inc., 1975. ISBN: 0080181627.

[22] Amazon.com. *Alexa Prize*. URL: https://developer.amazon.com/alexaprize.

[23] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-term Memory". In: vol. 9. 1997, pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[24] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*. 2014. arXiv: 1412.3555 [cs.NE].

[25] John F. Kolen and Stefan C. Kremer. “Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies”. In: 2001, pp. 237–243. DOI: 10.1109/9780470544037.ch14.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need.* 2017. arXiv: 1706.03762 [cs.CL].

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition.* 2015. arXiv: 1512.03385 [cs.CV].

[28] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. *Layer Normalization.* 2016. arXiv: 1607.06450 [stat.ML].

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* 2019. arXiv: 1810.04805 [cs.CL].

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* 2019. arXiv: 1907.11692 [cs.CL].

[31] Rico Sennrich, Barry Haddow, and Alexandra Birch. *Neural Machine Translation of Rare Words with Subword Units.* 2016. arXiv: 1508.07909 [cs.CL].

[32] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. “Language Models are Unsupervised Multitask Learners”. In: (2019).

[33] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP.* Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://aclanthology.org/W18-5446.

[34] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. *XLNet: Generalized Autoregressive Pretraining for Language Understanding.* 2020. arXiv: 1906.08237 [cs.CL].

[35] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. “RACE: Large-scale ReAding Comprehension Dataset From Examinations”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 785–794. DOI: 10.18653/v1/D17-1082. URL: https://aclanthology.org/D17-1082.

[36] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. *SQuAD: 100,000+ Questions for Machine Comprehension of Text.* 2016. arXiv: 1606.05250 [cs.CL].

[37] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. *The Pushshift Reddit Dataset.* 2020. arXiv: 2001.08435 [cs.SI].

[38] Anuradha Welivita, Yubo Xie, and Pearl Pu. *Fine-grained Emotion and Intent Learning in Movie Dialogues*. 2020. arXiv: 2012.13624 [cs.CL].

[39] Elliott R, Bohart AC, Watson JC, and Greenberg LS. "Empathy". In: *Empathy. Psychotherapy (Chic)* 48 (Mar. 2011). DOI: 10.1037/a0022187.

[40] Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. *Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping*. 2020. arXiv: 2002.06305 [cs.CL].

[41] Dan Hendrycks and Kevin Gimpel. *Gaussian Error Linear Units (GELUs)*. 2020. arXiv: 1606.08415 [cs.LG].

[42] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[43] Chia-Wei Liu, Ryan Lowe, Iulian V. Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. *How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation*. 2017. arXiv: 1603.08023 [cs.CL].

[44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040.

[45] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72. URL: https://aclanthology.org/W05-0909.

[46] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[47] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL].

[48] Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. *Relevance of Unsupervised Metrics in Task-Oriented Dialogue for Evaluating Natural Language Generation*. 2017. arXiv: 1706.09799 [cs.CL].

[49] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. *Evaluation of Text Generation: A Survey*. 2021. arXiv: 2006.14799 [cs.CL].

[50] Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. "Unifying Human and Statistical Evaluation for Natural Language Generation". 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, June 2019, pp. 1689–1701. DOI: 10.18653/v1/N19-1169. URL: https://aclanthology.org/N19-1169.

[51] Jean-Yves Magadur, Frederic Gavignet, Francois Andry, and Francis Charpentier. "A French oral dialogue system for flight reservations over the telephone". In: *Proc. 3rd European Conference on Speech Communication and Technology (Eurospeech 1993)*. 1993, pp. 1789–1792.

[52] Fabrizio Morbini, Eric Forbell, David DeVault, Kenji Sagae, David Traum, and Albert Rizzo. "A Mixed-Initiative Conversational Dialogue System for Healthcare". In: *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Seoul, South Korea: Association for Computational Linguistics, July 2012, pp. 137–139. URL: `https://aclanthology.org/W12-1620`.

[53] Aparup Khatua, Erik Cambria, Apalak Khatua, and Iti Chaturvedi. "Let's Chat about Brexit! A Politically-Sensitive Dialog System Based on Twitter Data". In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2017, pp. 393–398. DOI: `10.1109/ICDMW.2017.57`.

[54] Oriol Vinyals and Quoc Le. *A Neural Conversational Model*. 2015. arXiv: `1506.05869 [cs.CL]`.

[55] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. "Deep Learning–based Text Classification: A Comprehensive Review". In: *ACM Computing Surveys (CSUR)* 54.3 (2021), pp. 1–40.

[56] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. *Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models*. 2016. arXiv: `1507.04808 [cs.CL]`.

[57] Chen Xing, Wei Wu, Yu Wu, Ming Zhou, Yalou Huang, and Wei-Ying Ma. *Hierarchical Recurrent Attention Network for Response Generation*. 2017. arXiv: `1701.07149 [cs.CL]`.

[58] Raymond Nickerson, Susan Butler, M. Carlin, J. Decety, and W. Ickes. "Empathy and knowledge projection". In: *The Social Neuroscience of Empathy* (Jan. 2011), pp. 43–56.

[59] Karen K. Liu. "Embedded empathy in continuous, interactive health assessment". In: *CHI Workshop on HCI Challenges in Health Assessment*. 2005.

[60] Michael McTear, Zoraida Callejas, and David Griol. *The Conversational Interface: Talking to Smart Devices*. 1st. Springer Publishing Company, Incorporated, 2016. ISBN: 3319329650.

[61] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. *Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory*. 2018. arXiv: `1704.01074 [cs.CL]`.

[62] Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. "Automatic Dialogue Generation with Expressed Emotions". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 49–54. DOI: `10.18653/v1/N18-2008`. URL: `https://aclanthology.org/N18-2008`.

[63] Xianda Zhou and William Yang Wang. *MojiTalk: Generating Emotional Responses at Scale*. 2018. arXiv: `1711.04090 [cs.CL]`.

[64] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. *Affect-Driven Dialog Generation*. 2019. arXiv: 1904.02793 [cs.CL].

[65] Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. "Generating Responses with a Specific Emotion in Dialog". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3685–3695. DOI: 10.18653/v1/P19-1359. URL: https://aclanthology.org/P19-1359.

[66] Lei Shen and Yang Feng. "CDL: Curriculum Dual Learning for Emotion-Controllable Response Generation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 556–566. DOI: 10.18653/v1/2020.acl-main.52. URL: https://aclanthology.org/2020.acl-main.52.

[67] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. *Affective Neural Response Generation*. 2017. eprint: 1709.03968 (cs.CL).

[68] Jingyuan Li and Xiao Sun. "A Syntactically Constrained Bidirectional-Asynchronous Approach for Emotional Conversation Generation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 678–683. DOI: 10.18653/v1/D18-1071. URL: https://aclanthology.org/D18-1071.

[69] Peixiang Zhong, Di Wang, and Chunyan Miao. "An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss". In: *AAAI*. 2019, pp. 7492–7500. URL: https://doi.org/10.1609/aaai.v33i01.33017492.

[70] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. *Knowledge Bridging for Empathetic Dialogue Generation*. 2021. arXiv: 2009.09708 [cs.CL].

[71] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. "Eliciting Positive Emotion through Affect-Sensitive Dialogue Response Generation: A Neural Network Approach". In: *AAAI*. 2018.

[72] Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. *Generating Empathetic Responses by Looking Ahead the User's Sentiment*. 2021. arXiv: 1906.08487 [cs.CL].

[73] Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. *EmpDG: Multiresolution Interactive Empathetic Dialogue Generation*. 2020. arXiv: 1911.08698 [cs.CL].

[74] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. *The Design and Implementation of XiaoIce, an Empathetic Social Chatbot*. 2019. arXiv: 1812.08989 [cs.HC].

[75] Mark Davis. "A Multidimensional Approach to Individual Differences in Empathy". In: *JSAS Catalog Sel. Doc. Psychol.* 10 (Jan. 1980).

[76] Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and David Murphy. "Therapist empathy and client outcome: an updated meta-analysis". English. In: *Psychotherapy* 55.4 (Dec. 2018), pp. 399–410. ISSN: 0033-3204. DOI: 10.1037/pst0000175.

Appendix A

Optional Reading

A.1 The Cleaning Procedure for the RED Dialogs

As the behavior of our trained models highly relies on the data we use, we must do some pre-processing to remove the unnecessary parts. To attain this, we further clean the dataset with the following steps:

- Remove the duplicate turns under the the same conversation IDs.
- Remove the irrelevant tags such as HTML, URL, or any other meaningless marks appeared in browsers.
- Remove usernames and subreddit names.
- Remove the texts inside any squared bracket.
- Remove the bots' generated text starting with "I'm a bot...".
- Remove utterances in a dialog with the total length greater than 100 tokens.
- Remove the toxic words or toxic sentences of listener turns if the detecting probability is greater than 0.95.

Lastly, whenever we remove a dialog turn, we also discard all the following turns in the same dialog.

A.2 More Sample Responses

Table A.1 and A.2 lists some more examples of the generated responses performing well and poorly separately indicated in human evaluation. Besides, Table A.3 points out the solid examples where *EPIMEED+* may potentially generate generic responses making human raters indecisive to verify their quality.

Context:
Some background: My cat Archie likes to play with those hard colored gum balls almost exclusively. It's helped us grow considerably. It's given us tools to work through bad habits taught to us by our parents and helped us avoid the toxic pitfalls that we might have made through our insecurities. To my husband, I love you.
Emoprepend (sympathizing): I'm so sorry for your loss.
MEED2 (acknowledging): That's a good idea.
EPIMEED (acknowledging): That's awesome! I'm glad you're still with us!
EPIMEED+ (acknowledging): I'm glad you're still with us. You're not alone. We're here if you want to talk, we're all here for you. Keep it up, it's worth it.
Ground-truth (trusting): It's all about perspective. I tell myself this all the time. Life is hard but it sure is better to go through it with someone by your side, faults and all.

Context:
School is almost over. Things still haven't fixed with my best friend. I want to talk to her about it but I fear it'll just make my situation worse than it already is. We talk here and there.. but it was nothing compared to how our relationship was before. I have only Monday to make it right, then it's Finals and summer; no chance after Monday.
Emoprepend (sympathizing): I'm so sorry you're going through this. I hope things get better for you soon.
MEED2 (questioning): How long have you been together?
EPIMEED (questioning): Do you want to talk about it?
EPIMEED+ (questioning): Wanna talk about it? I'm here if you want to talk about it. I'm here to listen if you need someone to talk to. I don't know what you're going through, but I'm glad you're still here. You're not alone.
Ground-truth (hopeful): Think about it this way, what chances do you have after Monday? None like you said. Don't be afraid, just say what you need to say.

Context:
My cat died today, and I can't get it out of my mind. I keep think of him running on a sunny hill but that only makes me cry more. I got to hold him as he was being put down. I felt the life drain from his body. RIP Banx
Emoprepend (sympathizing): I'm so sorry for your loss.
MEED2 (sympathizing): I'm so sorry for your loss.
EPIMEED (sympathizing): I'm so sorry for your loss.
EPIMEED+ (sympathizing): I'm so sorry for what you're going through. I've been feeling this way for a while. I hope you'll be able to reach out to someone who's going through something similar to you's making you feel this way? Have you looked into getting help?
Ground-truth (faithful): When a pet or a love one dies I look up to sky and find a orange star... Why? Well, because I like to believe that the new Star is their soul and that it's watching me from above.

Table A.1: More samples of the generated responses showing *EPIMEED+* performs well.

Context:
- I look forward to this time, everyday. First I message her to wish her good morning and then I slowly touch myself. I keep touching myself until I ejaculate or until she comes back from the bathroom. She has no idea that I'm only staying to give my son a family. She cooks well though.
- So overwhelming, no support from anyone. I can't fucking take it anymore.
- Wuht?

Emoprepend (caring): I'm here for you if you need someone to talk to.

MEED2 (questioning): If you don't mind me asking, why do you think you can't take it anymore?

EPIMEED (questioning): Don't do it.

EPIMEED+ (questioning): How? How?

Ground-truth (suggesting): Try vitamins B6 and B12. It will help you with your problem.

Context:
i see everyone else with nice things. I'm 34yrs old with no assets, income or potential career path. my life is over.

Emoprepend (sympathizing): I'm sorry you're going through this. I hope things get better for you soon.

MEED2 (questioning): I see everyone else with nice things. My life is over. my life is over.

EPIMEED (agreeing): Me too.

EPIMEED+ (agreeing): I'm in the same boat as you. I've been in the same position as you're not alone in this way. It's like you's the only thing that's making you feel this way. I don't know what's going through. You're right, it's hard to talk to people's way too much you can't. I know what you

Ground-truth (jealous): I'm 32 and on the verge of falling into your situation. I live paycheck to paycheck, have a great salary, but most of it goes to credit cards. It's hard, watching everyone I know around my age in better places, financially speaking. It's bad enough comparing myself using other ridiculous measures of worth, but the money thing really messes with my head.

Context:
my old phone hadn't ever cracked, i got my new phone and within the first day shattered it. i need to be cheered up :(

Emoprepend (sympathizing): I'm so sorry you're going through this. I hope things get better for you soon.

MEED2 (questioning): What happened to your phone and within the first day of the first day ruined it?

EPIMEED (wishing): Congrats! I'm so happy for you!

EPIMEED+ (wishing): I'm glad you're still here

Ground-truth (questioning): Warranty?

Table A.2: More samples of the generated responses showing *EPIMEED+* performs poorly.

Context:
I was alone all day. I feel alone every day though. I saw people interact with each other all day. Sometimes I can go months without contacting old friends. I can't cry though.

Emoprepend (sympathizing): I'm so sorry you're going through this. I hope things get better for you soon.

MEED2 (questioning): Do you want to talk about it?

EPIMEED (agreeing): I'm in the same boat. You're not alone.

EPIMEED+ (agreeing): I know exactly how you feel. I'm in the same boat as you. You're not alone. I don't know what you're going through, but that's what I've done. It's not worth it, but it's worth it.

Ground-truth (suggesting): Pm me where you at. I'm on break too. If we're close we can go get cupcakes or chipotle while we ride the public transport :D

Context:
ugh and honestly maybe I'm not even bisexual maybe I'm just touch-starved and really lonely because I haven't had a relationship in a long time.

Emoprepend (questioning): Why haven't you had a relationship in a long time?

MEED2 (questioning): Why not just touch-starved and really lonely because I haven't had a relationship in a long time?

EPIMEED (questioning): Do you want to talk about it?

EPIMEED+ (questioning): I don't know what you're going through, but I'm glad you're still here. You're not alone, but you're not alone. What's going to happen?

Ground-truth (grateful): Hello fellow bisexual :) I've found that it's a difficult landscape to navigate because sometimes I'm not g̈ay enougḧfor the LGBTQA community (and have found a ton of backlash with lesbians) but not s̈traight enougḧfor straight people (and have received lots of biphobia with both men and women there). Finally found an LGBTQA choir that has been really accepting of me. Biphobia and bi-erasure is a very real thing.

Context:
For several years, I have been feeling mildly depressed at times and I kinda just kept to myself thinking I was having bad days. It got worse when I found out some bad news from my family over the Christmas break. It's harder to shake off this feeling. After some long talks with family members and my girlfriend, I have finally decided to go to a counselor. Wish me luck!

Emoprepend (wishing): Good luck!

MEED2 (wishing): Good luck!

EPIMEED (wishing): Good luck!

EPIMEED+ (wishing): I'm glad you're still here! Keep it up! Keep up the good work! Keep your chin up!

Ground-truth (encouraging): sending u all the good vibes i've got bud! I hope it goes well :)

Table A.3: More samples of the generated responses showing *EPIMEED+* makes ambiguous responses difficult for human raters to confirm ratings.