Medical Dialogue Generation via Dual Flow Modeling

Kaishuai Xu[1], **Wenjun Hou**[1,2*], **Yi Cheng**[1*], **Jian Wang**[1], **Wenjie Li**[1]

[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[2]Research Institute of Trustworthy Autonomous Systems and
Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen, China
{kaishuaii.xu, alyssa.cheng, jian-dylan.wang}@connect.polyu.hk,
houwenjun060@gmail.com, cswjli@comp.polyu.edu.hk

Abstract

Medical dialogue systems (MDS) aim to provide patients with medical services, such as diagnosis and prescription. Since most patients cannot precisely describe their symptoms, dialogue understanding is challenging for MDS. Previous studies mainly addressed this by extracting the mentioned medical entities as critical dialogue history information. In this work, we argue that it is also essential to capture the transitions of the medical entities and the doctor's dialogue acts in each turn, as they help the understanding of how the dialogue flows and enhance the prediction of the entities and dialogue acts to be adopted in the following turn. Correspondingly, we propose a Dual Flow enhanced Medical (DFMED) dialogue generation framework. It extracts the medical entities and dialogue acts used in the dialogue history and models their transitions with an entity-centric graph flow and a sequential act flow, respectively. We employ two sequential models to encode them and devise an interweaving component to enhance their interactions. Experiments on two datasets demonstrate that our method exceeds baselines in both automatic and manual evaluations.

1 Introduction

Medical dialogue systems (MDS) have drawn considerable research attention with the increasing demand for telemedicine, especially after the outbreak of the COVID-19 pandemic (Zeng et al., 2020; Liu et al., 2020; Zhou et al., 2021; He et al., 2022; Xia et al., 2020; Yan et al., 2022), as they can provide much more people with in-time and affordable access to medical services such as health consultation, diagnosis, and prescription.
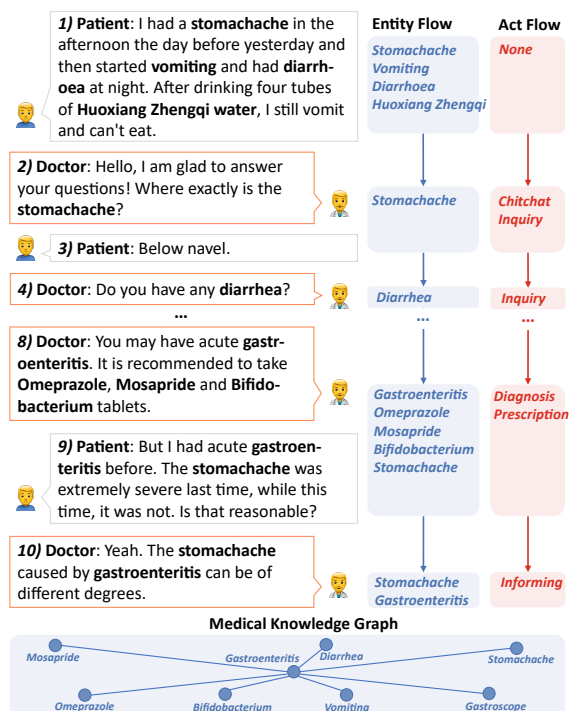

Figure 1: An example of a medical dialogue. *Diagnosis* and *Prescription* are short for *Make a diagnosis* and *Prescribe medications*, respectively.

For MDS, an efficient understanding of the dialogue history is challenging, as patients usually cannot describe their symptoms precisely and tend to convey lots of redundant information unnecessary for diagnosis (Liu et al., 2020; Mengel et al., 2002). To extract the critical information in the lengthy dialogue, previous research focused on identifying the important medical entities mentioned in the context, such as diseases, medicine, and symptoms (Liu et al., 2020; Lin et al., 2021; Liu et al., 2021).

In our work, we argue that capturing the transitions of the medical entities and the doctor's dialogue acts in each turn (as depicted in Figure

1) is also essential for the construction of MDS, which was largely overlooked by previous studies. In medical dialogues, the flows of medical entities and dialogue acts both follow particular transition patterns. For the medical entity flow, entities to appear in the following utterance are usually closely related to the ones mentioned recently. As in Figure 1, the entities mentioned in adjacent dialogue rounds are logically related, being neighboring nodes in the medical knowledge. For the dialogue act flow, though variations are allowed to some extent, it usually needs to follow the medical consultation framework suggested in Silverman et al. (2016). Modeling these two types of transitions would be helpful in dialogue understanding as they effectively capture how the dialogue history flows. Moreover, learning their transition patterns would also enhance the prediction of the dialogue acts and the medical entities to be adopted in the future turn.

Based on the above intuition, we propose a **D**ual **F**low enhanced **Med**ical (DFMED) dialogue generation framework. At each dialogue turn, it extracts the medical entities and the dialogue acts used in the dialogue history, and models their transitions with an entity-centric graph flow and a sequential act flow, respectively. Two sequential models are constructed to encode their transitions, with an interweaving component to enhance their interactions. The output representations are then used to predict the entities and the acts to be adopted in the following turn, which are employed to guide the response generation through gate control. Our main contributions are summarized as follows:

- We propose a novel MDS framework, DFMED, which models the transitions of medical entities and dialogue acts via step-by-step interweaving.
- We summarize the dialogue acts in the medical consultation scenario grounded on the medical documentation standards, SOAP notes (Cameron and Turtle-Song, 2002), including *make a diagnosis*, *prescribe medications*, etc.
- Experimental results show the superiority of DFMED over the previous frameworks and demonstrate the effectiveness of introducing the medical entity and dialogue act flows.

2   Related Work

Medical dialogue systems aim to provide medical services for patients. Early studies focus on automatic diagnosis in the form of a task-oriented dialogue system, and the purpose is to collect hidden symptoms in minimal turns and make a diagnosis at the end (Liao et al., 2020; Lin et al., 2019; Chen et al., 2021; Liu et al., 2022). Wei et al. (2018) creates a dataset annotated with symptom phrases and constructs a reinforcement learning based MDS. Xu et al. (2019) improves the topic transition in MDS by introducing a medical knowledge graph. With the release of large-scale medical dialogue datasets (e.g., MedDialog (Zeng et al., 2020), MedDG (Liu et al., 2020), and KaMed (Li et al., 2021)), dialogue response generation attracts increasing attention. Liu et al. (2020) frames medical dialogue generation as entity prediction and entity-aware response generation. Furthermore, Liu et al. (2021) unifies the dialogue context understanding and entity reasoning through a heterogeneous graph. Li et al. (2021) considers medical entities in patient and doctor utterances as states and actions and presents semi-supervised variation reasoning with a patient state tracker and a physician action network. The proposed model, VRBot, achieves comparable performance without entity supervision. Lin et al. (2021) analyses a low-resource challenge in medical dialogue generation and develops an entity-involved meta-learning framework to enhance diagnostic experience sharing between different diseases.

Although many studies have tried to improve medical dialogue generation by incorporating predicted medical entities, they simply treat entities in different turns as nodes in one entity graph with no entity transition modeling. Besides, few works focus on sequential entity-guided dialogue act prediction and sequential act-involved entity selection. Our framework exploits the transition and interaction of entity and act flows to strengthen dialogue understanding and guide response generation.

3   Preliminary

Problem Formulation.   We define a medical dialogue as $U=\{(P_k, D_k)\}_{k=1}^{T}$, where P and D represent utterances from patients and doctors. Each utterance contains several medical entities $E=\{e_i\}$, and each doctor utterance is annotated with multiple dialogue acts $A=\{a_j\}$. Given the dialogue history $U_t=\{P_1, D_1, ..., P_t\}$, the system is supposed to generate the t-th doctor utterance D_t.

Dialogue Acts.   We summarize several common dialogue acts implied in medical dialogues. One type is medical-related dialogue acts. We design

acts that represent a function in a medical documentation standard, the SOAP note (Cameron and Turtle-Song, 2002), and occur throughout the dialogue. For example, *State a required medical test* and *Prescribe medications* in the "Plan" function of the SOAP note are included in our designed acts. The other type is general open-domain dialogue acts. We choose some acts introduced by Zhao et al. (2022) and further refine them, such as merging acts that behave as social obligation management into *Chitchat*. Eventually, we obtain 7 dialogue acts for flow modeling, and later experiments demonstrate the effectiveness of guidance for response generation. Detailed description of each dialogue act can be found in Appendix A.2.

4 Method

The proposed method learns two flows (i.e., a medical entity flow and a dialogue act flow) to model the propagation of medical entities and medical-related interactions and guide response generation with the corresponding hints. As shown in Figure 2, the overall framework contains two modules. The Dual Flow Modeling module learns the entity and act transitions and infers probable entities and acts. The Response Generation module outputs a response under the guidance of the selected entities and predicted acts.

4.1 Dual Flow Modeling

Figure 3 left displays the architecture of the Dual Flow Modeling module. To extract flow transitions, we encode the medical entities and the dialogue acts in a sequential way. Besides, at each turn, entity and act embeddings are mutually integrated through an interweaving component, named Interweaver. The context states S^c produced by a context encoder also play a role in the integration. After encoding the whole dialogue history, we obtain the entity state S_t^e and act state S_t^a. The entity states are adopted to select the relevant entities. The act states are sent to a multi-label act predictor to estimate the next acts.

We apply BERT (Devlin et al., 2019) as the context encoder to encode the dialogue history. All utterances are concatenated with a unique role token (i.e., "[P]" or "[D]") to separate the patient and doctor utterances. We compute the context states with different history ranges $U_k=\{P_1, D_1, ..., P_k\}(k \in [1, t])$ for supporting the interweaving at each turn. The mean embedding of tokens in history U_k is ap-

Figure 2: The overall framework of DFMED.

plied as the k-th turn context state $S_k^c \in \mathbb{R}^d$, where d is the dimension of the state.

4.1.1 Entity Flow

To learn the sequential transition of medical entities, we create the Entity Flow that encodes an entity graph transition and selects the most relevant entities. Specifically, the flow consists of an entity graph for each turn. Medical entities of each dialogue turn and entities from an external knowledge graph with a one-hop connection to the entities in dialogue are included in the graph of the k-th turn, denoted as G_k. Since we assume entities hop along the links on the graph, the ones in the sub-graph partially provide future transition hints.

We define graphs that include entities until the t-th turn as $G_{\leq t}$. For all entities in graph $G_{\leq t}$, we use the same encoder as above to get entity embeddings. Instead of randomly initializing an embedding, BERT-based encoding keeps token-level semantics. Then, the average token embedding of each entity is employed as the raw embedding, denoted as $h^{e_0} \in \mathbb{R}^d$. Finally, Graph Attention Network (GAT) (Velickovic et al., 2018) is implemented to merge neighboring information for each entity:

$$\alpha_{ij}^k = \frac{\exp\left(\sigma_1\left(a^\mathsf{T}[W^k h_i^{e_0} \| W^k h_j^{e_0}]\right)\right)}{\sum_{\mu \in \mathcal{N}_i} \exp\left(\sigma_1\left(a^\mathsf{T}[W^k h_i^{e_0} \| W^k h_\mu^{e_0}]\right)\right)}, \tag{1}$$

$$h_i^e = \left[\sigma_2\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^k W^k h_j^{e_0}\right)\right]_{k=1}^K, \tag{2}$$

where $h_i^e \in \mathbb{R}^d$ is the updated entity embedding, K is the number of heads, $a \in \mathbb{R}^{2d}$ is a trainable weight, $W^k \in \mathbb{R}^{d_h \times d}$ is the linear transformation matrix for the k-th head, σ_1 and σ_2 are the activation function, and $\mathcal{N}_i$ represents neighboring entities that connect to entity i with one hop. The updated embedding is used to compute each turn's overall graph embedding via mean pooling:

$$\bar{h}_k^e = \frac{1}{|G_k|} \sum_{i \in G_k} h_i^e, k \in [1, t], \tag{3}$$
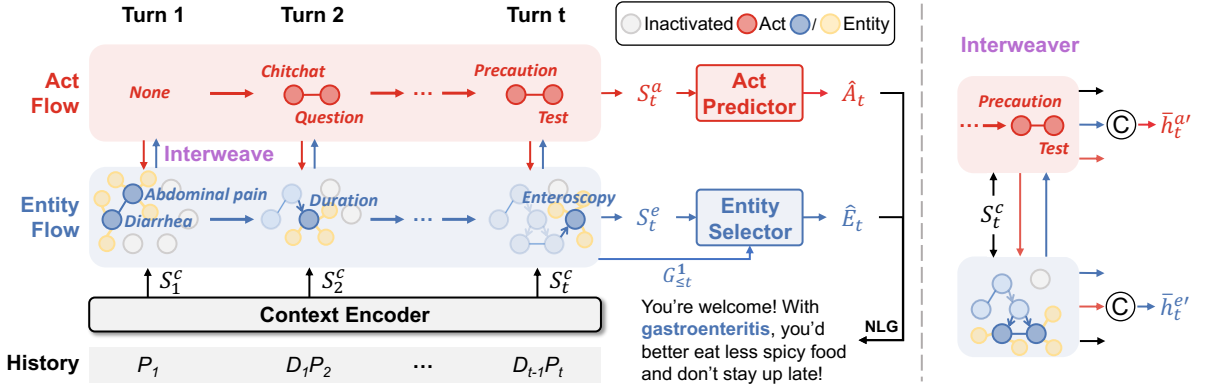
Figure 3: **Left**: The architecture of Dual Flow Modeling. It includes the dialogue act flow modeling in red and the medical entity flow modeling in blue. The one-hop sub-graph nodes of entities in dialogue are marked in yellow. **Right**: Detailed structure for flow interweaving. Each turn's graph embedding attends to previous act embeddings, and each turn's sequence embedding attends to previous entity embeddings.

where $\bar{h}_k^e \in \mathbb{R}^d$ is the entity graph embedding, t represents the turn of the target response. Following Tu et al. (2022), we employ a GRU to model the entity graph transition throughout the dialogue. The GRU takes all the previous graph embeddings $\{\bar{h}_1^e, \bar{h}_2^e, ..., \bar{h}_t^e\}$ as input and produces the entity state as follows:

$$S_t^e = \text{GRU}_{\text{Entity}}(S_{t-1}^e, \bar{h}_t^e), \qquad (4)$$

where $S_t^e \in \mathbb{R}^d$ is the final hidden state of the GRU. We denote S_t^e as the entity state that implies clues for the entity transition in the previous context.

Then, we apply the entity state S_t^e to calculate relevant scores for candidate entities. These entities are from the sub-graph of entities in a historical dialogue context. The score is defined as follows:

$$\text{Score} = \langle S_t^e, h_i^e \rangle, i \in G_{\leq t}^{\mathbf{1}}, \qquad (5)$$

where $\langle, \rangle$ represents a similarity function, h_i^e is the entity embedding, $G_{\leq t}^{\mathbf{1}}$ is the one-hop sub-graphs for entities until the t-th turn. We select top-k relevant entities $\hat{E}_t$ to guide response generation.

4.1.2 Act Flow

To learn the medical-related interactions of the doctor, we design the Act Flow that encodes act sequences and predicts the next acts. Specifically, the flow is composed of the act sequence of each turn. We first randomly initialize the trainable embedding of each dialogue act, denoted as $h^a \in \mathbb{R}^d$. The act sequence of the k-th turn can be defined as $A_k = \{h_1^a, h_2^a, ..., h_{n_a^k}^a\}$, where n_a^k represents the number of dialogue acts for each turn. Then, we compute the act sequence embedding $\bar{h}_k^a \in \mathbb{R}^d, k \in [1, t]$ through mean pooling.

Similar to the entity flow, we employ a GRU to model dialogue act transition. With all sequence embeddings as input, the final hidden state of the GRU is calculated and denoted as the act state:

$$S_t^a = \text{GRU}_{\text{Act}}(S_{t-1}^a, \bar{h}_t^a), \qquad (6)$$

where $S_t^a \in \mathbb{R}^d$ is the act state of the t-th turn. Then, the multi-act probability of the t-th turn is computed based on the act state S_t^a with a sigmoid and linear transformation layer:

$$\text{Prob} = \text{sigmoid}(W_a S_t^a + b_a), \qquad (7)$$

where $W_a \in \mathbb{R}^{n_a \times d}$ and $b_a \in \mathbb{R}^{n_a}$ are model parameters, and n_a denotes the number of candidate dialogue acts. The predicted dialogue acts $\hat{A}_t$ are obtained through an appropriate threshold.

4.1.3 Flow Interweaving

To achieve the integration of these two flows, we present an Interweaver to extend the entity graph embedding and act sequence embedding, as shown in Figure 3 right. This component incorporates the dialogue context into entity/act states and integrates entity/act sequential information from each other.

For the entity flow, we first fuse the historical context into the entity graph embedding. The context-aware graph embedding of each turn is computed via cross-attention:

$$\alpha_{ki} = \text{softmax}(\frac{Q_k^{c\top} K_i^e}{\sqrt{d}}), \qquad (8)$$

$$\bar{h}_k^{e^c} = \sum_{i \in G_k} \alpha_{ki} V_i^e, k \in [1, t], \qquad (9)$$

where $\bar{h}_k^{e^c} \in \mathbb{R}^d$ is the context-aware graph embedding at the k-th turn, K_i^e and V_i^e are linear projected vectors based on the entity embedding h_i^e, and Q_k^c is based on the context state S_k^c. We denote this operation on S_k^c and $[h_i^e]_{i \in G_k}$ as $\text{CA}(S_k^c, [h_i^e]_{i \in G_k})$. Then, the act transition is fused into the graph embedding via CA as follows:

$$\bar{h}_k^{e^a} = \text{CA}(\bar{h}_k^e, [h_j^a]_{j \in A_{\leq k}}), k \in [1, t], \quad (10)$$

where $\bar{h}_k^{e^a} \in \mathbb{R}^d$ is the act-aware graph embedding, and $A_{\leq k}$ represents act sequences until the k-th turn. The final entity graph embedding is defined as the concatenation of three embeddings:

$$\bar{h}^{e'} = [\bar{h}^e; \bar{h}^{e^c}; \bar{h}^{e^a}], \quad (11)$$

where $\bar{h}_k^{e'} \in \mathbb{R}^{3d}$ is the extended graph embedding that incorporates historical dialogue context and act transition pattern.

For the act flow, we apply the CA operation to compute the context-aware and entity-aware act sequence embedding following the same way. The context-aware sequence embedding $\bar{h}_k^{a^c}$ is based on S_k^c and $[h_j^a]_{j \in A_k}$, and the entity-aware sequence embedding $\bar{h}_k^{a^e}$ is based on $\bar{h}_k^a$ and $[h_i^e]_{i \in G_{\leq k}}$. These two embeddings incorporate the historical dialogue context and entity transition pattern. Then, the final sequence embedding $\bar{h}^{a'} \in \mathbb{R}^{3d}$ is concatenated as follows:

$$\bar{h}^{a'} = [\bar{h}^a; \bar{h}^{a^c}; \bar{h}^{a^e}]. \quad (12)$$

We send the extended embeddings $\bar{h}^{e'}$ and $\bar{h}^{a'}$ instead of the pure embeddings to the above GRUs, allowing two flow models to acquire context information and lead each other.

4.1.4 Training Objective

The training of the dual flow modeling module includes two tasks, medical entity ranking and multi-act classification. The first one follows a contrastive learning (Gao et al., 2021) way with a negative log likelihood loss:

$$\mathcal{L}_e = -\sum_t^T \sum_{t^+} \log \frac{e^{\langle S_t^e, h_{t^+}^e \rangle}}{\sum_{i^- \in G_{\leq t}^{\mathbf{1}}} e^{\langle S_t^e, h_{i^-}^e \rangle}}, \quad (13)$$

where $h_{t^+}^e$ is the embedding of a target entity mentioned in the t-th response, $\langle , \rangle$ is the dot product operation that calculates relevant scores (see Eq. 5), and $G_{\leq t}^{\mathbf{1}}$ denotes the one-hop sub-graphs for
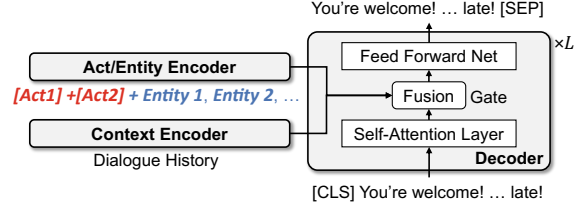

Figure 4: The detail of Response Generation.

the dialogue history until the t-th turn. We randomly select several entities instead of the whole sub-graph as negative entities. The second one is defined as a multi-label classification with a binary cross-entropy loss:

$$\mathcal{L}_a = \sum_t^T \sum_j^{n_a} \text{BCE}(\hat{y}_{t,j}^a, y_{t,j}^a), \quad (14)$$

where $\hat{y}_{t,j}^a$ is the probability of the dialogue act j for the t-th response (see Eq. 7), and $y_{t,j}^a$ is the ground-truth act label. The overall training objective of the dual flow modeling module can be calculated as:

$$\mathcal{L}_F = \lambda_e \mathcal{L}_e + \lambda_a \mathcal{L}_a, \quad (15)$$

where λ_e and λ_a are weights for each task.

4.2 Flow-guided Response Generation

After training the dual flow modeling module, we exploit the top-k relevant entities $\hat{E}^t$ (see Sec. 4.1.1) and predicted dialogue acts $\hat{A}^t$ (see Sec. 4.1.2) to guide response generation. We first encode the entity/act and dialogue history separately, allowing a relatively complete dialogue context. Then, these two types of information is merged into the decoder via a fusion component.

Act-Entity Fusion As shown in Figure 4, the dialogue acts and entities are concatenated into one token sequence. We assign a unique token to each dialogue act. Given the entity/act sequence and dialogue history sequence as input, the corresponding encoder produces final hidden states, H^{ea} and H^c. We fuse two types of information through a gate mechanism in the decoder:

$$h_w^{l,ea} = \text{CA}(h_w^l, H^{ea}), \quad (16)$$

$$h_w^{l,c} = \text{CA}(h_w^l, H^c), \quad (17)$$

$$g_w^l = \text{sigmoid}(W^l h_w^{l,c}), \quad (18)$$

$$h_w^{l'} = \text{FFN}(g_w^l h_w^{l,c} + (1 - g_w^l) h_w^{l,ea}), \quad (19)$$

where $h_w^{l'}$ is the final hidden state at the w-th token after fusion, h_w^l denotes the output of the l-th self-attention layer, FFN is the feed forward network, and W^l is the trainable parameter. We adopt the final hidden state $h_w^{L'}$ of the decoder to compute the probability distribution of the next token:

$$p(D_{t,w+1}) = \text{softmax}(W_d h_w^{L'} + b_d), \quad (20)$$

where W_d and b_d are linear mapping parameters.

Training and Inference When training the response generation module, we use the selected top-k entities from the dual flow modeling module and ground truth dialogue acts as encoder input. The training objective is defined as follows:

$$\mathcal{L}_G = -\sum_t^T \sum_{w=1}^w \log p(D_{t,w+1}), \quad (21)$$

where $p(D_{t,w+1})$ denotes the probability of the next token in the t-th response. Then in inference, we apply the top-k entities and predicted dialogue acts to generate the next response.

5 Experiments

5.1 Dataset

Our experiments are conducted on two medical dialogue datasets, MedDG (Liu et al., 2020) and KaMed (Li et al., 2021). The MedDG dataset contains 17K dialogues, focusing on 12 diseases in the gastroenterology department. The medical entities mentioned in the dialogues are annotated in MedDG. We split the dataset into 14862/1999/999 dialogues as the training/validation/test sets, following its original division. The KaMed dataset contains over 63K dialogues, covering diverse diseases in about 100 hospital departments. We filter out some dialogues with privacy concerns in KaMed (see Appendix A.3) and obtain 29,159/1,532/1,539 dialogues as the training/validation/test sets. The final data accounts for around 51% of the total.

5.2 Baseline models

We compare DFMED with five baseline models. (1) **Seq2Seq** (Sutskever et al., 2014) is an RNN-based sequence to sequence model with an attention mechanism. (2) **HRED** (Serban et al., 2016) is a hierarchical RNN that models a dialogue as a token sequence and an utterance sequence. (3) **GPT-2** (Radford et al., 2019) is a transformer decoder-based language model. (4) **BART**

(Lewis et al., 2020) is a transformer-based encoder-decoder model. (5) **VRBot** (Li et al., 2021) is a medical dialogue generation model with patient entity tracking and doctor entity learning.

For the experiments on the MedDG dataset, we extend Seq2Seq, HRED, GPT-2, and BART with entity hints, following Liu et al. (2020). Specifically, the extracted medical entities are appended at the end of the input sequence. In the following, these models with entity modeling are displayed with the **-Entity** suffix (e.g., Seq2seq-Entity).

5.3 Evaluation Metrics

Automatic Evaluation. We evaluate modules in the proposed framework separately. For the dual flow modeling module, the top-20 recall rate (R@20) of target entities and the weighted F1 score (Weighted-F1) of different dialogue acts considering the act imbalance are adopted. For the response generation module, BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores in different n-grams (i.e., B-1, B-2, B-4, R-1, and R-2) are adopted to assess the response quality. Besides, we also measure the precision, recall, and F1 of entities (i.e., E-P, E-R, and E-F1) in responses to demonstrate the reliability following Liu et al. (2020).

Human Evaluation. We chose 100 cases at random and invited three annotators to evaluate them manually. Results of the proposed framework are compared with different baseline models. We use three metrics to evaluate all of the generated responses based on past research (Liu et al., 2020; Li et al., 2021): *sentence fluency* (FLU), *knowledge accuracy* (KC), and *entire quality* (EQ). On a 5-point Likert scale, from 1 (worst) to 5 (best), three annotators are asked to score these responses.

5.4 Implementation Details

For all baselines, we implement the open-source algorithm following Liu et al. (2020) and Li et al. (2021). We use the MedBERT[1] pretrained in the medical domain as the backbone of the dual flow modeling module. To extract the medical entities in the dialogue history, we refer to the medical knowledge graph CMeKG[2] and extract the text spans that match the string of nodes in the graph. Entities with a one-hop connection to the historical entities are target ones for entity flow modeling. Besides, following Yan et al. (2022), to identify

---

[1] https://github.com/trueto/medbert
[2] http://cmekg.pcl.ac.cn/

Methods		B-1	B-2	B-4	R-1	R-2	E-P	E-R	E-F1
w/o Pre-training	Seq2Seq	28.55	22.85	15.45	25.61	11.24	16.79	10.44	12.88
	Seq2Seq-Entity	29.13	23.22	15.66	25.79	11.42	**23.79**	15.89	19.06
	HRED	31.61	25.22	17.05	24.17	9.79	15.56	10.12	12.26
	HRED-Entity	32.84	26.12	17.63	24.26	9.76	21.75	15.33	17.98
	VRBot	29.69	23.90	16.34	24.69	11.23	18.67	9.72	12.78
w/ Pre-trained LM	GPT-2	35.27	28.19	19.16	28.74	13.61	18.29	14.45	16.14
	GPT-2-Entity	34.56	27.56	18.71	28.78	13.62	21.27	17.10	18.96
	BART	34.94	27.99	19.06	29.03	**14.40**	19.97	14.29	16.66
	BART-Entity	34.14	27.19	18.42	28.52	13.67	23.49	16.90	19.66
	DFMED	**42.56**†	**33.34**†	**22.53**†	29.31	14.21	22.48	**22.84**†	**22.66**†
	w/o Act Flow	36.79	29.18	19.81	**29.45**	14.26	22.73	21.70	22.20
	w/o Entity Flow	42.14	32.83	21.95	29.26	13.73	16.86	22.06	19.11
	w/o Interweaving	42.35	33.02	22.19	29.02	14.11	22.14	20.62	21.34

Table 1: Automatic evaluation results on MedDG. $\dagger$ denotes statistically significant differences ($p < 0.05$).

Methods	B-1	B-2	B-4	R-1	R-2
Seq2Seq	23.52	18.56	12.13	23.56	8.67
HRED	26.75	21.08	13.91	22.93	7.80
VRBot	30.04	23.76	16.36	18.71	7.28
GPT-2	33.76	26.58	17.82	26.80	10.56
BART	33.62	26.43	17.64	27.91	11.43
DFMED	**40.20**†	**30.97**†	**20.76**†	**28.28**†	**11.54**†
w/o Act Flow	35.47	28.11	18.78	27.97	11.45
w/o Entity Flow	39.14	29.92	19.73	27.17	10.47
w/o Interweaving	39.34	30.45	20.38	28.03	11.39

Table 2: Automatic evaluation results on KaMed. $\dagger$ denotes statistically significant differences compared to all baselines ($p < 0.05$).

Methods	FLU	KC	EQ
BART	3.82	1.86	3.06
BART-Entity	3.85	2.03	3.35
DFMED	**4.00**	**2.14**	**3.61**
Gold	4.12	3.97	4.35

Table 3: Human evaluation results on MedDG.

epochs with a batch size of 4. We implement all experiments on a single RTX 3090 GPU.

6 Results and Analysis

6.1 Automatic Evaluation

The overall comparison of DFMED and other baseline models on the MedDG dataset is illustrated in Table 1, and the KaMed dataset is in Table 2. The observations from the comparison are as follows:

(1) Our proposed framework DFMED outperforms these baseline models in most metrics. Specifically, on the MedDG dataset, it is 8.42%, 6.15%, 4.11%, and 0.79% higher than the best baseline model, BART-Entity, on B-1, B-2, B-4, and R-1. On the KaMed dataset, DFMED outperforms BART by 6.58%, 4.54%, 3.12%, and 0.37% on B-1, B-2, B-4, and R-1, indicating better similarity to the content of ground truth responses. Besides, on the MedDG dataset, DFMED exceeds BART-Entity by 3.00% on E-F1, meaning that it can generate responses with more accurate entity mentions. The above increases are because DFMED learns the transition of a medical entity flow and a dialogue act flow and predicts the entities and acts to guide response generation. Moreover, interweaving two flows enhances the prediction of entities and acts.

(2) DFMED effectively fuses the guidance from

the dialogue acts, we apply an open-source pseudo-labeling algorithm[3] to automatically label each utterance. We train the dual flow modeling module with the AdamW (Loshchilov and Hutter, 2019) optimizer. The learning rate and batch size are 4e-5 and 12 with 1000 warmup steps. The best loss weights λ_e and λ_a are 1 and 0.05 through grid searching. After training ten epochs, checkpoints with the highest average F1 for act prediction and the highest recall rate for top-20 entity selection on the validation set are selected. We select the threshold for each dialogue act, which achieves the best F1 score for the corresponding act on all validation samples. Then, we use Chinese pre-trained BART$_{base}$[4] model with a six-layer encoder and a six-layer decoder for the response generation module. The entity/act encoder and context encoder share the same encoder. We adopt the AdamW optimizer and set the learning rate to 3e-5 with 2000 warmup steps. The model is trained for ten

---

[3]https://github.com/yanguojun123/Medical-Dialogue
[4]https://huggingface.co/fnlp/bart-base-chinese

Methods	MedDG			KaMed		
	Weighted-F1	R@20	B-4	Weighted-F1	R@20	B-4
DFMED	**62.83**	**56.53**	**22.53**	**55.81**	**52.23**	**20.76**
w/o Flow Modeling	62.09	54.74	22.11	55.16	51.31	20.21
w/o Interweaving	62.13	54.98	22.19	55.17	51.44	20.38
w/o Entity attends to Act	62.37	55.75	22.34	55.43	51.84	20.62
w/o Act attends to Entity	62.43	55.91	22.40	55.62	52.10	20.67

Table 4: Results of Dual Flow Modeling with ablation. The best results are in **bold**.

medical entities and dialogue acts. Compared to GPT-2 and BART, which have a performance drop on BLEU with the incorporation of medical entities (i.e., **-Entity**), DFMED shows increases in all these metrics. The main reason is that the force incorporation of entities may reduce response fluency. In DFMED, the dialogue act guidance influences attention to entities in the encoder and implicitly prevents the enforcement. The gate mechanism in the decoder controls the proportion of information from entities and acts. Besides, dialogue acts also provide essential content for responses without entity hints. This improvement is significant as about half of the responses in datasets do not match an entity in CMeKG via automatic string matching.

6.2 Human Evaluation

We select methods with high accuracy on the MedDG dataset to conduct a human evaluation, as shown in Table 3. Our framework displays an overall better response quality. Especially on the EQ, DFMED performs significantly better than baselines due to the incorporation of dialogue acts. The Cohen's Kappa coefficient equals 0.53 and indicates a moderate agreement (Cohen, 1968).

6.3 Analysis of Dual Flow Modeling

To further explore the effectiveness of our method, we investigate the following variants of DFMED: (1) **w/o Flow Modeling**, where we use a context state produced by mean pooling of all hidden states of dialogue context tokens to rank entities and predict dialogue acts. (2) **w/o Interweaving**, which removes the interweaving between entities and acts. (3) **w/o Entity attends to Act**, which removes the interweaving from act sequences. (4) **w/o Act attends to Entity**, which removes the interweaving from sequential entity graphs.

Table 4 shows the ablation study results. We observe drops in all metrics with the ablation variants, indicating the effectiveness of our proposed module. Specifically, the result of *w/o Flow Mod-*

eling significantly drops on the entity recall rate and slightly drops on the overall act prediction F1 compared to the full model. It demonstrates that flow modeling can be necessary for learning the transition of entities and acts in dialogues. Besides, comparison among variants of the interweaver illustrates that incorporating sequential entity graphs and act sequences assists in the transition. Notably, the act sequence is more conducive to the transition of the entity and act flows. It may be because the entity graph introduce noise entities from patient utterances. These entities are more variable and can be deviated from the main content.

6.4 Case Study

Patient: Hello, doctor! I feel a little stomachache and suffocated. The stomachache is intermittent, sometimes it hurts, sometimes it doesn't. When I am not active, the pain is not obvious, but when I move, bloating makes me sick. I still have diarrhea. :(
Doctor: Hi, is it colic? Does stomachache have anything to do with eating?
Patient: No! It hurts before and after eating.
...
Patient: I've never had a gastroscopy.

Gold Response: You may have gastritis or stomach cramps and need to do a gastroscopy or a barium meal first! If diagnosed with the above diseases, you can take omeprazole, Kangfuxin liquid, and belladonna tablets following specific instructions.

BART: I suggest you try some omeprazole and hydrotalcite chewable tablets.
BART-Entity: It is recommended to have a gastroscopy to see if there is any history of gastritis or gastric ulcer.

Selected Entities: gastritis, gastroscopy, omeprazole,...;
Predicted Acts: [diagnosis], [prescription], [test].
Ours (DFMED): [You may have gastritis or gastroesophageal reflux.] [It is recommended to take omeprazole, mosapride, hydrotalcite.] [If the symptoms are not relieved, it is recommended to do a gastroscopy.]

Table 5: Case study. The responses are generated by different models, where key entities are underlined.

A case generated by the above methods is illustrated in Table 5. Compared to baseline models, DFMED can produce responses more consistent

with the dialogue context reflected by medical entities and dialogue acts. We can observe that the dual flow modeling module correctly predicts all acts and selects several medical entities, although two medications (e.g., "mosapride") are different from ground truth mentions. Then, the beneficial guidance from medical entities and acts is employed to generate the next response. The response generation module fuses the entities and acts and produces the response containing these two hints.

7 Conclusion

In this paper, we propose a dual flow enhanced medical dialogue generation framework, DFMED, that models a medical entity flow and a dialogue act flow to improve the relevant entity selection and dialogue act prediction. Besides, we design an interweaver to strengthen interactions of two flows. The selected entities and predicted acts are applied to guide response generation. Experiments validate the effectiveness of our DFMED on two datasets.

Limitations

Although our proposed framework beats several baseline methods for medical dialogue generation, there is still room for progress. We exploit an entity flow and a dialogue act flow to improve dialogue understanding and guide response generation. However, our summarized dialogue acts are limited in the types and granularity of functions they denote. We can manually annotate more medical-related dialogue acts in our future research following the SOAP notes. Besides, more medical knowledge with different formats, such as medical articles and medical examination reports, can be incorporated. Finally, it is crucial to recognize the potential risks associated with system utilization and the possibility of patient privacy leakage. A collaborative approach involving both dialogue systems and medical professionals should be considered. This will ensure that responses are endorsed by physicians and stringently overseen by reliable authorities.

Ethics Statement

Our proposed system aims to provide medical services, such as diagnosis and prescription, for patients who suffer from certain diseases. All datasets have been anonymized when released in dataset papers. However, since we train the model with limited and incomplete samples in two datasets, the generated responses may involve misleading information about diagnosis, treatment, and precautions. We recommend that users adopt the system as an auxiliary tool and go to the hospital for help if necessary. Besides, when interacting with the system, there is a risk of sensitive information leak (e.g., gender as reported by users). It can be addressed by adopting anonymous technology in the future. Thus, we strongly advise users to consider the ethical implications of the generated responses carefully. Furthermore, the scientific artifacts that we used are freely available for research, including NLTK, ROUGE, Transformers, and other GitHub codes. And this paper's use of these artifacts is consistent with their intended use.

Acknowledgment

References

Susan Cameron and Imani Turtle-Song. 2002. Learning to write case notes using the soap format. *Journal of Counseling & Development*, 80(3):286–292.

Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. 2021. Diaformer: Automatic diagnosis via symptoms sequence generation. *ArXiv*, abs/2112.10433.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhenfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chen, Guandong Xu, and Jian Wu. 2022. Dialmed: A dataset for dialogue-based medication

recommendation. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 721–733. International Committee on Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. 2021. Semi-supervised variational reasoning for medical dialogue generation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 544–554. ACM.

Kangenbei Liao, Qianlong Liu, Zhongyu Wei, Baolin Peng, Qin Chen, Weijian Sun, and Xuanjing Huang. 2020. Task-oriented dialogue system for automatic disease diagnosis via hierarchical reinforcement learning. *ArXiv*, abs/2004.14254.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13362–13370. AAAI Press.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. 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)*, pages 5033–5042, Hong Kong, China. Association for Computational Linguistics.

Wenge Liu, Yi Cheng, Hao Wang, Jianheng Tang, Yafei Liu, Ruihui Zhao, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. "my nose is running.""are you also coughing?": Building A medical diagnosis agent with interpretable inquiry logics. *ArXiv*, abs/2204.13953.

Wenge Liu, Jianheng Tang, Xiaodan Liang, and Qingling Cai. 2021. Heterogeneous graph reasoning for

knowledge-grounded medical dialogue system. *Neurocomputing*, 442:260–268.

Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *ArXiv*, abs/2010.07497.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mark B Mengel, Warren Holleman, and Scott A Fields. 2002. *Fundamentals of clinical practice*. Springer Science & Business Media.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Jonathan Silverman, Suzanne Kurtz, and Juliet Draper. 2016. *Skills for communicating with patients*. crc press.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Quan Tu, Shen Gao, Yanran Li, Jianwei Cui, Bin Wang, and Rui Yan. 2022. Conversational recommendation via hierarchical information modeling. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2201–2205. ACM.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 1062–1069. AAAI Press.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7346–7353. AAAI Press.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2022. Remedi: Resources for multi-domain, multi-service, medical dialogues. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 3013–3024. ACM.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Jianqiao Zhao, Yanyang Li, Wanyu Du, Yangfeng Ji, Dong Yu, Michael R. Lyu, and Liwei Wang. 2022. Floweval: A consensus-based dialogue evaluation framework using segment act flows. *CoRR*, abs/2202.06633.

Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, Yichen Zhang, Qingyang Wu, Zhou Yu, Kun Xu, Eric Xing, and Pengtao Xie. 2021. On the generation of medical dialogs for COVID-19. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 886–896, Online. Association for Computational Linguistics.

A Appendix

A.1 Details of Packages

We use the NLTK package in version 3.4.1 for calculating BLEU scores, the Pyrouge package in version 0.1.3 for calculating ROUGE scores, and the Transformers package in version 4.21.3.

A.2 Details of Dialogue Acts

Dialogue Acts	MedDG	KaMed
Inquire	25.49%	20.95%
Make a diagnosis	6.72%	8.64%
Prescribe medications	10.12%	13.74%
State a required medical test	4.25%	8.29%
Provide daily precautions	7.51%	5.29%
Inform	29.91%	30.04%
Chitchat	15.98%	13.04%

Table 6: The proportion of each dialogue act in the MedDG and KaMed datasets.

In this section, we will describe the detail of our summarized dialogue acts. The seven dialogue acts can be divided into two types: (i) medical-related functions and (ii) general open-domain dialogue acts. The specific meaning of each act is interpreted as follows:

Medical-related functions. (1) **Inquire**. The doctor asks questions about the history of the present illness (e.g., the location and duration of one symptom), previous surgery and medical conditions, current medications, allergies, etc. It corresponds to the "Subjective" function of the SOAP note (Cameron and Turtle-Song, 2002). (2) **Make a diagnosis**. The doctor makes a differential diagnosis based on historical dialogue context. It corresponds to the "Assessment" function of the SOAP note. (3) **Prescribe medications**. The doctor provide medication names and instructions. (4) **State a required medical test**. The doctor explains which tests are required and why each one was chosen to resolve diagnostic ambiguities; besides, what the next step would be if the results are positive or negative. (5) **Provide daily precautions**. The doctor explains the things that need to be paid attention to every day. Acts (3), (4), and (5) correspond to the "Plan" function of the SOAP note.

General open-domain dialogue acts. (6) **Inform**. The doctor tells the patient some information that he assumes to be correct. (7) **Chitchat**. The doctor expresses welcome, goodbye, apology and thanks to the patient.

The proportion of each act in the MedDG and KaMed datasets is shown in the Table 6. Examples for each dialogue act are listed as follows:

1. **Inquire**: "Hello, do you usually have diarrhea?", "Have you taken any medicine before?";

2. **Make a diagnosis**: "You may have gastroenteritis", "You may have allergic rhinitis, which is easy to get sick this season.";

3. **Prescribe medications**: "Please take Amoxicillin capsule 1.0g 2 times a day and Clarithromycin tablet 0.5g 2 times a day. Both are taken after meals.";

4. **State a required medical test**: "If you often feel sick to your stomach, you can do a gastroscopy.", "If your condition does not improve, I suggest you do a gastroscopy and Helicobacter pylori detection.";

5. **Provide daily precautions**: "Please drink plenty of water, eat more fruits and vegetables. And try to have a morning poop.";

6. **Inform**: "Migraine is a primary headache whose etiology and pathogenesis are not fully understood.", "It can be effective in five days, and individual differences are relatively large.";

7. **Chitchat**: "You're welcome, and I wish you a speedy recovery!", "Thank you so much", "Hello!";

A.3 Extra Process of the KaMed dataset

Patient: Can I apply a facial mask if I have pimples on my face?
Doctor: Hello, do you have a picture to show? How long has it been? Are there any discomforts?
Patient: The image is not available for privacy concerns.
Doctor: Based on your description, it seems like acne, also known as pimples or blackheads.
...

Table 7: An example of the filtered dialogues.

We filter out dialogues that include texts such as "The image is not available for privacy concerns" and "The voice is not available for privacy concerns" since these dialogues are incomplete and hard to understand. We have obtained approximately 51% samples for our experiments. The reason is as follows.

The KaMed dataset is collected from an online medical consultation platform. The raw dialogues contain multi-modal information such as pictures (e.g., medical examination reports and photos of body parts) and voice messages. These messages are crucial for dialogue development since the doctor will respond to the picture or voice. For example, they will discuss the result of an examination report, or patients will directly use voice messages instead of texts to express their condition. However, this information is replaced by meaningless text such as "The image is not available" for privacy concerns when collecting the dataset. Thus, the dialogue context is incomplete and difficult to understand. We argue that filtering dialogues that contain these texts can help us build a more robust model for dialogue understanding.

A   For every submission:

☑ **A1.** Did you describe the limitations of your work?
8 Limitations.

☑ **A2.** Did you discuss any potential risks of your work?
9 Ethics Statement.

☑ **A3.** Do the abstract and introduction summarize the paper's main claims?
0 Abstract; 1 Introduction.

☒ **A4.** Have you used AI writing assistants when working on this paper?
Left blank.

B   ☑ Did you use or create scientific artifacts?

5.1 Dataset; 5.2 Baseline models; 5.3 Evaluation Metrics.

☑ **B1.** Did you cite the creators of artifacts you used?
5.1 Dataset; 5.2 Baseline models; 5.3 Evaluation Metrics.

☑ **B2.** Did you discuss the license or terms for use and / or distribution of any artifacts?
9 Ethics Statement.

☑ **B3.** Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
9 Ethics Statement.

☑ **B4.** Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
9 Ethics Statement.

☑ **B5.** Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
5.1 Dataset.

☑ **B6.** Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
5.1 Dataset; Appendix A.2 Details of Dialogue Acts.

C   ☑ Did you run computational experiments?

5 Experiments.

☑ **C1.** Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5.4 Implementation Details.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5.4 Implementation Details.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
6 Results and Analysis.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A.1 Details of Package.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.