Interspeech 2024
1-5 September 2024, Kos, Greece

Applying Reinforcement Learning and Multi-Generators for Stage Transition in an Emotional Support Dialogue System

Jeremy Chang, Kuan-Yu Chen, Chung-Hsien Wu

Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
jeremychang9, kuanyu.theme, chunghsienwu @gmail.com

Abstract

The use of empathetic dialogue systems has grown recently. However, establishing them for users experiencing mental depression requires more advanced consoling skills. In this paper, a dialogue system based on Emotional Support was developed. The system offers coping strategies through stages designed to address users' distress in long-term conversations. It employs a recurrent-based approach integrated with reinforcement learning for a decision model, which selects a generator from three specialized conditional generation models to generate empathetic responses. Experimental results showed improvements in BLEU, Rouge-L, and Distinct-n metrics compared to the baseline. On average, the system's BLEU score increased by 0.87, Rouge-L by 1.85, Distinct-1 by 0.69, and Distinct-2 by 2.26. As a result, the system generates responses aligned with Emotional Support skills, ultimately comforting the user's distress.

Index Terms: emotional support, empathetic responses, reinforcement learning, long-term dialogue system

1. Introduction

The primary goal of an empathetic dialogue system is to understand users' expressions and provide human-like responses, demonstrating emotional understanding and empathy. Current approaches often focus on recognizing users' emotions and context-relevant knowledge to offer comfort [2-5]. However, the empathy and relevance of these responses remain inadequate. For example, phrases like "I'm sorry to hear that," and "That's great" are frequently the only type of response generated [6]. Hence, the strategy and variety for empathetic response generation should become more complex [7-9]. In this study, an Emotional Support dialogue system [10] was proposed to enrich the strategy of generating empathetic responses. The system generates different strategies of responses based on three Emotional Support Stages [11] to effectively comfort the users (help-seekers) and decrease their emotion intensity.

The task of Emotional Support involves expressing empathy and emotions to alleviate the user's distress. It utilizes support strategies across three stages: Exploration, Comforting, and Action (To avoid confusion with reinforcement learning, we will be using the term 'Suggesting' instead of 'Action' throughout the remainder of the paper.) Emotional Support Conversations dataset (ESConv) [10] combines emotional chatting and empathetic responding, demanding a deep understanding to offer effective aid.

To integrate Emotional Support into the system, this study created three specialized generators, each corresponding to a specific stage. Each generator has a unique conditional model, enhancing response generation within their stages. Additionally, the system employs a novel recurrent-based approach integrated with reinforcement learning for a decision model. This decision model selects the suitable generator based on the conversation, determining the response generation strategy. Consequently, the proposed system generates diverse responses that progressively alleviate users' distress through Emotional Support stages. Refer to Figure 1 for the system overview.

2. Proposed methods

2.1. Conditional models training

There are three conditional models: a problem detection model, an emotion detection model, and an intent generation model. The problem and emotion detection models utilize the BERT-base model [12]. For problem detection, we employ an augmented version of ESConv, consisting of sentences labeled with 12 types of problems. The augmentation details are explained in Section 3. For emotion detection, the Go-Emotions dataset [13] is employed, which consists of sentences labeled with 27 emotion categories. Equation (1) illustrates the formula for mapping BERT to the problem types and emotion classes using multi-class classification along with the cross-entropy loss function.

$$L_{CE}(\hat{y}, y) = -\frac{1}{N}\sum_{i=1}^{N} y^{(i)} log\hat{y}^{(i)} \quad (1)$$

The intent generation model utilizes the DialoGPT model as its backbone. This model is also trained on an augmented version of ESConv, comprised of knowledge extracted from COMET [14]. As depicted in (2), training involves utilizing the negative log-likelihood loss function. This leads the model to generate the user's intent sentence, such as "to move forward" or "to find a place to run".

$$L_{NLL}(\hat{y}, y) = -\frac{1}{N}\sum_{i=0}^{N} P(y_i) \cdot logsoftmax\big(P(\hat{y}_i)\big) \quad (2)$$

10.21437/Interspeech.2024-882

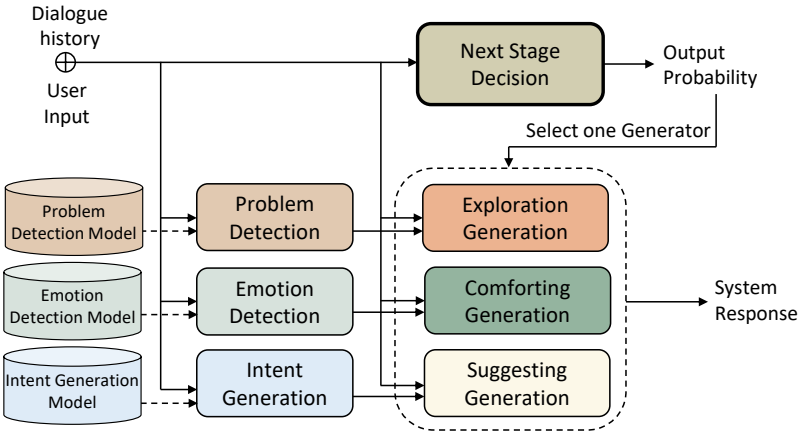

Figure 1: *System overview.*

2.2. Conditional generation models training

There are three distinct DialoGPT [1] models within the system, each dedicated to a specific stage: Exploration, Comforting, and Suggesting. To improve the response quality, conditions from the models described in Section 2.1 are integrated into the generation. These conditions are appended to the input sequence's end using special tokens, similar to the prefix technique employed in T5 [15]. This integration ensures that each generation model produces unique responses aligned with the corresponding stages.

To train the exploration generator, sentences are extracted from ESConv's exploration section, each labeled with the "Questioning" and "Restatement or Paraphrasing" strategies. This generator generates responses that explore the user's information. The comforting generator is trained with the sentences from ESConv's comforting section labeled with "Reflection of feelings," "Self-disclosure," and "Affirmation and Reassurance" strategies. Its responses are designed to resonate with the user's emotional state. For the suggesting generator, sentences are extracted from ESConv's suggesting section, each labeled with the "Providing Suggestion" and "Information" strategies. This generator produces responses that guide the user in resolving their problem.

2.3. Conditional generation models training

The Next Stage Decision Model aims to effectively employ Emotional Support skills during conversations by determining the most appropriate stage, as shown in Figure 2. It consists of two training phases: the pretraining and the Reinforcement Learning phase.

The motivation behind using Reinforcement Learning (RL) lies in the assumption that determining a perfect strategy is difficult. In certain cases, multiple strategies could all be appropriate. By incorporating a user model and RL, the system becomes more adaptable to various conversations.

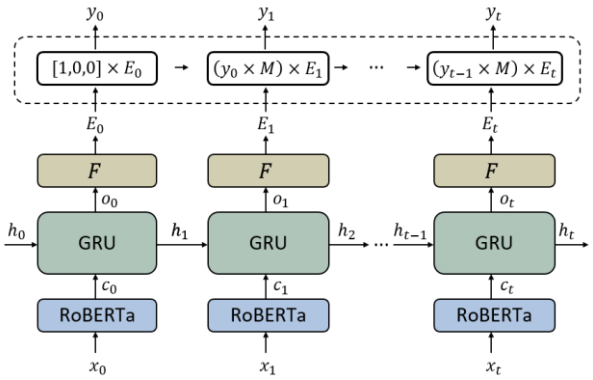

Figure 2: *Recurrent stage probability estimation.*

In the pretraining phase, we utilize RoBERTa [16] and GRU [17] in conjunction with a Markov chain matrix to estimate the probability of stage transitions. We analyze ESConv and establish a Markov chain that captures the transitions between the three stages. The resulting Markov chain is visualized in Figure 3. Using the obtained Markov chain, we derive a 3×3 transition matrix, denoted as *M*. This matrix allows us to estimate the probabilities of transitioning between stages.

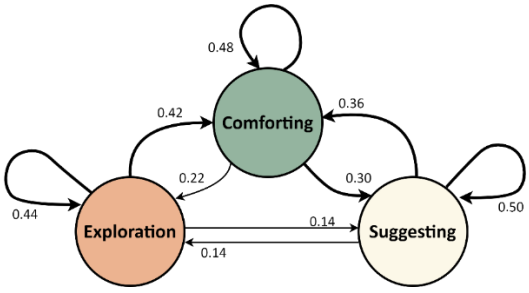

Figure 3: *Markov chain of ESConv dataset.*

By utilizing RoBERTa, the model encodes and obtains context embedding c_t. Next, context embedding c_t is fed into GRU. It generates an output vector o_t, along with a hidden state h_t that is carried forward to the next iteration. Mapping the output vector o_t to a 3×3 dimensional space with each element

representing a specific state. Then, we apply the softmax function to each row, resulting in an emission matrix E_t.

To estimate the probability of each stage, we update the previous stage distribution y_{t-1} by matrix multiplying it with the transition matrix M. This step considers the Markov chain information and captures local dependencies to derive the next possible stage distribution. Finally, the updated stage distribution is a matrix multiplied by the emission matrix E_t to derive the stage probability distribution y_t.

During the Reinforcement Learning phase, a user model is trained and frozen to simulate a depressed help-seeker. This enables our system to interact with the user model, gaining rewards for RL and improving the decision model. The training process for the user model mirrors that of the generators. It is also a DialoGPT model, trained using sentences from help-seekers in ESConv. This self-play framework is inspired by Li et al. [18], as depicted in Figure 4. It comprises three main components: dialogue history, agent, and environment.

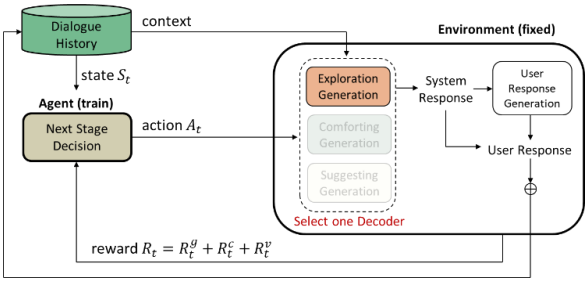

Figure 4: *Self-play Reinforcement Learning framework.*

The Reinforcement Learning begins by passing the current dialogue history, represented as the state S_t, into the agent. The agent then produces a stage probability distribution, from which we sample an action A_t. This action corresponds to selecting a generator from the three conditional models for generating the response, which is passed to the user model. This allows us to obtain a simulated user response, mimicking the behavior of a help-seeker. The system and user responses are appended to the dialogue history, with the history window set to 5. At this point, we calculate a set of rewards R_t for the current time step. The agent's training employs REINFORCE [19], optimizing the parameters θ to maximizes the total rewards across all conceivable trajectories, as formulated in (3). E represents the expected value, R represents the reward, and τ signifies the trajectory of each episode, which is the sequence of states, actions, and rewards encountered by the agent during an episode.

$$\theta = argmax_\theta(E[R(\tau|\pi(A|S,\theta))]) \quad (3)$$

To ensure that the agent achieves the goal of reducing the user's emotional distress, this paper employs a goal reward R_t^g along with reward shaping [20], which is a technique that involves offering small rewards whenever the agent makes correct decisions. This paper incorporates a conversation flow reward R_t^c and a valence reward R_t^v. The total reward is depicted as (4).

$$R_t = R_t^g + R_t^c + R_t^v \quad (4)$$

Goal Reward (R_t^g): The goal reward assesses the agent's effectiveness in alleviating the user's emotional distress, calculated only when the whole conversation ends. We utilize a BERT-based valence prediction model trained on the Stanford Sentiment Treebank dataset [21]. The reward is determined by the valance value of the first user sentence $Val(x_0)$ and the final user sentence $Val(x_{T-1})$, scaled by λ as shown in (5). In this paper, λ is set to 100.

$$R_t^g = \lambda * \big(Val(x_t) - Val(x_0)\big), \ if\ t = T-1 \quad (5)$$

Conversation Flow Reward (R_t^c): By analyzing dialogues in ESConv, we noted that user reactions can be expected following responses, leading to mood improvement. To model this, we train a discriminator via the Next Sentence Prediction task. This discriminator gauges whether a given user sentence x_t aligns with anticipated the conversation flow, wherein the historical user sentences are concatenated into a sequence. We employ RoBERTa as the discriminator's core, fine-tuning it using ESConv. The formula is illustrated in (6). m is the sliding window of the conversation flow.

$$R_t^c = \begin{cases} 1, & if\ Discriminator([x_{t-m-1}, x_{t-1}], x_t) = \\ -1, & if\ Discriminator([x_{t-m-1}, x_{t-1}], x_t) = \end{cases} \quad (6)$$

Valence Reward (R_t^v): The emotional valence of the user can be expected based on the responses. We observe that during the exploration stage, the user's valence generally decreases, whereas during the comforting and suggesting stages, it either increases or remains unchanged. To calculate the valence reward R_t^v, we employ the valence prediction model mentioned in the goal reward section. By comparing the valence value between the previous user sentence and the current user sentence, the valence reward is obtained. The formulation is represented in (7).

$$R_t^v = \begin{cases} 1, & if\ Val(x_t) < Val(x_{t-1}) \cdot stage = explore \\ 1, & if\ Val(x_t) \geq Val(x_{t-1}) \cdot stage = comfort \\ 1, & if\ Val(x_t) \geq Val(x_{t-1}) \cdot stage = suggest \\ -1, & Otherwise \end{cases} \quad (7)$$

3. Dataset

Emotional Support Conversations (ESConv) includes annotations for support strategies, user emotion intensity, and problem types. The sentences were classified into seven strategies: Questioning, Restatement or Paraphrasing, Reflection of Feelings, Self-disclosure, Affirmation and Reassurance, Providing Suggestion, and Information. These strategies can be mapped into three Emotional Support stages. Each dialogue was further categorized into 12 problem types. The emotion intensity of the user was labeled both at the start and the end of the conversation.

To address the data imbalance, in this study, ChatGPT [22] is utilized for dataset augmentation, yielding a total of 14,380

Table 1: *Performance of the response generation and the next stage decision model.*

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	Distinct-n	Stage Acc.
BlenderBot-Joint [10]	19.94	7.93	3.96	2.29	18.08	3.6/21.87	-
DialoGPT [1]	12.61	7.34	5.27	3.97	16.37	4.21/22.19	-
+ RoBERTa	13.16	7.72	5.56	4.19	17.26	4.31/22.38	48.42
+ GRU	13.57	7.89	5.67	4.27	17.41	4.78/22.97	49.79
+ Markov Chain	13.75	7.95	5.69	4.28	17.64	4.82/23.32	50.44
+ Goal Reward	13.92	7.99	5.70	4.28	17.85	4.79/24.03	50.90
+ Valance Reward	14.18	8.08	5.74	4.29	17.93	4.73/23.76	51.55
+ Conv Flow Reward	**14.26**	**8.19**	**5.84**	**4.38**	**18.22**	**4.90/24.45**	**52.44**

Table 2: *Generation example.*

Role	Stage	Dialogue
User	-	Hello, my name is Rose, how are you today?
System	Exploration	I am well thanks. **How are you doing today?** (Questioning)
User	-	I am doing alright. I am having a lot of anxiety.
System	Exploration	That is not good. **What seems to be causing the anxiety?** (Questioning)
User	-	My hair is unruly because of my appearance.
System	Comforting	**It must be really hard for you to feel this way.** I'm sorry to hear that. (Reflection of Feelings)
User	-	It is really hard for me. I feel like I have too much hair.
System	Comforting	**It can be overwhelming to maintain all that hair.** Sounds like you're trying to get it cut to be less messy. (Affirmation and Reassurance)
User	-	Yes, I am.
System	Suggesting	Have you **thought about doing a fade haircut? (**Providing Suggestion)
User	-	I haven't tried that. But it sounds like a great idea!

problem type sentences. For identifying the help-seeker's intent, COMET is employed, specifically utilizing the "xIntent" relation, to further enhance the dataset.

4. Experimental results

As illustrated in Table 1, we conducted comparison and ablation studies on the next stage decision model, comparing our system to the baselines, which were fine-tuned on ESConv. Using the next stage decision model showed statistically significant improvement compared to DialoGPT, as confirmed by paired t-tests with p-values < 0.05 and the bootstrapping approach [23] to compute confidence intervals. When compared to using only RoBERTa, the integration of GRU and Markov chain led to a noteworthy 2.02% accuracy improvement in a 3-class stage classification task, positively impacting our dialogue system's performance.

On the other hand, BlenderBot-Joint [10] showed great performance in BLEU-1, BLEU-2, and Rouge-L, generating responses that resembled the golden responses. However, the diversity of the generated responses was not as good as our system, as indicated by BLEU-3, BLEU-4, and Distinct-n.

The findings emphasized the effect of both goal reward and reward shaping rewards for Reinforcement Learning. Notably, relying solely on goal reward without the guidance of reward shaping resulted in challenges for training the next stage decision model. This underscores that reward shaping contributed to improvements in both stage decision accuracy and the overall system performance.

Table 2 presents a generation example from our system. The generated responses successfully followed to the ideal stage transition and employed strategies based on Emotional Support. The system began with the strategy of exploration to understand the situation, provided empathetic comfort to the user, and concluded with a practice suggestion.

5. Conclusion

This paper introduces an Emotional Support dialogue system that employs multiple generators to generate triple-stage responses: exploration for gathering user information, comforting for emotional solace, and suggesting for aiding problem-solving.

Our system employs a combined training approach, beginning with a recurrent-based method for managing stage transitions, followed by Reinforcement Learning to mitigate user emotional distress. The evaluation, which includes BLEU, Rouge-L, and Distinct metrics, demonstrates notable enhancements compared to DialoGPT. Specifically, our system attains an average improvement of 0.87 in BLEU, 1.85 in Rouge-L, 0.69 in Distinct-1, and 2.26 in Distinct-2. These results validate the effective implementation of our Emotional Support strategy within our system.

6. References

[1] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and W. B. Dolan, "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*, 2020, pp. 270-278.

[2] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5370-5381.

[3] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Information Fusion,* vol. 64, pp. 50-70, 2020.

[4] S. Sabour, C. Zheng, and M. Huang, "Cem: Commonsense-aware empathetic response generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, vol. 36, no. 10, pp. 11229-11237.

[5] L. Wang, J. Li, Z. Lin, F. Meng, C. Yang, W. Wang, and J. Zhou, "Empathetic Dialogue Generation via Sensitive Emotion Recognition and Sensible Knowledge Selection," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 4634-4645.

[6] X. Gao, Y. Zhang, M. Galley, C. Brockett, and W. B. Dolan, "Dialogue Response Ranking Training with Large-Scale Human Feedback Data," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 386-395.

[7] E. M. Smith, M. Williamson, K. Shuster, J. Weston, and Y.-L. Boureau, "Can You Put it All Together: Evaluating Conversational Agents' Ability to Blend Skills," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 2021-2030.

[8] Y.-H. Wang, J.-H. Hsu, C.-H. Wu, and T.-H. Yang, "Transformer-based empathetic response generation using dialogue situation and advanced-level definition of empathy," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2021: IEEE, pp. 1-5.

[9] J.-H. Hsu, J. Chang, M.-H. Kuo, and C.-H. Wu, "Empathetic Response Generation based on Plug-and-Play Mechanism with Empathy Perturbation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 2023.

[10] S. Liu, C. Zheng, O. Demasi, S. Sabour, Y. Li, Z. Yu, Y. Jiang, and M. Huang, "Towards Emotional Support Dialog Systems," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 3469-3483.

[11] C. E. Hill, *Helping skills: Facilitating, exploration, insight, and action, 3rd ed* (Helping skills: Facilitating, exploration, insight, and action, 3rd ed.). Washington, DC, US: American Psychological Association, 2009.

[12] J. D. M.-W. C. Kenton and L. K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171-4186.

[13] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4040-4054.

[14] J. D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, and Y. Choi, "(Comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, no. 7, pp. 6384-6392.

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research,* vol. 21, pp. 1-67, 2020.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692 doi:10.48550/arXiv.1907.11692.

[17] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724-1734.

[18] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao, "Deep Reinforcement Learning for Dialogue Generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 1192-1202.

[19] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning,* vol. 8, pp. 229-256, 1992.

[20] Y. Wu and Y. Tian, "Training agent for first-person shooter game with actor-critic curriculum learning," in *International Conference on Learning Representations*, 2016.

[21] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631-1642.

[22] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774 doi: 10.48550/arXiv.2303.08774.

[23] Ferrer, L. and Riera, P. Confidence Intervals for evaluation in machine learning [Computer software]. https://github.com/luferrer/ConfidenceIntervals.