LaSF: Look a Step Further for Empathetic Dialogue Generation

Aiqi Cui, Zhen Lei, Ziyi Liu, Feifan Yan
{ac4788,zl3028,zl2888,fy2241}@columbia.edu

1 Introduction

Providing accessible mental health support is a challenging problem all over the world because in-person treatments are generally expensive, time-consuming and require expertise. Due to this reason, online peer-to-peer support platforms are becoming increasingly popular, and some of these platforms are also trying to build AI assistants that can assist or even replace human experts to make mental health care services more accessible to people. Studies have shown that empathetic interactions play an important role in symptom improvement in mental health support and are significant in building therapeutic alliance and rapport (Elliott et al., 2011). Nevertheless, highly empathetic personalized conversations are still rare on online support platforms.

Recently, some dialogue systems see empathy as a significant factor both to mental health care field as well as to the quality of response in general chatbots (Huang et al., 2019). Zhou et al. (2017) proposed a method to address the emotion factor in large-scale conversation generation. Lin et al. (2019) proposed that being empathetic not only requires the ability of generating emotional responses, but more importantly, requires the understanding of user emotions and replying appropriately and accordingly. Majumder et al. (2020) argued that empathetic responses often mimic the emotion of the user to a varying degree, depending on its positivity or negativity and content. Sharma (2020) proposed a new rewrite task to improve empathy in responses.

In this paper, we propose an intuitive method to improve empathy in our generated responses, which is to **look a step further (LaSF)**. Intuitively, we know that one is empathetic if he can put himself in someone else's shoes, but how can an agent put herself in the user's shoes? We assume that if an agent can know the reaction of the user, in other words, she can predict the user's response, then she may use the user's response to check whether her current response is appropriate or not.

The major components of this project include: (1) Researching and finding a pre-trained dialog model as the base model suitable for the mental health support task and find useful datasets. We experimented with DialogGPT and Blender as the base models. Details about the dataset will be discussed in later section. (2) Conducting language modeling using an empathy dataset to fine-tune the base model to create our baseline model. (3) Using an empathy ranker/classifier model to evaluate baseline model responses. The empathy score given by the ranker/classifier will be the baseline score. (4) Refining the baseline model using dialog planning by leveraging reinforcement learning techniques. Specifically, we incorporated the sentiment score of the user response, instead of the whole sentence, as reward for the agent to find a better policy to generate more empathetic expressions.(5) Evaluating the LaSF method performance and compared it against the performance of other models.

2 Related Work

2.1 Natural Language Processing for Mental Health

With the rapid development of NLP technologies, there have been growing interests in the area of applying such technologies to mental health problems, aiming at providing better support for patients. Studies have shown that online chat-based mental health support can be as effective as face-to-face counseling sessions (Hoermann et al., 2017). The accessibility and convenience of online support have promoted the popularity of peer support platforms like TalkLife. At the same time, since people

on such platforms do not generally possess professional knowledge of psychological counseling, the outcomes of these sessions might not be ideal. To address this issue, researchers have also focused on identifying helpful conversation strategies. Althoff et al. (2016) conducted a large-scale computational discourse analysis of counseling conversations using various NLP techniques such as sequence-based conversation models and language model comparisons and discovered various strategies that can potentially lead to better outcomes. More recently, Sharma et al. (2020) studied the importance of empathetic conversations in mental health support through a multi-task RoBERTa-based bi-encoder model.

However, as Sharma et al. (2020) pointed out, users do not self-learn to express empathy over time, which shows a promising research direction of developing dialog systems with empathy for mental health support. Recent works incorporating this aspect have demonstrated positive results. Sharma et al. (2021) used deep reinforcement learning to transform generated responses to more empathetic ones. Roller et al. (2020) also showed that the model fine-tuned on an empathetic dataset generated more empathetic responses. We would like to build on their work by also considering user reactions to the generated responses, to output more empathetic and caring responses.

2.2 Empathetic Dialogue Generation

Recognizing partner's feelings and emotions in a conversation, a key communicative skill for humans, has long been a challenge for dialogue agents. It is common for daily conversations to be prompted by people sharing their personal experiences, and a response acknowledging any implied feelings would sound more satisfying. Rashkin et al. (2019) proposed EmpatheticDialogues, a dataset dedicated to facilitate training a dialogue agent to generate responses with empathy as well as a benchmark to gauge a model's ability to respond in an empathetic way.

However, merely generating emotional responses is not enough. What's more important is to correctly understand user emotions and reply accordingly. Mixture of Empathetic Listeners (MoEL) is capable of detecting the most apt emotion of a speaker and generating an appropriate response (Lin et al., 2019).

More recent study argues that rather than being treated uniformly, emotions of the user are often mimicked by empathetic responses to a varying degree (Majumder et al., 2020). By adopting techniques as emotion grouping, i.e. positive & negative groups, and emotion mimicry, the model MIME manages to achieve better performance than the SOTA model MoEL in both empathy and relevance, though there still remains improvements for fluency.

2.3 Dialog Planning

In human conversations, "put yourself in someone else's shoes" is a common communication strategy to achieve empathy. People predict how their counterparts may react to their opinion and use the predication to phrase the conversation in order to best reflect their thoughts. Inspired by this idea, AI planning, a long-standing subarea of artificial intelligence focusing on the construction of sequences of actions to achieve a goal, has been recently introduced to the field of dialog systems to simulate the response prediction logic (Russell and Norvig, 2002; Stent et al., 2004; Jiang et al., 2019; Walker et al., 2007).

By formulating the human conversation as a Markov decision process (MDP), this kind of problem is usually solved by reinforcement learning techniques (Zhang et al., 2020b; Hancock et al., 2019). In earlier works, the idea of planning has been used to improve dialog generation (Stent et al., 2004; Walker et al., 2007). As a domain-specific application, Yarats and Lewis (2018) utilized dialog planning for negotiation tasks, in which the agent, at each state, simulates the complete dialogue for several candidate responses and select the response with the highest reward. In a more recent work, Jiang et al. (2019) introduced a novel way to teach the agent to look ahead more efficiently by only foreseeing several turns and use the limited information for the policy model.

3 Methodology

In **look a step further (LaSF)**. we leveraged reinforcement learning (RL) to let the system learn to predict the user reaction based on the current dialog context, which include the last user utterance and the currently generated reply. Details of the policy model and reward function will be discussed in the following sections.

3.1 Single Turn Policy Gradient

We used an adapted version of policy gradient method in reinforcement learning to fine-tune our original generative dialog model. Our steps include: (1) getting an utterance from user. (2) our chatbot generate a response conditioned on the user utterance. (3) our reward function outputs a score of the generated response. (4) use the outputted score as reward to fine-tune our policy (generation model). Figure 2 shows the process. The algorithm is shown below.

Algorithm 1: REINFORCE

Input: policy network parameter θ, training dataset D, number of iterations T, empathy predictor P

while $t \leq T$ **do**
  sample a start sentence s from D;
  get generated sentence g and logits $\nabla\pi_\theta$ through $\theta(s)$;
  get reward r through $P(r)$;
  optimize $-r * \nabla\pi_\theta$;
  $t \leftarrow t + 1$;
end

- **State:** The state for our problem is the utterance from user.
- **Action:** The action for our problem is the whole utterance the policy generated, because reward cannot be computed for a single word.
- **Reward:** Reward is given by an outside model to evaluate the future influence of the generated utterance.
- **Policy:** Policy is a pre-trained chatbot, which is Blender-small.

3.2 Reward Function: Sentiment Prediction

Since BERT (Devlin et al., 2018) has shown state-of-the-art performance on many NLPs tasks, including question answering, named entity recognition, and sentiment analysis (Shin et al., 2019), we decided to build a sentiment predictor based on BERT.

To train the predictor, we first created a new dataset for sentiment prediction by reforming the EmpatheticDialogues(ED) dataset (Rashkin et al., 2019). We chose this dataset because we want the predictor learn sentiment as well as how people show sentiment in empathetic conversation settings compared to every-day conversations. Specifically, we separated the multi-turn dialogs in ED into (s_i, l_i) pairs, which represents the dialog context by aggregating the previous turn's utterances from the speaker and listener. Then, we extracted the reply s_{i+1} from the speaker and evaluated its sentiment with VADER, a lexicon and rule-based sentiment analysis tool (Hutto and Gilbert, 2014), to get a reply sentiment score.

As Figure 1 shows, with the ED sentiment dataset, we trained the predictor by fine-tuning BERT with dialog context and sentiment scores to predict sentiment scores between 1 and 0, where 1 indicates reply with positive sentiment, or satisfaction, and 0 indicates dissatisfaction. The predictor achieved around 75% accuracy.

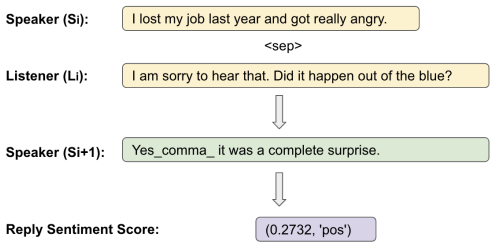

Figure 1: BERT-based Sentiment Prediction

3.3 Optimization

Our objective function is:

$$L = -r * \nabla\pi(\overline{a}|\theta) \quad (1)$$

where r is the score computed by the look ahead sentiment prediction, $\overline{a}$ is the average logits of output words, and $\nabla\pi(\overline{a}|\theta)$ is the gradient of output words given the parameters of pretrained-model, Blender-small.

4 Experiments

In this section, we provide details of the experiments we have performed, including the dataset, evaluation metrics, baseline settings, and results.

4.1 Dataset

Empathetic Dialogues Rashkin et al. (2019) constructed a dataset of 24850 one-to-one conversations between humans. Before each conversation, the speaker chose a emotion label and wrote about a situation in which they felt that way. During the conversation the speaker would describe it to the listener, and the listener would

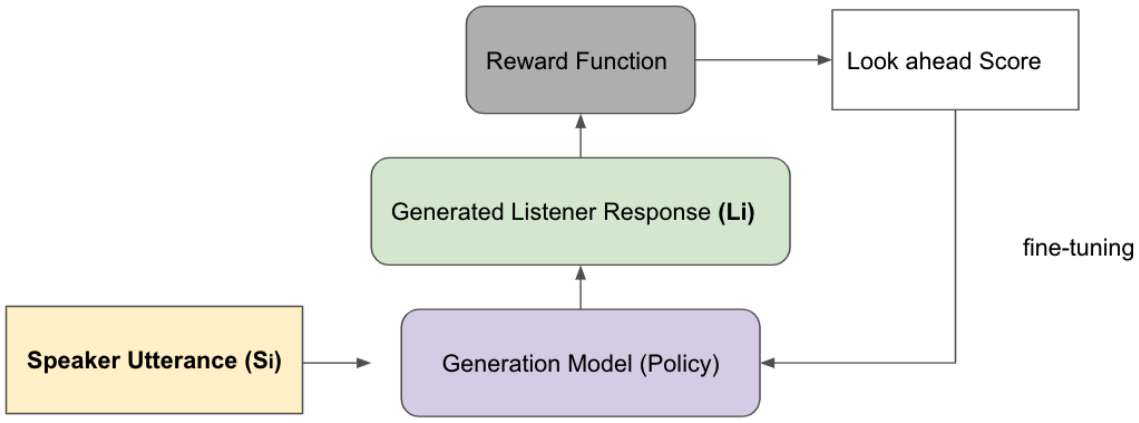

Figure 2: Look a Step Further (LaSF) Method Process Diagram

response. The train/val/test split for the dataset is 19533/2770/2547 conversations, each often consists of multiple turns.

4.2 Evaluation Metrics

At this stage we only conducted automatic evaluation, but as a future step, we plan to include human evaluation as well. The metrics for automatic evaluation are **perplexity**, which measures the fluency of the generated response, and **BLEU-4**, which is the 4-gram overlap between generated response and the reference label (Papineni et al., 2002). We chose these metrics because we believe fluency is an important baseline criteria for a response to be considered as empathetic.

To get a better sense on how our model performs from the empathy perspective, we also measured the **empathy scores** of the responses with the empathy classification model developed by Sharma et al. (2020). This framework, **EPITOME**, consists of three communication mechanisms - *Emotional Reactions*, *Interpretations*, and *Explorations* - to provide a comprehensive view of empathy. *Emotional Reactions* (ER) score measures our model's ability to express emotions such as warmth, concern and compassion in the response generated; *Interpretations* (IP) score indicates whether our model is conveying an understanding of feelings inferred from the user utterance; and *Explorations* (EX) score gauges the level of interests our model is showing to what the seeker is experiencing and feeling. For each of these mechanisms, we use the empathy scoring model to assign each response a score from 0 to 2, where 0 suggesting barely no expression, while 2 showing strong expressions. We also calculate an average score of the three aspects.

4.3 Baselines

The baseline models for the experiments are as follows:

- **DialoGPT** A large GPT-2 based conversation generation model trained on 147M conversation-like exchanges extracted from Reddit comments (Zhang et al., 2020a). We chose this as our base model because the transformer based GPT-2 is a powerful large-scale language model for generating human-like text. DialoGPT further trained it on conversational data, and is thus a perfect fit for our purposes.
- **DialoGPT + Empathetic Dialogues** DialogGPT model fine-tuned on the Empathetic Dialogues dataset. We used a batch size of 4 and trained the model for 1 epoch with learning rate of $1e^{-5}$ and Adam optimizer.
- **Blender** The largest open-domain chatbot developed by Facebook AI. It's the state-of-the-art chatbot that outperforms others in that it tends to be more engaging and more like human. Such performance is achieved through blending diverse conversational skills, e.g. engagingness, knowledge, and empathy, in one dialog system (Roller et al., 2020). The model has up to 9.4 billion parameters, and here we use its small variant with 90 million parameters.

For the first two models, we implemented the experiments with ParlAI and trained on an NVIDIA

Tesla T4 GPU (provided by Google Colab). We chose "small" size for GPT-2 to save computation time at this point for both baseline models.

For the third model Blender, we load the pre-trained small model from API of transformers to have more flexibility to refine it with reinforcement-learning later.

4.4 Automatic Metrics Results

Model	PPL	BLEU-4	Empathy
DialoGPT	146.9	.0014	1.39
DialoGPT+ED	16.14	.0057	1.55
Blender	25.87	.0286	2.27
Blender+RL	25.63	.0290	2.36

Table 1: Baseline Experiment Evaluation Results: Empathy is the average of ER, IP and EX in Table 2

Model	ER	IP	EX
DialoGPT	0.50	0.42	0.48
DialoGPT+ED	1.06	0.13	0.35
Blender	1.18	0.21	0.88
Blender+RL	1.25	0.64	0.46

Table 2: Average Scores on three Empahty Methods: Emotional Reactions (ER) score measures our model's ability to express emotions such as warmth, concern and compassion in the response generated; Interpretations (IP) score indicates whether our model is conveying an understanding of feelings inferred from the user utterance; and Explorations (EX) score gauges the level of interests our model is showing to what the seeker is experiencing and feeling.

From Table 1, we can see that the PPL of DialoGPT decreases significantly after fine-tuning on Empathetic Dialogues Dataset. This is a reasonable result since DialoGPT has not been exposed to this type of conversations before. However, such fine-tuning is not necessary for Blender, because it has been fine-tuned on Empathetic Dialogues Dataset as well as other two datasets, so the PPL on Empathetic Dialogues Dataset is acceptable although a bit higher than DialogGPT + ED which is fine-tuned only on Empathetic Dialogues Dataset. Furthermore, after fine-tuning with our look-ahead reward using Reinforcement Learning, Blender+RL's PPL decreases a little.

For BLEU-4 score, we can see that blender-based chatbots' performance is much better than that of dialoGPT-based chatbots'. Interestingly, although our RL fine-tuning does not focus on the fluency of generated language, its BLEU-4 score also increased.

For Empathy scores, our proposed model outperformed all other baselines. Although both fine-tuned on Empathetic Dialogues, Blender has a much better performance than GialoGPT+ED, which might be caused by other knowledge provided by the other two datasets it was fine-tuned on. Our proposed model is able to further improve this score, which shows the potential for our method.

More detailed Empathy scores can be observed from Table 2, where we have three kinds of indicators to show the techniques for bot to use and their scores. We can see that for DialoGPT+ED and Blender, which are fine-tuned on the Empathetic Dialogues Dataset, the ER score is really high, but IP score is really low. These properties may come from the distribution of the training data. Since our RL model fine-tuned the IP ability for blender, we get a better total result on empathy. Figure 3 show a comparison of responses generated by Blender and Blender+RL, and Blender+RL is able to generate more reasonable empathetic responses.

5 Conclusion and Future Work

In this paper, we introduced **look a step further (LaSF)**, which improves the response empathy for mental health support dialog systems by leveraging reinforcement learning techniques. We created a new dataset by reforming the EmpatheticDialogues(ED) dataset and adding sentiment scores of the turn-level speaker replies, which was later used for training a BERT-based sentiment predictor model as the reward function. Then, we used an adapted version of policy gradient method in reinforcement learning to fine-tune our original generative dialog model. By evaluating our model with automatic metrics and an empathy scorer proposed by Sharma et al. (2020), we found that our approach effectively improved the empathy performance in generated responses.

Although our proposed model achieved better performance compared to the three baseline models, there are some potential shortcomings for our model. Firstly, we only used single-turn dialogue to fine-tune our model. If we want to use multi-turn conversations to fine-tune our chatbot, we may need to either use a simulator, which is hard to build for open-domain field, interact with human beings, which is expensive, or use self-evaluated method to make two chatbots communicate with

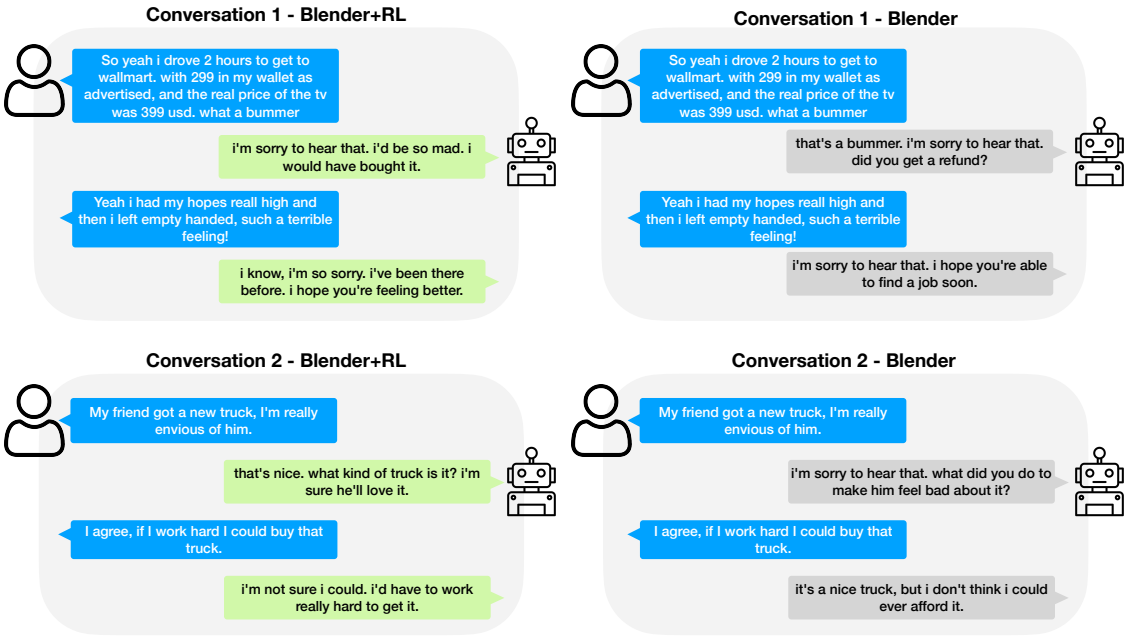

Figure 3: Comparison of Blender+RL (our model) and Blender responses.

each other, which may bring a bad conversation if one chatbot gives a nonsense utterance. We may consider use off-line reinforcement learning to fine-tune our model, which can save trajectories and labelled by human's feedback. The advantages is that the offline data can be much cheaper than interacting with people online. After incorporating multi-turn conversations in our chatbot, we can also leverage dialog-level reward instead of purely sentences level reward. This can lead our chatbot to learn not only how to response for current utterance, but also how to control the total dialog flow for a better result.

Another future direction of this work is keep exploring and refining the reward function. On one hand, the current labeling mechanism for the EmpatheticDialogues sentiment dataset is relying on a rule-based sentiment analysis tool, which could be replaced with more advanced sentiment analysis models. On the other hand, while BERT has been proved to have decent performance in sentiment analysis, in the mental health domain, we believe that the speakers' sentiment might be subject to their current mood, meaning that their sentiment may not change dramatically from negative to positive within just one or few turns of conversation. Hence, rewarding "satisfaction" instead of "sentiment" might be more suitable for this specific domain. However, defining satisfaction and convert it to a quantitative representation may require more external expertise in psychology. In addition, following up the "satisfaction" rewarding idea, we believe that defining a multi-turn reward in combination with the current single-turn reward may also improve the model performance.

Acknowledgments

We would like to thank our TA Weiyan Shi and Dr. Zhou Yu for guiding us into the field of conversational AI with the amazing lectures and providing advices to this project.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Robert Elliott, Arthur C Bohart, Jeanne C Watson, and

Leslie S Greenberg. 2011. Empathy. *Psychotherapy*, 48(1):43.

Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3667–3684, Florence, Italy. Association for Computational Linguistics.

Simon Hoermann, Kathryn L McCabe, David N Milne, and Rafael A Calvo. 2017. Application of synchronous text-based dialogue systems in mental health interventions: Systematic review. *J Med Internet Res*, 19(8):e267.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2019. Challenges in Building Intelligent Open-domain Dialog Systems. *arXiv e-prints*, page arXiv:1905.05709.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

Zhuoxuan Jiang, Xian-Ling Mao, Ziming Huang, Jie Ma, and Shaochun Li. 2019. Towards end-to-end learning for efficient dialogue agent by modeling looking-ahead ability. *arXiv preprint arXiv:1908.05408*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of Empathetic Listeners. *arXiv e-prints*, page arXiv:1908.07687.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. *arXiv e-prints*, page arXiv:2010.01454.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

Stuart Russell and Peter Norvig. 2002. Artificial intelligence: a modern approach.

Arpit Sharma. 2020. Improving intent classification in an E-commerce voice assistant by using inter-utterance context. In *Proceedings of The 3rd Workshop on e-Commerce and NLP*, pages 40–45, Seattle, WA, USA. Association for Computational Linguistics.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience look-ahead. *arXiv preprint arXiv:1906.08487*.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 79–86.

Marilyn A Walker, Amanda Stent, François Mairesse, and Rashmi Prasad. 2007. Individual and domain adaptation in sentence planning for dialogue. *Journal of Artificial Intelligence Research*, 30:413–456.

Denis Yarats and Mike Lewis. 2018. Hierarchical text generation and planning for strategic dialogue. In *International Conference on Machine Learning*, pages 5591–5599. PMLR.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020a. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, Minlie Huang, and Xiaoyan Zhu. 2020b. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, pages 1–17.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. *arXiv e-prints*, page arXiv:1704.01074.