Knowledge-Infused Topic Model for Empathetic Dialogue Response

Po-Chuan Chen Jen-Tzung Chien
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
E-mail: {present90308.ee11, jtchien}@nycu.edu.tw

***Abstract*—Empathetic dialogue generation aims to produce the responses that reflect an understanding of dialogue content in a multi-turn dialogue system through emotional expression. Traditional methods focus on enhancing the emotional response prediction or improving the dialogue generation either through the reinforcement learning or based on the topic models. This study employs a graph neural network to learn the characteristics and behaviors over empathetic dialogue interactions. Such a network is further integrated with recent advancements in topic model scheme to enhance the dialogue topic representation. This model is additionally fused with the external knowledge to impose relevance in the generated response. Experimental results demonstrate that the proposed method surpasses some related methods in automatic and human evaluations, showing improvement in terms of empathy, relevance, fluency, text perplexity, generation diversity and emotion classification accuracy.**

I. Introduction

Open-domain dialogue systems [1] have undergone remarkable advancements, transitioning from early rule-based models [2] to sophisticated neural network architectures [3], [4] that support a broad range of conversational topics and contexts. While substantially skilled at generating diverse and coherent responses, these systems often fail to deliver conversations that exhibit genuine empathy, a critical component to foster an emotional and intelligent interaction. Empathy in a dialogue system can sufficiently enhance user satisfaction and build trust and close relationships in human-computer interactions. Empathy becomes even more crucial as users increasingly seek more personalized and responsive interactions. Empathy in a dialogue system is not just about understanding the words a user says but also about grasping the underlying emotions and responding in a way that acknowledges and respects those feelings. An example is shown in Figure 1(a) where an empathetic dialogue involves complex processes such as emotion detection, sentiment analysis, and context-aware response generation. Integrating these elements allows a dialogue system to interact beyond a pure transactional exchange towards a meaningful and supportive conversation that can significantly improve user experience.

Recent approaches to empathetic dialogue generation [5], [6] were ranged from developing the emotion recognition models [7], [8] to designing the affective computation components [9], and integrating the multi-modality data [10]. Emotion recognition would like to identify and interpret the emotional states conveyed in user inputs, enabling the system to tailor its responses. Affective computing was implemented to extend this model by incorporating physiological and contextual data, which helped comprehensively understand the emotional states of a user. This study proposes a new topic model [11]–[13] for empathetic dialogue response based on a knowledge-infused topic model (KITM). This model is learned to grasp the conversation contexts and the emotional nuances, as well as to produce the contextually appropriate and emotionally meaningful responses by integrating dialogue topics [14], [15] and external knowledge [16] where the topic and knowledge are aligned with the user emotional states, respectively. This paper further addresses the challenges in empathetic dialogue generation with a precise emotion detection. This study presents an advanced algorithm to balance between linguistic creativity and conversational constraints for empathetic dialogue generation. The experiments on EmpatheticDialogues dataset [17] were evaluated in terms of different metrics.

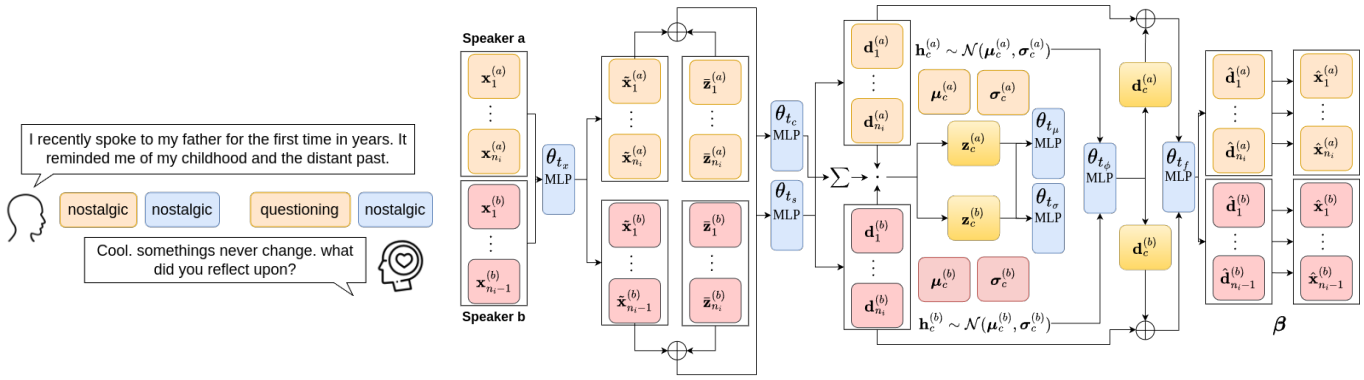
II. Knowledge-Infused Topic Model

Figure 2 shows presents an empathetic response generation based on an encoder-decoder-based transformer where a knowledge-infused process is implemented to build emotion flow model for two-speaker dialogue control based on graph convolution network. This study exploits a conversational topic model enriched by external knowledge where emotion prediction, dialogue understanding and generation are developed.

A. Inputs/Outputs in Empathetic Dialogue Generation

The task of empathetic dialogue generation [18], [19] involves producing emotional and contextual responses for interactive communications where the user information $\mathbf{X}^u$ and external knowledge $\mathbf{X}^k$ are merged. Inputs and responses of a dialogue system are first defined. This work adopts the common-sense decoder (COMET) [16] to generate five responses $\mathbf{X}^k$ based on the predefined questions for events related to 'intent', 'need', 'effect', 'react' and 'want'. User inputs consist of the dialogue utterances $\mathbf{X}^u = [\mathbf{x}^u_1, \ldots, \mathbf{x}^u_{N-1}]$ and the external knowledge responses $\mathbf{X}^k$ from COMET. In addition, system responses consist of the generated reply $\widehat{\mathbf{y}}$, the prediction of emotions and intents for each dialogue turn $\widehat{\mathbf{y}}^e = [\widehat{\mathbf{y}}^e_1, \ldots, \widehat{\mathbf{y}}^e_N]$, and the overall emotion-intent prediction for the entire conversation $\widehat{\mathbf{y}}^\star$. This study introduces six objectives $\{\mathcal{L}_e, \mathcal{L}_r, \mathcal{L}_d, \mathcal{L}_f, \mathcal{L}_g, \mathcal{L}_t\}$, which are jointly optimized to build an empathetic dialogue system based on the emotion-aware inputs $\{\mathbf{X}^u, \mathbf{X}^k\}$ and responses $\{\widehat{\mathbf{y}}, \widehat{\mathbf{y}}^e, \widehat{\mathbf{y}}^\star\}$.

979-8-3315-7206-8/25/$31.00



(a) (b)

Fig. 1: (a) Illustration for an example of a two-speaker dialogue. Blue tags show overall dialogue emotion while yellow tags show personal utterance emotion. (b) An overview of topic reconstruction for a two-speaker dialogue through the operations of concatenation ($\oplus$), summation (Σ) and dot product ($\cdot$).

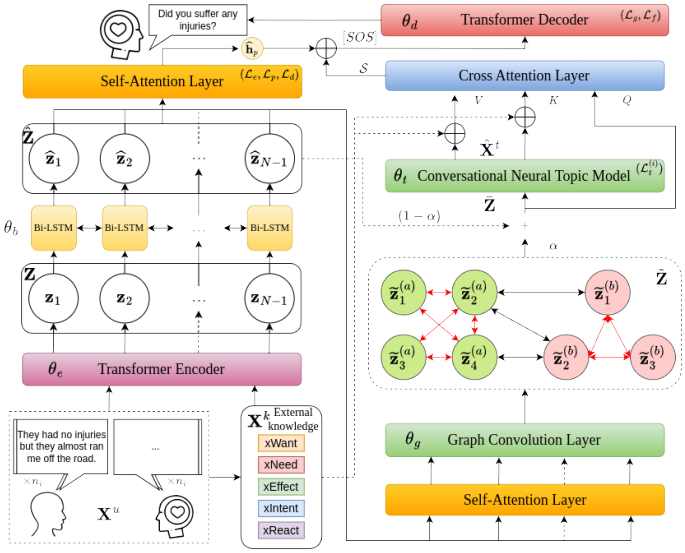


Fig. 2: An overview of encoder-decoder framework for empathetic dialogue generation via knowledge-infused topic model where a graph encoder for $\widetilde{\mathbf{Z}}$ is used.

B. Knowledge-Infused Emotion Representation

1) Conversation and knowledge encoder: A transformer encoder $f_{\theta_e}(\cdot)$ [20] is utilized to extract features or embeddings from both input utterances $\mathbf{X}^u$ and external knowledge $\mathbf{X}^k$ derived from COMET. This encoding process extract the input composite embeddings, which are represented as a concatenation of two encoding matrices $f_{\theta_e}(\mathbf{X}^u)$ and $f_{\theta_e}(\mathbf{X}^k)$ in $\mathbf{Z} = \{f_{\theta_e}(\mathbf{X}^u), f_{\theta_e}(\mathbf{X}^k)\}$ where $\mathbf{Z}$ consists of a sequence of $N-1$ vectors $\langle \mathbf{z}_1, \ldots, \mathbf{z}_{N-1} \rangle$ for $N-1$ turns. This is an encoding for user conversation and external knowledge from $\mathbf{X}^u$ and $\mathbf{X}^k$, respectively.

2) Emotion flow contextual model: Next, a bidirectional long short-term memory (Bi-LSTM) $f_{\theta_b}(\cdot)$ is introduced to capture the dependencies and contexts from past to future utterances within a dialogue trajectory consisting of utterances in N turns with their emotions. The source input for this process is a concatenation of the encoded utterances and the relevant external knowledge. This process represents the emotion flow through Bi-LSTM as $\widehat{\mathbf{Z}} = f_{\theta_b}(\mathbf{Z})$ where $\widehat{\mathbf{Z}} = [\widehat{\mathbf{z}}_1, \ldots, \widehat{\mathbf{z}}_{N-1}]$. This process is crucial for emotion flow representation, ensuring that historical and forthcoming contexts in two directions are considered.

3) Utterance emotion classification: Furthermore, the fine-grained emotion-intent classification task is performed by minimizing the cross-entropy loss between the emotion prediction output $\widehat{\mathbf{y}}^e$ and the one-hot emotion ground-truth $\mathbf{y}^e$. The probability of the predicted emotion $\widehat{\mathbf{y}}^e_j$ of the utterances in individual turn $\widehat{\mathbf{z}}_j$ is computed by using the softmax function where the linear transformation $\mathbf{W}_e$ is applied over Bi-LSTM outputs $\widehat{\mathbf{z}}_j$. The probability in turn j is expressed by

$$p(\widehat{\mathbf{y}}^e_j) = \text{Softmax}(\mathbf{W}_e \widehat{\mathbf{z}}_j) \tag{1}$$

which is used to calculate the emotion-intent classification loss from N turns in a form of cross-entropy error function

$$\mathcal{L}_e = -\sum_{j=1}^{N-1} \mathbf{y}^e_j \log p(\widehat{\mathbf{y}}^e_j). \tag{2}$$

4) Response emotion prediction: Another objective function is introduced to optimize the overall response of emotion prediction with totally N turns. To do so, the cross-entropy loss between the emotion-intent distribution of the predicted response $\widehat{\mathbf{y}}^e_N$ and the ground-truth label of target response $\mathbf{y}^e_N$ is minimized. The self-attention mechanism aggregates the encoded representations of the previous utterances and turns, denoted as $\widehat{\mathbf{h}}_p = \text{Att}(\langle \widehat{\mathbf{z}}_1, \ldots, \widehat{\mathbf{z}}_{N-1} \rangle)$. The probability of the predicted response emotion $p(\widehat{\mathbf{y}}^e_N)$ is then computed by using the softmax function on linear transformation $\mathbf{W}_p$ over the aggregated hidden representation $\widehat{\mathbf{h}}_p$ as

$$p(\widehat{\mathbf{y}}^e_N) = \text{Softmax}(\mathbf{W}_p \widehat{\mathbf{h}}_p). \tag{3}$$

The classification loss for response emotion $\mathcal{L}_r$ is defined as

$$\mathcal{L}_r = -\mathbf{y}_N^e \log p(\widehat{\mathbf{y}}_N^e). \tag{4}$$

5) Dialogue emotion prediction: In a conversation, the dialogue contains information from each utterance and the overall emotion-intent $\mathbf{y}^\star$ for the entire dialogue flow. Cross-entropy loss is minimized to optimize the summarization of conversational emotion. In the process, the self-attention mechanism aggregates the encoded representations of all utterances, denoted as $\widehat{\mathbf{h}}_d = \mathrm{Att}(\langle \widehat{\mathbf{z}}_1, \ldots, \widehat{\mathbf{z}}_{N-1} \rangle)$. Notably, self-attention layers are individual for response emotion $\widehat{\mathbf{h}}_p$ and dialogue emotion $\widehat{\mathbf{h}}_d$. The probability of the predicted emotion-intent for the dialogue $p(\widehat{\mathbf{y}}^\star)$ is then computed by using the softmax function on the linear transformation $\mathbf{W}_d$ over the aggregated hidden representation $\widehat{\mathbf{h}}_d$ in a form of

$$p(\widehat{\mathbf{y}}^\star) = \mathrm{Softmax}(\mathbf{W}_d \widehat{\mathbf{h}}_d). \tag{5}$$

The loss for predicting dialogue emotion $\mathcal{L}_d$ is measured by

$$\mathcal{L}_d = -\mathbf{y}^\star \log p(\widehat{\mathbf{y}}^\star). \tag{6}$$

This objective function ensures accurate summarization of the overall emotion-intent for the entire dialogue. The three loss functions $\mathcal{L}_e$, $\mathcal{L}_r$, and $\mathcal{L}_d$ are then combined to form the classification loss $\mathcal{L}_c$ for knowledge-infused emotion prediction.

6) Multi-role graph encoder: The multi-role graph encoder $f_{\theta_g}(\cdot)$ updates the representations according to multi-role relations where each speaker has n_i utterances. Representations $\widetilde{\mathbf{Z}} = \{\widetilde{\mathbf{z}}_j^{(i)}\}_{j=1}^{N-1}$ are defined for $\widetilde{\mathbf{z}}_j$ of turn j with n_i utterances of a speaker i. Under each representation $\widetilde{\mathbf{z}}_j^{(i)} = \{\widetilde{\mathbf{z}}_k^{(i)}\}$, $\widetilde{\mathbf{z}}_k^{(i)}$ is associated with a speaker ID $i \in a, b$ for two speakers and k is the index of utterances. This study considers two key relations including *intra-speaker*, ensuring topic consistency with bidirectional edges within a window size, and *inter-speaker*, determining topic shifts with bidirectional edges within a window size. The updated representation is $\widetilde{\mathbf{z}}_k^{(i)} = f_{\theta_g}(\mathbf{z}_k^{(i)})$ in $\widetilde{\mathbf{Z}}$. The graph encoder is implemented by a graph convolution network (GCN) [21], [22].

C. Knowledge-Infused Topic Representation

1) Conversational neural topic model (ConvNTM): ConvNTM $f_{\theta_t}(\cdot)$ generates the bag-of-words (BoW) representations $\widehat{\mathbf{X}}^t$ from topic representation $\bar{\mathbf{Z}}$ [23] and weighted matrix $\boldsymbol{\beta}$. The topic representation $\bar{\mathbf{Z}}$ is a combination of $\widehat{\mathbf{Z}}$ and $\widetilde{\mathbf{Z}}$ with α set as 0.5 in $\bar{\mathbf{Z}} = (1-\alpha)\cdot\widehat{\mathbf{Z}} + \alpha \cdot \widetilde{\mathbf{Z}}$. The topic-based loss of a speaker i is yielded as a variational upper bound of log likelihood of data $\mathbf{w} \in \widehat{\mathbf{X}}^t$ with the latent topic variable $\mathbf{d}_c^{(i)}$, modeled by Gaussian prior $p(\cdot)$ and variational posterior $q(\cdot)$ with parameters $\{\boldsymbol{\mu}_c^{(i)}, \boldsymbol{\sigma}_c^{(i)}\}$ and the weight parameters $\boldsymbol{\beta}$

$$\begin{aligned}\mathcal{L}_t^{(i)} = -\mathbb{E}_{q(\mathbf{d}_c^{(i)}|\boldsymbol{\mu}_c^{(i)}, \boldsymbol{\sigma}_c^{(i)})} \Bigg[\sum_{k=1}^{n_i} \sum_{w} \log p(\mathbf{w}|\mathbf{d}_c^{(i)}, \boldsymbol{\beta})) \\ + w_{kl} \cdot D_{\mathrm{KL}}(q(\mathbf{d}_c^{(i)}|\boldsymbol{\mu}_c^{(i)}, \boldsymbol{\sigma}_c^{(i)}) \| p(\mathbf{d}_c^{(i)}))\Bigg]\end{aligned} \tag{7}$$

where w_{kl} is a hyperparameter and $D_{\mathrm{KL}}(\cdot)$ is the Kullback-Leibler (KL) divergence. Following the topic reconstruction process in Figure 1(b), the embedding $\widetilde{\mathbf{x}}_k^{(i)}$ is extracted as BoW features using multi-layer perceptron (MLP) $f_{\theta_{t_x}}(\cdot)$. The role topic $\mathbf{d}_k^{(i)}$ is learned via $f_{\theta_{t_s}}(\cdot)$ by incorporating the source latent variable $\bar{\mathbf{z}}_k^{(i)}$. The dialogue history $\mathbf{z}_c^{(i)}$ is then computed as

$$\mathbf{z}_c^{(i)} = \tanh\left(\sum_{k=1}^{n_i} f_{\theta_{t_c}}(\langle \widetilde{\mathbf{x}}_k^{(i)}, \bar{\mathbf{z}}_k^{(i)} \rangle) \cdot \mathbf{d}_k^{(i)}\right). \tag{8}$$

The dialogue topic probability $\mathbf{d}_c^{(i)}$ is determined by using $f_{\theta_{t_\phi}}(\cdot)$ with latent variable $\mathbf{z}_c^{(i)}$ from $\mathcal{N}(\boldsymbol{\mu}_c^{(i)}, \boldsymbol{\sigma}_c^{(i)})$. $\boldsymbol{\mu}_c^{(i)}$ and $\log \boldsymbol{\sigma}_c^{(i)}$ are computed using $f_{\theta_{t_\mu}}(\cdot)$ and $f_{\theta_{t_\sigma}}(\cdot)$, respectively. BoW is reconstructed with weight parameters $\boldsymbol{\beta}$

$$\widehat{\mathbf{x}}_k^{(i)} = \mathrm{Softmax}(f_{\theta_{t_f}}(\langle \mathbf{d}_k^{(i)}, \mathbf{d}_c^{(i)} \rangle \boldsymbol{\beta})). \tag{9}$$

In the knowledge-infused topic model, the reconstructed BoW representations $\widehat{\mathbf{X}}^t = \{\widehat{\mathbf{x}}_1^{(i)}, \ldots, \widehat{\mathbf{x}}_{n_i}^{(i)}\}$ for each speaker are aggregated, informing responses with specific topics and contextual knowledge. The training loss $\mathcal{L}_t^{(i)}$ in Eq. (7) includes a cross-entropy loss for reconstructing the BoW representations and a KL divergence term for regularizing the variational posterior of latent topic variables $\mathbf{d}_c^{(i)}$, and the hyperparameter w_{kl} was fixed as 0.01. The overall loss $\mathcal{L}_t$ is the sum of the losses for both speakers, i.e. $\mathcal{L}_t = \mathcal{L}_t^{(a)} + \mathcal{L}_t^{(b)}$.

2) Learning external knowledge in empathetic generation: The process of combining external knowledge results in using cross-attention mechanism to select relevant information $\mathcal{S} = \mathrm{CrossAtt}(\bar{\mathbf{Z}}, \langle \mathbf{X}^k, \widehat{\mathbf{X}}^t \rangle, \langle \mathbf{X}^k; \widehat{\mathbf{X}}^t \rangle)$ through query, key and value. For text generation, the frequency-aware cross-entropy [24] is used to penalize high-frequency tokens via the loss function $\mathcal{L}_f = -\sum_{t=1}^{T}\sum_{i=1}^{V} w_i \epsilon_t(\mathbf{c}_i) \log p(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{X}^u)$ where w_i is the frequency weight for token $\mathbf{c}_i$ from V vocabulary words and $\epsilon_t(\mathbf{c}_i)$ checks if token $\mathbf{c}_i$ matches the ground truth. The standard negative log-likelihood loss is $\mathcal{L}_g = -\sum_{t=1}^{T} \log p(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{X}^u)$. Overall, the model objective combines different losses including the classification loss $\mathcal{L}_c = \mathcal{L}_e + \mathcal{L}_r + \mathcal{L}_d$ and the text generation losses into a unified function $\mathcal{L} = \alpha \cdot \mathcal{L}_c + \beta \cdot \mathcal{L}_f + \gamma \cdot \mathcal{L}_g + \delta \cdot \mathcal{L}_t$ where $\alpha = 1$, $\beta = 1$, $\gamma = 1.5$ and $\delta = 1$ were empirically chosen.

III. Experiments

A. Experimental Settings

The EmpatheticDialogues dataset [17] was collected to train a dialogue system to understand and respond empathetically to human emotions in conversations. This dataset was designed to evaluate system ability to engage in emotion-aware dialogue, improving the machine capacity to interpret and react to various emotional cues during interactions. Several related models were included for comparative study. KEMP [25] employed a graph encoder to handle the emotional dependencies in conversations. CEM [26] integrated the situational and emotional data to boost the dialogue empathy. SEEK [27] leveraged the interplay between knowledge and emotion for predicting the responses. CASE [28] aligned the cognitive and emotional aspects for nuanced dialogue. ECORE [29] employed a multi-resolution emotion graph and a correlation-enhanced decoder.

TABLE I: Examples of generated responses in two-speaker dialogue by using different methods under different emotions.

Context	*Speaker*: I just moved to this neighborhood and some dumb criminals shot one of my neighbors and ran into the woods!
	Listener: That's not good. Do you own a gun?
	Speaker: I do! I want to be able to protect my son.
Speaker Emotion	Angry
CEM [26]	I am sure he was a good parent.
KEMP [25]	I am glad you are okay.
SEEK [27]	That's a lot of people in the world.
CASE [28]	That's good, I hope he gets old and get him back.
ECORE [29]	I am glad you are ok.
KITM	I agree with you.
Golden	That is always the number one goal.
Context	*Speaker*: Yeah about 10 years ago I had a horrifying experience. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost ran me off the road.
	Listener: Did you suffer any injuries?
	Speaker: No, I was not hit. It turned out they were drunk. I felt guilty but realized it was his fault.
Speaker Emotion	Guilty
CEM	I haven't seen that too.
KEMP	I am glad you are okay.
SEEK	I am glad you did not get it?
CASE	Oh no, that is terrible.
ECORE	That is terrible. I am glad you got it back.
KITM	Oh my gosh, I am so sorry to hear that.
Golden	Why did you feel guilty? People really shouldn't drive drunk.

Evaluation metrics include both automatic and human assessments using our KITM and other methods. Automatic metrics are accuracy (Acc.) (%) for emotion-intent tagging, perplexity [30] (PPL) for text fluency and coherence, distinct-n (Dist-n) [31] for response diversity, and BLEU [32] for comparing the generated responses to reference texts. Human evaluations was to rate the coherence (Coh.), empathy (Emp.), and fluency (Flu.) on a scale of 1 to 5, providing a qualitative measure of the quality of dialogue. Four professional annotators were hired to judge the quality of the generated response in human evaluation.

TABLE II: Results on different automatic and human metrics.

Method	PPL ↓	Dist-1 ↑	Dist-2 ↑	Acc. ↑	Coh.	Emp.	Flu.
Transfomer	37.68	0.45	2.02	-	-	-	-
CEM	37.03	0.66	2.99	36.44	2.78	2.65	2.85
KEMP	36.89	0.61	2.65	37.58	2.63	2.63	2.7
SEEK	37.09	0.73	3.23	41.85	2.08	2.08	2.1
CASE	36.4	0.65	3.37	37.81	2.48	2.55	2.75
ECORE	**33.31**	0.72	3.49	39.56	3.03	2.83	3.08
KITM	36.06	**0.88**	**3.58**	**42.04**	**3.25**	**3.18**	**3.48**

B. Experimental Results

Table I shows the dialogue responses using different methods to reply a speaker about a crime issue under different emotions. Each model provides a unique response where only KITM agrees with the speaker's fear, but still lacking emotional support. Table II reports the results on automatic and human evaluations. In automatic evaluation, KITM obtains with the highest scores in distinctiveness and accuracy among different methods, while ECORE has the lowest perplexity. In human evaluation over different related methods, KITM leads in Coh., Emp., and Flu., with ECORE showing good performance in Coh. and Emp. but weak performance in Flu.

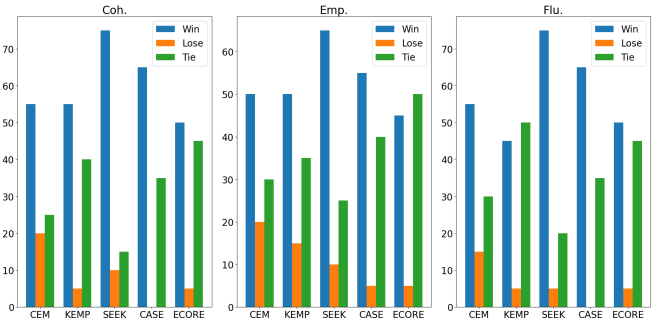


Fig. 3: Human A/B test (%) by using KITM relative to other methods on the metrics of coherence, empathy and fluency.

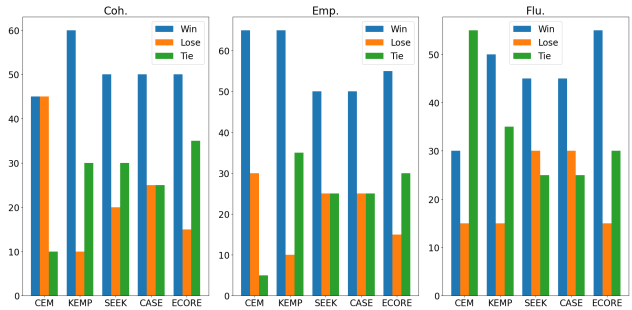


Fig. 4: LLM A/B test (%) by using KITM relative to other methods on the metrics of coherence, empathy and fluency.

TABLE III: Ablation study on knowledge-infused topic model.

Method	PPL ↓	Dist-1 ↑	Dist-2 ↑	Acc. (%) ↑
KITM	**36.06**	0.88	3.58	42.04
w/o Topic	36.77	0.61	2.44	**43.58**
w/o Topic Loss	36.44	**0.89**	**3.6**	42.89
w/o Graph	36.56	0.61	2.42	41.48
w/o Graph & Topic	36.86	0.55	2.16	39.94

Figure 3 highlights the KITM's consistent superiority in the A/B tests, particularly in Coh. and Flu., often outperforming the other methods like CEM, SEEK, and CASE. In this comparison, KITM and ECORE are found to be closely matched, with several ties showcasing their competitive relation. Overall, KITM emerges as a robust method across both automatic and human evaluation metrics, excelling in generating coherent, empathetic, and fluent outputs. Figure 4 shows the results of an A/B test comparing the performance of KITM with several baseline large language models (LLMs). The evaluation focuses on three metrics including coherence, empathy, and fluency. For this comparison, we utilize four distinct LLMs

including GPT-3.5 Turbo, LLaMA 3 (8B), Gemini, and Claude 3.5. These models serve as the benchmarks for assessing the performance of KITM across different metrics. The averaged scores from all four models provide a comprehensive baseline for comparison across the three evaluation aspects.

Table III presents an ablation study on KITM, showing its performance across various metrics. It is found that KITM achieves the competitive PPL, Dist-1 score, Dist-2 score, and accuracy. The accuracy increases when topics are excluded from the model, indicating a trade-off between including topics and maintaining accuracy. This finding suggests that while topics enhance the diversity in dialogue generation, but may also introduce complexity that slightly reduces overall predictive accuracy. In addition, the importance of merging two-speaker graph encoder is obvious from this comparison.

IV. Conclusions

In conclusion, this paper has presented a learning approach to empathetic dialogue generation, producing the contextually appropriate and the emotionally reasonable responses in a dialogue system. The knowledge-infused topic model was developed to enhance the empathetic dialogue by leveraging a multi-role graph neural network to understand speaker roles and relationships, integrating a conversational neural topic model for maintaining the topic consistency and synthesizing the external knowledge. The emotion prediction of individual dialogue turns and overall dialogue was implemented. The effectiveness of this approach was demonstrated through comparative study in both automatic and human evaluations on an empathetic dialogue dataset.

References

[1] M. Rohmatillah and J.-T. Chien, "Advances and challenges in multi-domain task-oriented dialogue policy optimization," *APSIPA Transactions on Signal and Information Processing*, vol. 12, no. 1, 2023.

[2] R. Higashinaka, K. Imamura, et al., "Towards an open-domain conversational system fully based on natural language processing," in *Proc. of International Conference on Computational Linguistics*, 2014.

[3] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2015.

[4] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proc. of AAAI Conference on Artificial Intelligence*, 2016, pp. 3776–3783.

[5] P.-C. Chen, M. Rohmatillah, Y.-T. Lin, and J.-T. Chien, "Convcounsel: A conversational dataset for student counseling," in *Proc. of Conference of the Oriental COCOSDA*, 2024, pp. 1–6.

[6] M. Rohmatillah, B. G. Ngo, W. Sulaiman, P.-C. Chen, and J.-T. Chien, "Reliable dialogue system for facilitating student-counselor communication," in *Proc. of Annual Conference of International Speech Communication Association*, 2024, pp. 1003–1004.

[7] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2019.

[8] X. Zhang and Y. Li, "A cross-modality context fusion and semantic refinement network for emotion recognition in conversation," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2023.

[9] A. S. Raamkumar and Y. Yang, "Empathetic conversational systems: A review of current advances, gaps, and opportunities," *IEEE Transactions on Affective Computing*, 2022.

[10] M. Jung, Y. Lim, S. Kim, J. Y. Jang, S. Shin, and K.-H. Lee, "An emotion-based Korean multimodal empathetic dialogue system," in *Proc. of Workshop on When Creative AI Meets Conversational AI*, 2022.

[11] J.-T. Chien, "Topic modeling for speech and language processing," in *Modern Methodology and Applications in Spatial-Temporal Modeling*, pp. 87–111. Springer, 2016.

[12] J.-T. Chien and C.-H. Lee, "Deep unfolding for topic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 318–331, 2018.

[13] J.-T. Chien, "Hierarchical theme and topic modeling," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 565–578, 2016.

[14] H. Sun, Q. Tu, J. Li, and R. Yan, "ConvNTM: Conversational neural topic model," in *Proc. of the AAAI Conference on Artificial Intelligence*, 2023, pp. 13609–13617.

[15] J.-T. Chien, "Bayesian nonparametric learning for hierarchical and sparse topics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 422–435, 2018.

[16] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4762–4779.

[17] H. Rashkin, E. M. Smith, M. Li, and Y-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2019.

[18] J.-T. Chien and Y.-C. Wu, "Empathetic response generation via regularized Q-learning," in *Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2024, pp. 1–6.

[19] T.-C. Luo and J.-T. Chien, "Variational dialogue generation with normalizing flows," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7778–7782.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of International Conference on Neural Information Processing Systems*, 2017.

[21] T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. of International Conference on Learning Representations*, 2016.

[22] J.-T. Chien and C.-W. Tsao, "Graph evolving and embedding in transformer," in *Proc. of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, pp. 538–545.

[23] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, pp. 993–1022, 2003.

[24] S. Jiang, P. Ren, C. Monz, and M. de Rijke, "Improving neural response diversity with frequency-aware cross-entropy loss," in *The World Wide Web Conference*, 2019, pp. 2879–2885.

[25] Q. Li, P. Li, Z. Ren, P. Ren, and Z. Chen, "Knowledge bridging for empathetic dialogue generation," in *Proc. of AAAI Conference on Artificial Intelligence*, 2022, pp. 10993–11001.

[26] S. Sabour, C. Zheng, and M. Huang, "CEM: Commonsense-aware empathetic response generation," in *Proc. of AAAI Conference on Artificial Intelligence*, 2022, pp. 11229–11237.

[27] 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, pp. 4634–4645.

[28] J. Zhou, C. Zheng, B. Wang, Z. Zhang, and M. Huang, "CASE: Aligning coarse-to-fine cognition and affection for empathetic response generation," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2023.

[29] F. Fu, L. Zhang, Q. Wang, and Z. Mao, "E-CORE: Emotion correlation enhanced empathetic dialogue generation," in *Proc. of Conference on Empirical Methods in Natural Language Processing*, 2023.

[30] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker, "Perplexity—a measure of the difficulty of speech recognition tasks," *Journal of the Acoustical Society of America*, vol. 62, no. S1, pp. S63–S63, 1977.

[31] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.