Focus-aware Response Generation in Inquiry Conversation

Yiquan Wu[1], **Weiming Lu**[1*], **Yating Zhang**[2], **Adam Jatowt**[3]
Jun Feng[4], **Changlong Sun**[12], **Fei Wu**[1*], **Kun Kuang**[1*]
[1]Zhejiang University, Hangzhou, China
[2]Alibaba Group, Hangzhou, China
[3]University of Innsbruck, Austria
[4]State Grid Zhejiang Electric Power Co., LTD, Hangzhou, China
{wuyiquan, luwm, kunkuang}@zju.edu.cn, yatingz89@gmail.com, jatowt@acm.org
changlong.scl@taobao.com, junefeng.81@gmail.com, wufei@cs.zju.edu.cn

Abstract

Inquiry conversation is a common form of conversation that aims to complete the investigation (e.g., court hearing, medical consultation and police interrogation) during which a series of focus shifts occurs. While many models have been proposed to generate a smooth response to a given conversation history, neglecting the focus can limit performance in inquiry conversation where the order of the focuses plays there a key role. In this paper, we investigate the problem of response generation in inquiry conversation by taking the focus into consideration. We propose a novel Focus-aware Response Generation (FRG) method by jointly optimizing a multi-level encoder and a set of focal decoders to generate several candidate responses that correspond to different focuses. Additionally, a focus ranking module is proposed to predict the next focus and rank the candidate responses. Experiments on two orthogonal inquiry conversation datasets (judicial, medical domain) demonstrate that our method generates results significantly better in automatic metrics and human evaluation compared to the state-of-the-art approaches.

1 Introduction

Thanks to the high effectiveness of machine learning techniques, natural language processing (NLP) has made tremendous progress in a variety of tasks; for example, in conversation response generation which empowers many applications such as chatbots (e.g., Siri). The performance of response generation was significantly improved after applying neural network models such as recurrent neural networks (RNN) (Cho et al., 2014; See et al., 2017) and Transformers (Vaswani et al., 2017; Ji et al., 2020). However, existing studies on response generation mainly concentrate on relevance and fluency, rarely paying attention to the focus, which is important from the viewpoint of the rationality of generated responses.

Inquiry conversation (inquiry dialogue) is a common form of conversation that aims to complete the investigation (Hamami, 2014) (e.g., court hearing, medical consultation, police interrogation). Focus shifts tend to occur often in inquiry conversation, and their order plays a key role. For example, as shown in Fig.1, a judge will not issue the verdict before the defendants have finished pleading, and the doctor will not prescribe drugs before stating a diagnosis. The latent focuses of utterances often affect dialogue development, and hence it is beneficial to incorporate the notion of focus in the response generation process.

In this paper, we focus on response generation in inquiry conversation, and we aim at improving the rationality of generated content. For practical reasons, we only generate the responses of the leading role speaker in a conversation (e.g., judge, doctor). When addressing this problem one faces the following challenges: (1) **The focuses are sequential yet latent.** The next response should be generated considering the focuses underlying in the conversation history, and the next focus needs to be predicted. (2) **The focuses are discrete and different focuses correspond to different responses.** Thus, the generator needs to determine the focus before and then generate a response guided by the established focus.

To address these challenges, we propose a novel focus-aware response generation (FRG) method by jointly optimizing a multi-level encoder, a set of focal decoders (to generate responses with different focuses), and a synergistic focus ranking module. Specifically, the multi-level encoder is designed to better learn the latent focuses from the conversation history based on the aggregated characteristics of speakers and the content in each block (defined in Sec.3) through a speaker level attention layer and a block level attention layer. Then, each decoder in the set of focal decoders generates a candidate response guided by its corresponding focus. Finally,

Findings of the Association for Computational Linguistics: ACL 2023, pages 12585–12599

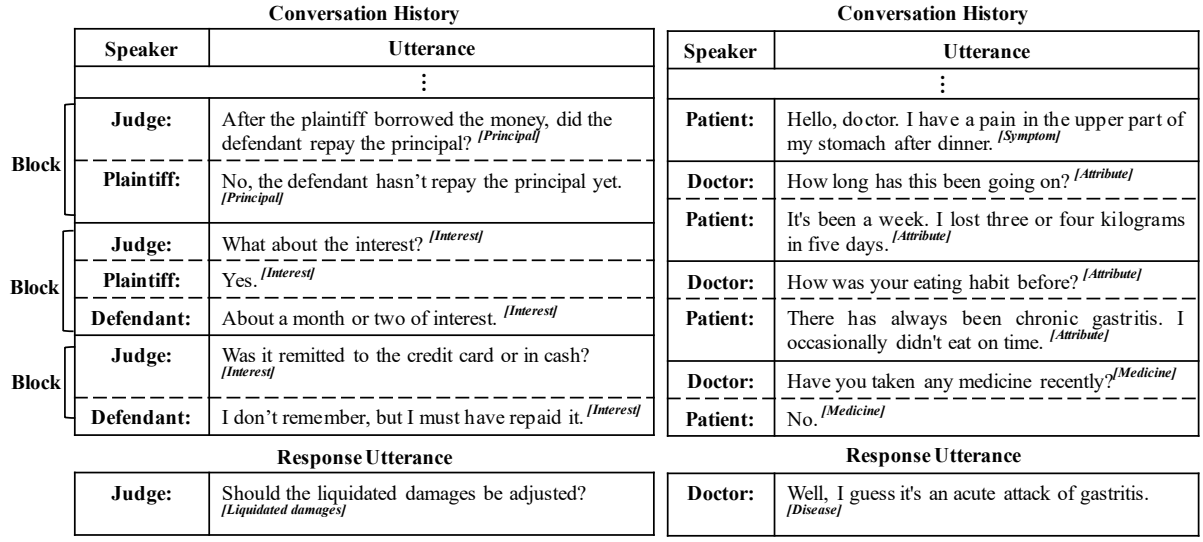

Figure 1: Examples of response generation. The left column is a conversation in a court hearing, where the focus shifts from *Principal* to *Liquidated damages*. The right one is a conversation in medical consultation, where the focus shifts from *Symptom* to *Disease*. Note that only the focuses in response utterances are given as annotations. The focuses in the Conversation history are actually not labeled in the dataset.

the focus ranking module ranks all the candidate responses generated by the focal decoders and predicts the next focus for the final output.

To test the proposed method, we employ two inquiry conversation datasets from two diverse domains - court hearing and medical consultation. Due to the difficulty and high cost of annotating focuses in different domains which typically require input from domain experts, we use a two-stage training paradigm to assure the generalizability of our method. We first warm-up the decoders together with a large number of unlabeled data to ensure the generation ability, and then we fine-tune them separately on a small number of labeled data to ensure the generation quality of particular focus. Extensive experiments show that the proposed FRG model achieves the best performance on both automatic metrics and human evaluation.

To sum up, our contributions are as follows:

- We investigate the response generation task in inquiry conversation by involving the focus in the generation process.
- We propose a novel focus-aware response generation (FRG) method by jointly optimizing a multi-level encoder, a set of focal decoders, and a synergistic focus ranking module.
- We validate the performance of the proposed method with extensive experiments on two orthogonal inquiry conversation datasets. The experiments indicate the high domain adaptability of our approach.
- To motivate other researchers to investigate this task, we make the code publicly available [1].

2 Related Work

2.1 Conversational NLG

Neural language generation (NLG) has been widely studied and applied in many tasks including machine translation (Wu et al., 2016; He et al., 2018; Shen et al., 2019), question answering (McCann et al., 2018; Bagchi and Wynter, 2013) and text summarization (Rush et al., 2015; Liu and Lapata, 2019; Wu et al., 2020, 2022). Existing NLG methods can be divided into rule-based and neural-based. The rule-based methods generate content through manually formulated templates (Yang et al., 2019; Becker, 2002). Such responses tend to be smooth and regular, but the cost of formulating templates is quite high. The neural-based methods take the advantage of deep learning (Shen et al., 2021; Zhang et al., 2022a,b; Li et al., 2022a,b; Zhang et al., 2023; Qian et al., 2023; Ma et al., 2021), which requires far less labor and enables flexibility. Bahdanau et al. (2015) firstly applied the attention mechanism into the NLG task. See et al. (2017) proposed a Pointer-Generator Networks (PGN), which can solve the Out-Of-Vocabulary (OOV) problem.

[1] https://github.com/wuyiquan/FRG

In conversational scenarios, many relevant NLG techniques have been also proposed, such as dialogue summarization (Chen and Yang, 2020), chatbots (Li et al., 2016), and response generation (Zhou et al., 2018b). In our work, we focus on the task of response generation for inquiry conversation.

2.2 Response Generation

Response generation is a key task in NLG, which aims to generate a response based on the conversation history (Zhou et al., 2018a,b; Zeng and Nie, 2021). Several approaches have been proposed to improve generation performance. Xing et al. (2017) proposed Topic-Aware Neural Response Generation (TAS2S) which incorporates pre-processed topic words to generate the response. Lau et al. (2017) introduced a Topically Driven Neural Language Model (TDLM) method, which can generate a response based on the predicted topic embedding. Lei et al. (2021) applied a hierarchical speaker-aware encoder to model the conversation. Zhao et al. (2017) propose a dialogue act-guided generation work, which aims to improve the diversity of the response. Wu et al. (2021) proposed a controllable grounded response generation framework, which uses an explicit hint phrase to generate. Due to the popularity of pre-training, several pre-trained models have been employed for response generation task, such as TransferTransfo (Wolf et al., 2019) and DialoGPT (Zhang et al., 2020b).

In this work, we emphasize the focus shifts among the blocks in the conversation, therefore the block level attention module is proposed for capturing their sequences. In addition, our model uses a set of focal decoders to generate a ranking list of responses corresponding to the predicted focus which is more applicable in practical use.

3 Problem Formulation

In this section, we define the problem of response generation in inquiry conversation. We first describe the key concepts as below:

Inquiry conversation is a form of conversation that aims to complete an investigation (Hamami, 2014) (e.g., court hearing, medical consultation).

Focus is the center of the conversation at a certain stage of its progress. The focuses tend to shift during the conversation.

Leading role is the speaker (e.g., interrogating speaker such as a judge, doctor) who controls the focus shifts in the inquiry conversation.

Block consists of several consecutive utterances and is regarded as the smallest unit of the focus shifting. Therefore, the conversation can be divided into several blocks according to the actions of the leading role speaker.

Response utterance refers to the interrogating utterance of the leading role speaker (examples are shown in Figure 1).

The problem of response generation in inquiry conversation is then defined here as follows:

Given the conversation history $\mathbf{U} = \{(\mathbf{u}_t, s_t)\}_{t=1}^{n_u}$ where $\{(\mathbf{u}_t, s_t)\}$ is tth pair of utterance u_t and the role of speaker s_t, the task is to determine the next focus f and based on it generate the corresponding response denoted as $\mathbf{r} = \{w_t\}_{t=1}^{m}$ for the leading role.

4 Method

In this section, we describe our focus-aware response generation (FRG) model. Fig.2 shows the overall framework. Our model consists of a shared multi-level encoder, a focus ranking module, and a set of focal decoders. The model works in a multi-task learning manner. The ranking module and decoders take the output of the encoder as an input.

4.1 Multi-level Encoder

The multi-level encoder consists of four layers, which encode the input from different levels. Firstly, we introduce two kinds of special tokens: (1) Speaker token <s> indicates the end of a speaker's utterance, where s is the id of the speaker. (2) Block token <b> refers to the end of a block. A block consists of several consecutive utterances with the same focus and is set automatically according to the speaking action of the leading role speaker (e.g., judge, doctor). For example, in Fig.2, the blocks are created every time the judge speaks. The input is transformed to:

$$I = \{\mathbf{u}_1, <s_1>, \mathbf{u}_2, <s_2>, <b>, \mathbf{u}_3, <s_3>, ..., \mathbf{u}_{n_u}, <s_{n_u}>, <b>\},$$

where $\mathbf{u}$ is the utterance, s is the corresponding speaker, n_u is the number of utterances. Note that since we only generate responses for the leading role speaker, I will always end with a <b>.

The input is a sequence of tokens. We then first transform the tokens into embeddings. The special tokens mentioned above are randomly initialized.

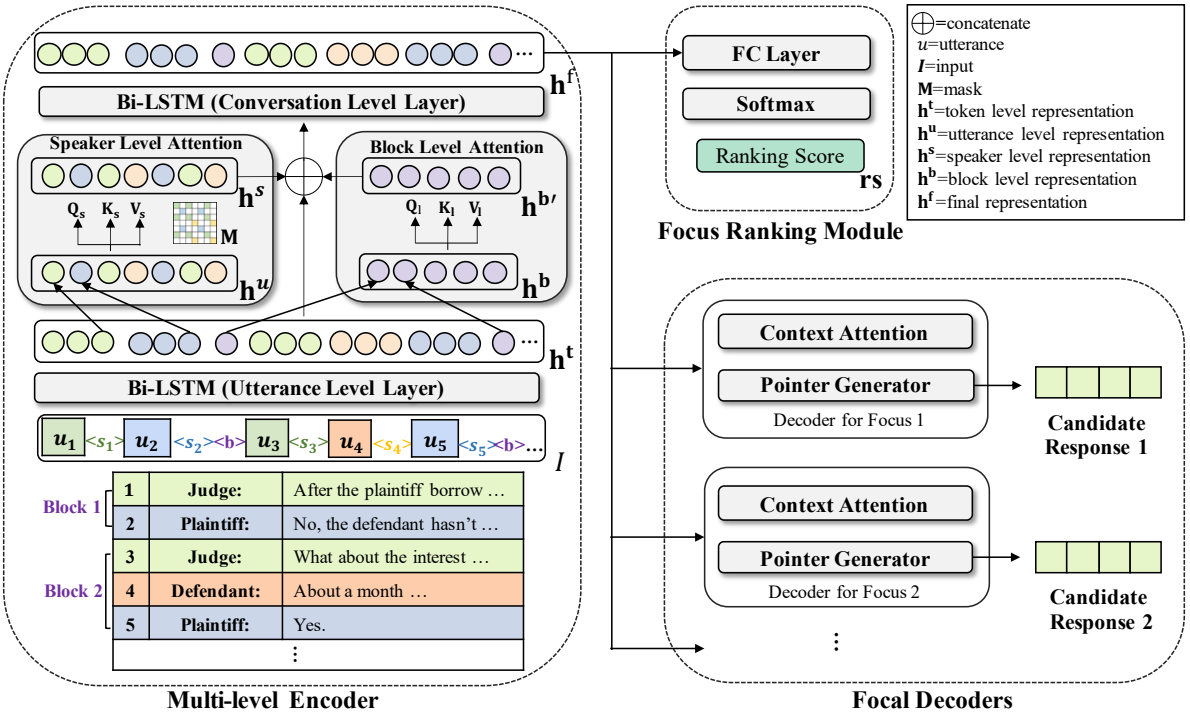

Figure 2: The architecture of FRG consists of a multi-level encoder, a set of focal decoders, and a synergistic focus ranking module.

4.1.1 Utterance Level Layer

In this layer, the embeddings of tokens are fed into a bidirectional LSTM (Bi-LSTM) (Huang et al., 2015), producing a token-level representation of the input $\mathbf{h^t} = \text{Bi-LSTM}(I)$.

To obtain a representation for each utterance, we take the output of the speaker token for that utterance. Thus, the utterance-level representation of the input is $\mathbf{h^u} = \{h_k^t\}, k \in X_S$, where X_S is the set of speaker token indices in I.

To obtain a representation for each block, we take the output of the block token for that focus block. Thus, the block-level representation of the input is $\mathbf{h^b} = \{h_k^t\}, k \in X_B$, where X_B is the set of block token indices in I.

4.1.2 Speaker Level Attention Layer

In the conversation, different speakers will play different roles. In order to obtain the speaker level representation, we create a special mask M according to the speaker's id. M is a matrix with the dimension of $[n_u, n_u]$. For any $m_{i,j}$ in M:

$$m_{i,j} = \begin{cases} 1 & s_i = s_j \\ 0 & s_i \neq s_j \end{cases}. \quad (1)$$

where s_i is the speaker of the utterance of $\mathbf{u_i}$.

Given the utterance-level representation $\mathbf{h^u}$ and the mask M, the speaker-level representation $\mathbf{h^s}$ is calculated as follows:

$$\mathbf{h^s} = \text{softmax}\left(\frac{Q_s^\top K_s M}{\sqrt{d_{ks}}}\right) V_s$$
$$Q_s = W_{Qs}\mathbf{h^u}, K_s = W_{Ks}\mathbf{h^u}, V_s = W_{Vs}\mathbf{h^u} \quad (2)$$

where W_{Qs}, W_{Ks}, W_{Vs} are learnable parameters, and d_{ks} is the dimension of K_s.

4.1.3 Block Level Attention Layer

In inquiry conversation, we assume the focus shifts only when the leading role speaker speaks, by which we divide the conversation history into several blocks.

Given the block-level representation $\mathbf{h^b}$, we run a self-attention on it, and the final block-level representation $\mathbf{h^{b'}}$ is calculated as follows:

$$\mathbf{h^{b'}} = \text{softmax}\left(\frac{Q_f^\top K_f}{\sqrt{d_{kf}}}\right) V_f$$
$$Q_f = W_{Qf}\mathbf{h^b}, K_f = W_{Kf}\mathbf{h^b}, V_f = W_{Vf}\mathbf{h^b} \quad (3)$$

where W_{Qf}, W_{Kf}, W_{Vf} are learnable parameters, and d_{kf} is the dimension of K_f.

4.1.4 Conversation Level Layer

In this layer, we concatenate the output of the former layers to get $\mathbf{h^{con}}$. For each h_i^t in the $\mathbf{h^t}$, we concatenate it with its corresponding speaker-level representation and block-level representation:

$$h_i^{con} = [h_i^t; h_{x(i)}^s; h_{y(i)}^{b'}] \tag{4}$$

where x is a function mapping the index of h^t and h^s, y is a function mapping the index of h^t and $h^{b'}$ and $[\cdot; ...; \cdot]$ represents the concatenation operation.

Then we use another Bi-LSTM layer to obtain the final representation of the input $\mathbf{h} = \text{Bi-LSTM}(\mathbf{h^{con}})$.

4.2 Focal Decoders

In order to make the model generate a reasonable response, we use a set of decoders with the same structure that aim to generate responses guided by different focuses. We call them focal decoders. Specifically, the number of decoders is equal to the number of predefined focuses.

Given the representation of the input $\mathbf{h}$ and the decoding state s_t, we apply the attention mechanism (Bahdanau et al., 2015). At each step t, the attention distribution a^t is calculated as follows:

$$\begin{aligned} e_i^t &= v^T \tanh\left(W_H h_i + W_S s_t + b_{\text{attn}}\right) \\ a^t &= \text{softmax}\left(e^t\right) \end{aligned} \tag{5}$$

where v, W_H, W_S, b_{attn} are learnable parameters.

The context vector h_t^* is the weighted sum of $\mathbf{h}$, such that $h_t^* = \sum_i a_i^t h_i$.

Then, the context vector h_t^* is concatenated with the decode state s_t and fed to linear layers to produce the vocabulary distribution p_{voc}:

$$p_{vo} = \text{softmax}(V'(V[s_t; h_t^*]) + b) + b') \tag{6}$$

where V, V', b, b' are all learnable parameters.

We use a generation probability (See et al., 2017) to solve the OOV problem. Given the context h_t^*, the decode state s_t and the decoder's input x_t, the generation probability p_{gen} is calculated as follows:

$$P_{gen} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x_t + b_{ptr}) \tag{7}$$

where w_{h^*}, w_s, w_x and b_{ptr} are learnable parameters, and σ is the sigmoid function.

The final probability for a word w for the current time step is obtained:

$$P(w) = P_{gen} * p_{voc}(w) + (1 - P_{gen}) \sum_{i:w_i=w} a_i^t \tag{8}$$

Given the same $\mathbf{h}$, the decoders will generate different outputs due to the different parameters. We explain how to warm-up and independently fine-tune the decoders in the training part.

4.3 Focus Ranking Module

Given the representation of the input $\mathbf{h}$, the focus ranking module will produce the probability of each focus through a fully connected layer and a softmax operation. The ranking score $\mathbf{rs} = \{rs_1, rs_2, ..., rs_{n_f}\}$ is obtained as $\mathbf{rs} = \text{softmax}(\text{FC}(\text{mean}(\mathbf{h})))$, where FC denotes a fully-connected layer. Then, the outputs of the decoders can be sorted by the $\mathbf{rs}$.

4.4 Two-Stage Training Paradigm

Since the annotation of the focus is difficult and costly, we adopt a two-stage training paradigm to assure the high generalization ability of our method.

In the first stage, we use a large number of unlabeled data to train the model without the ranking module, aiming to let the decoders acquire a good generation ability. Here, all the decoders share the same parameters.

For the decoders, the loss for time step t is the negative log-likelihood of the target word w_t^*:

$$\mathcal{L}_t = -logP(w_t^*), \tag{9}$$

and the overall generation loss is:

$$\mathcal{L}_{gen} = \frac{1}{T} \sum_{t=0}^{T} \mathcal{L}_t, \tag{10}$$

where T is the length of the response utterance.

In the second stage, we use a small number of labeled data to train the ranking module and fine-tune the encoder and decoders trained in the first stage. In this stage, each decoder corresponds to a different focus, and the decoders will be trained by the data annotated to their corresponding focus.

For the ranking module, we use cross-entropy as the loss function:

$$\mathcal{L}_{rank} = -\sum_{i=1}^{n_f} y_i \log\left(rs_i\right), \tag{11}$$

where $y_i = 1$ if $i = f$, otherwise, $y_i = 0$. f is the annotated focus. n_f stands for the number of focuses.

For the set of focal decoders, we take a mask operation when calculating the loss of each decoder.

The actual loss for the decoder d_i is:

$$\mathcal{L}_i = \begin{cases} \mathcal{L}_{gen}(i) & f = i \\ 0 & f \neq i \end{cases}, \quad (12)$$

where i is the corresponding focus of d_i and $\mathcal{L}_{gen}(i)$ is the generation loss of d_i.

Thus, the total loss in the second training stage is:

$$\mathcal{L}_{total} = \sum_{i=1}^{n_f} \mathcal{L}_i + \lambda * \mathcal{L}_{rank}, \quad (13)$$

where we set λ to $0.1 * n_f$.

4.5 Inference

During inference, the decoders apply beam search with the size of 4 to generate candidate outputs, which will be sorted by the ranking score **rs**.

5 Experiments

5.1 Dataset

We use the following two datasets for experiments: Court Hearing and Medical Consultation.

Court Hearing dataset:[2] Court hearing is a judicial event where the judge inquires the plaintiff and the defendant in order to clarify the facts of the case. The annotated data we use is released by Duan et al. (2019) [3]. The input is the conversation history, and the output is the next response utterance of the judge. There are seven focuses in this dataset: *Principal, Interest, Common debt claim, Guarantee liability, Liquidated damage, Creditor qualification, Limitation of action*.

Medical Consultation dataset:[4] Medical consultation is a conversation between a patient and a doctor. The annotated dataset we use is released by the competition: Conference on Knowledge Graph and Semantic Computing 2021 (CCKS21)[5]. There are five focuses for this dataset: *Symptom, Medicine, Test, Attribute, Disease*.

The statistics of the two datasets are shown in Tab.1. We randomly separate each dataset into a training set, a validation set, and a test set according to a ratio of 80%:10%:10%. The annotated data is ensured not to be in the test set.

[2]This dataset is provided by the High People's Court of a province in China.

[3]https://github.com/zhouxinhit/Legal_Dialogue_Summarization.

[4]The raw data can be downloaded in https://github.com/UCSD-AI4H/Medical-Dialogue-System.

[5]The data can be downloaded in https://www.biendata.xyz/competition/ccks_2021_mdg/data/

Type	CH	MC
# of Samples	240,000	100,000
# of Focuses	7	5
# of Annotations	7,000	5,000
Avg.# of tokens in input	106.9	90.3
Avg.# of speakers	2.47	2
Avg.# of tokens in response	13.6	13.1

Table 1: Statistics of the dataset. CH refers to court hearing and MC refers to medical consultation.

5.2 Evaluation Metrics

5.2.1 Automatic Evaluation

We adopt **ROUGE**[6], **BLEU** (Papineni et al., 2002) and **BERTScore** (Zhang et al., 2020a) as the automatic metrics. Specifically, we report the values of ROUGE-1 and ROUGE-L for ROUGE; BLEU-1 and BLEU-N (average of BLEU-1 to BLEU-4) for BLEU; P, R and F1 for BERTScore.

5.2.2 Human Evaluation

We conduct a human evaluation to analyze the quality of the generated responses. We randomly sample 500 test cases from each dataset. For each case, we present the responses generated by 5 representative methods[7] together with the ground truth to 5 annotators. The evaluation is conducted following two perspectives: (1) **Rationality level**. The rationality indicates the logical coherence between the conversation history and the generated response. Annotators are asked to give a score on the rationality of the generated response. (2) **Fluency level**. Annotators are asked to give a score on the fluency of the generated response. Both scores range from 1 to 5 (1 for the worst and 5 for the best).

5.3 Baselines

We employ the following methods as baselines for comparison with our approach:

L-Distance (Levenshtein distance) is used to measure the difference between two texts. Given the input of the test case, we find out the case in the training dataset with the smallest L-distance and take its response as the output. This method performs in a text retrieval manner. **LSTM+ATT** (Sutskever et al., 2014) and **PGN** (See et al., 2017) are RNN-based models. **T5** (Raffel et al., 2020) and **GPT-2** (Radford et al., 2019) are transformer-based models for NLG task. We also fine-tune them on the task datasets. **TransferTransfo** (Wolf

[6]https://pypi.org/project/rouge/

[7]We shuffle all the results to make fair evaluation for all the methods.

Methods	Court Hearing							Medical Consultation						
	ROUGE		BLEU		BERTScore			ROUGE		BLEU		BERTScore		
	R-1	R-L	B-1	B-N	P	R	F1	R-1	R-L	B-1	B-N	P	R	F1
L-Distance	10.7	10.5	27.8	1.3	60.0	62.1	61.0	9.5	9.1	31.1	2.3	62.0	62.2	62.1
LSTM+ATT (Bahdanau et al., 2015)	16.1	15.0	42.0	12.7	62.8	63.5	63.1	11.7	10.7	37.9	8.6	62.5	63.4	62.9
PGN (See et al., 2017)	17.3	15.5	43.3	17.4	65.1	64.1	64.6	12.0	10.7	40.5	9.8	63.5	63.0	63.2
GPT-2 (Radford et al., 2019)	16.4	14.6	39.6	13.9	63.6	64.3	63.9	13.0	11.5	36.1	10.5	63.3	63.3	63.0
T5 (Raffel et al., 2020)	15.8	14.4	38.1	12.8	62.4	62.3	62.3	11.3	10.2	34.7	9.6	63.0	63.1	63.0
TransferTrasnfo (Wolf et al., 2019)	16.5	14.8	41.4	14.0	64.2	63.5	63.8	13.2	12.1	37.8	12.7	64.2	64.5	64.3
DialogGpt (Zhang et al., 2020b)	16.6	15.3	42.0	13.9	63.6	63.6	63.6	13.5	11.3	37.5	12.4	63.3	63.5	63.3
† TDLM (Lau et al., 2017)	22.3	19.5	45.4	17.6	67.2	67.2	67.2	16.1	13.5	45.2	14.3	64.0	63.6	63.8
† TAS2S (Xing et al., 2017)	23.8	18.2	45.2	17.7	68.5	68.0	68.2	15.6	13.0	42.7	14.3	64.6	64.5	64.5
† MPG (Ide and Kawahara, 2021)	21.1	18.4	43.2	16.3	66.5	67.5	67.0	14.8	12.7	42.9	13.9	64.2	63.7	63.9
FRG w/o RM	18.6	16.3	44.9	16.7	66.7	65.7	66.2	12.5	11.3	37.8	11.6	63.3	65.5	64.4
† FRG w/o ML	30.1	26.8	55.4	25.8	68.3	66.6	67.4	16.6	15.6	44.6	13.3	65.8	66.8	66.3
† FRG w/o BL	31.6	27.0	58.0	26.5	68.3	68.6	68.4	16.6	15.7	40.2	15.4	66.2	66.0	66.1
† FRG w/o SL	32.3	28.1	57.4	27.0	67.0	67.0	67.0	17.0	16.2	43.1	16.0	65.7	65.8	65.7
† FRG-top1	**33.3**	**29.4**	**59.7**	**28.7**	**72.5**	**72.2**	**72.3**	**17.9**	**16.5**	**50.3**	**16.9**	**66.6**	**67.8**	**67.2**
† FRG-top3*	41.5	36.9	60.5	34.9	73.8	71.7	72.7	27.2	25.1	52.4	22.1	72.4	77.0	74.6

Table 2: Results on legal and medical datasets. † denotes models that use annotation data. * indicates the model that is not used for comparison.

Methods	Court Hearing		Medical Consultation	
	Rat.	Flu.	Rat.	Flu.
L-Distance	2.12	**4.01**	1.76	**4.17**
PGN	3.07	3.29	2.52	3.19
GPT-2	3.03	3.32	2.56	3.23
TAS2S	3.45	3.34	2.78	3.15
FRG-top1	**3.78**	3.49	**3.55**	3.43

Table 3: Results of human evaluation.

et al., 2019) and **DialoGPT** (Zhang et al., 2020b) are dialogue pre-trained models. We fine-tune them on task datasets. **TDLM** (Lau et al., 2017) predicts focus embedding first, then sends the focus embedding to the decoder to form responses. **TAS2S** (Xing et al., 2017) predicts focus words first, then takes the focus words as the external vocabulary to the decoder. **MPG** (Ide and Kawahara, 2021) uses multi-task learning to simultaneously predict the focus and generate the response.[8]

FRG-top1 indicates that we choose as the output the content generated by the decoder which has the highest ranking score, while **FRG-top3** means that we take the three top-ranked candidates at the same time. The latter simulates a practical scenario that a user could select an appropriate answer from the suggested candidates.

We also conduct the ablation experiments on **FRG-top1** as follows: **FRG w/o RM** removes the ranking module and replaces the set of decoders with a single decoder. **FRG w/o ML** removes the speaker level attention layer and block level attention layer. **FRG w/o BL** removes the block level attention layer. **FRG w/o SL** removes the speaker level attention layer.

5.4 Experimental Results

In this section we analyze the experimental results[9].

Quantitative evaluation. Tab.2 demonstrates the results of response generation on both Court Hearing and Medical Consultation datasets with ROUGE, BLEU, and BERTScore.

Based on the results, we make the following observations: (1) **L-Distance** method has the worst performance in both datasets, which means that simply retrieving the response from the dataset based on the context similarity is not promising. (2) RNN-based baselines and Transformer-based baselines achieve similar performance in this task yet much lower than the performance of **FRG**. It demonstrates that with the help of a multi-level encoder and focal decoders, **FRG** is capable of estimating the focus of the leading role speaker and thus generating more precise content. (3) Models that employ annotations achieve better performance, which proves the usefulness of considering the focus. (4) **TDLM** and **TAS2S** show that merging the focus embedding into the decoder brings only a small improvement, which suggests the positive effect of the focal decoders. (5) Moreover, **FRG** also seems to have good domain adaptability by achieving the best performance on both Court Hearing and Medical Consultation datasets compared with the baselines.

To investigate the effects of the number of annotations in the second training stage, we study the performance change in Fig. 4 and draw the following conclusions: (1) A small number of annotations can bring significant improvement to the

[8]The focus is called as topic in TDLM and TAS2S, while it is called as emotion label in MPG.

[9]The detailed parameter settings are shown in Appendix and the code will be released.

Conversation History

Speaker	Utterance
	⋮
Judge:	Was it the IOU written first or the loan delivered first?
Plaintiff:	I wrote the IOU first, then the defendant went to my home to get the money.
Judge:	What is the relationship between the plaintiff and the defendant?
Plaintiff:	We were classmates.
Judge:	Where was the money from?
Plaintiff:	Some was my own funds and others was from banks.
Judge:	How was the interest agreed?
Defendant:	The interest was 2.5%.

Response Utterance

Ground Truth	Has the defendant ever paid interest after issuing IOU?
L-Distance	Does the plaintiff have any other facts and evidence to supplement?
PGN	Plaintiff, what is your relationship with the defendant?
GPT-2	What is the relationship between the plaintiff and the defendant?
TAS2S	What about the interest?
FRG (ours)	Has the defendant paid any interest?

Conversation History

Speaker	Utterance
	⋮
Patient:	I feel short of breath and stuffy when I eat, and my feet and hands are soft.
Doctor:	It is recommended to seek medical advise in gastroenterology. How long has it been like this?
Patient:	It's been good and bad for half a year.
Doctor:	Do you have a regular diet?
Patient:	No, and I always like ice.
Doctor:	Do you work and rest regularly?
Patient:	Yes, I never stay up late.

Response Utterance

Ground Truth	It may be caused by eating habits. A gastroscope is recommended.
L-Distance	Yes, suitable for treatment.
PGN	It is recommended to see a doctor in the hospital.
GPT-2	Do you have a regular diet?
TAS2S	Hello, you can see a doctor in the Department of Gastroenterology.
FRG (ours)	I suggest you have a gastroscope.

Figure 3: Case study. The left (right) one refers to a case from Court Hearing (Medical Consultation) dataset.

model (e.g., boosting ROUGE-L from 16.3 to 25.8 for the Court Hearing dataset). With the increase of the number of annotations, the performance of the model continues to improve. (2) The effect of annotations on judicial domain data is stronger than that for the medical domain. It indicates that the number and the granularity of the focuses used may influence the performance.

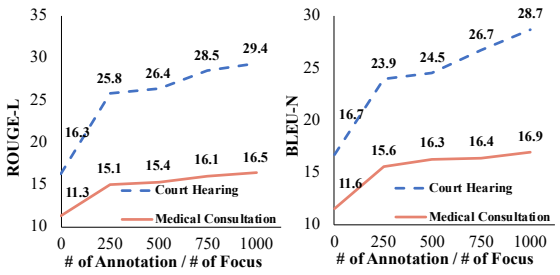

Figure 4: The effect of the number of annotations on ROUGLE-L (left) and BLEU-N (right) curves.

Qualitative evaluation We show the result of human evaluation in Tab. 3, and report the following observations: (1) Although **L-Distance** has high performance in fluency due to its retrieval method, it achieves very poor results in focus rationality. (2) Thanks to the focal decoders, **FRG** significantly improves the performance at the rationality level. (3) **FRG** also achieves better performance on fluency level compared to other generative methods. (4) Kappa coefficient κ between any two human annotators is above 0.8, which indicates the high quality of human evaluation.

Ablation Study We report the results of ablation study in Tab. 2 noticing a dramatic decrease in the performance of **FRG w/o RM** (e.g., decrease from 33.3 to 19.6 on R-1 in Court Hearing dataset), which points to the high importance of ranking module and focal decoders. Similarly, the **FRG w/o ML**, **FRG w/o BL** and **FRG w/o SL** also experience a decrease in performance, albeit less than **FRG w/o RM**. This confirms the effectiveness of the proposed block level attention layer and speaker level attention layer in the encoder.

5.5 Case Study

Fig. 3 shows two cases of the responses generated by our method (**FRG**) and by the four baseline methods to provide a more intuitive understanding of the performance of each method. We find that the output of **L-Distance** is irrelevant to the conversation history. The utterances generated by **PGN**, **GPT-2** and **TAS2S** are more likely to repeat the content already spoken in the conversation history. **FRG** is able to generate more reasonable content thanks to the guidance of the focus.

5.6 Error analysis

To explore the limitations of our model, we also analyze generated responses that had a high error rate[10], then we summarize the problems that occur, and also explore optimization solutions.

After conducting statistical analysis, we make the following observations: (1) **FRG** performs worse when external information needs to be used.

[10]We collect the samples in human evaluation whose either rationality or fluency score of FRG-top1 equals 1.

In the Court Hearing dataset, 27% of errors are related to this problem (e.g., "According to the law, the maximum interest rate shall not exceed four times the interest rate of similar bank loans."). At the same time, 38% of errors in Medical Consultation dataset are related to such problem (e.g., "According to the instructions, Trimebutine Maleate tablets and Golden Bifid can be taken after meals."). (2) 36% of errors in Court Hearing dataset and 47% of errors in Medical Consultation dataset occur when a long response needs to be generated (e.g., more than 25 tokens). (3) Long conversation history (e.g., more than 10 utterances) will also cause a high error rate. This is the case of 42% of errors in Court Hearing dataset and 53% of errors of Medical Consultation dataset.

To address these problems, constructing a retrieval database and enhancing the long dependence of language models can be promising for the future.

6 Conclusion and Future Work

In this paper, we investigate the response generation task in inquiry conversation from a focal view and propose a novel focus-aware response generation (FRG) method. We design a multi-level encoder to represent the conversation history at different levels, as well as a set of focal decoders to generate responses guided by different focuses. Thanks to the focus ranking module, the generated responses are sorted for the final output. The experiment results show the effectiveness of our method.

In the future, we will explore the following directions based on the FRG method: (1) Adding external knowledge to constrain the ranking module and (2) Using the feedback of users to optimize the ranking module in practical applications.

7 Limitations

In this section, we discuss the limitations of our work as follows:

- As described in the paper, our proposed method requires annotations of the latent focus; a small number of annotations (around 250 labeled samples per focus) can already bring a significant improvement (see Fig.4). Therefore when applying our approach to other domains it is necessary to prepare at least a few annotations.
- As mentioned in the error analysis section, the model is unable to generate unseen entities, such as specific drug names or laws. Further improvement should be made to solve this problem for practical use.

Acknowledgments

This work was supported in part by Key R&D Projects of the Ministry of Science and Technology (2020YFC0832500), National Natural Science Foundation of China (62006207, 62037001, U20A20387), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJUSIAS-0010), Project by Shanghai AI Laboratory (P22KS00111), Program of Zhejiang Province Science and Technology (2022C01044), the Fundamental Research Funds for the Central Universities (226-2022-00143, 226-2022-00142) and MOE Engineering Research Center of Digital Library.

Finally, we would like to thank the anonymous reviewers for their helpful feedback and suggestions.

References

Sugato Bagchi and Laura Wynter. 2013. Method for a natural language question-answering system to complement decision-support in a real-time command center. US Patent 8,601,030.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tilman Becker. 2002. Practical, template-based natural language generation with TAG. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks, TAG+ 2002, Venice, Italy, May 20-23, 2002*, pages 80–83. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4106–4118. Association for Computational Linguistics.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a*

Special Interest Group of the ACL, pages 1724–1734. ACL.

Xinyu Duan, Yating Zhang, Lin Yuan, Xin Zhou, Xiaozhong Liu, Tianyi Wang, Ruocheng Wang, Qiong Zhang, Changlong Sun, and Fei Wu. 2019. Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1361–1370. ACM.

Yacin Hamami. 2014. Inquiry in conversation: Towards a modelling in inquisitive pragmatics. *Logique et Analyse*, pages 637–661.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2018. Sequence to sequence mixture model for diverse machine translation. *arXiv preprint arXiv:1810.07391*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Tatsuya Ide and Daisuke Kawahara. 2021. Multi-task learning of generation and classification for emotion-aware dialogue response generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 119–125. Association for Computational Linguistics.

Changzhen Ji, Xin Zhou, Yating Zhang, Xiaozhong Liu, Changlong Sun, Conghui Zhu, and Tiejun Zhao. 2020. Cross copy network for dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1900–1910. Association for Computational Linguistics.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 355–365. Association for Computational Linguistics.

Yuejie Lei, Yuanmeng Yan, Zhiyuan Zeng, Keqing He, Ximing Zhang, and Weiran Xu. 2021. Hierarchical speaker-aware sequence-to-sequence model for dialogue summarization. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 7823–7827. IEEE.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1192–1202. The Association for Computational Linguistics.

Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, et al. 2022a. End-to-end modeling via information tree for one-shot natural language spatial video grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8707–8717.

Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Wenqiao Zhang, Jiaxu Miao, Shiliang Pu, and Fei Wu. 2022b. Hero: Hierarchical spatio-temporal reasoning with contrastive action correspondence for end-to-end video object grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3801–3810.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3728–3738. Association for Computational Linguistics.

Xinyin Ma, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Weiming Lu. 2021. MuVER: Improving first-stage entity retrieval with multi-view entity representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2617–2624, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *CoRR*, abs/1806.08730.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Peng Qian, Zhenguang Liu, Yifang Yin, and Qinming He. 2023. Cross-modality mutual learning for enhancing smart contract vulnerability detection on bytecode. In *Proceedings of the ACM Web Conference 2023*, pages 2220–2229.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389. The Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *International conference on machine learning*, pages 5719–5728. PMLR.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and label: A two-stage identifier for nested named entity recognition. 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)*, pages 2782–2794, Online. Association for Computational Linguistics.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 763–780. Association for Computational Linguistics.

Yiquan Wu, Yifei Liu, Weiming Lu, Yating Zhang, Jun Feng, Changlong Sun, Fei Wu, and Kun Kuang. 2022. Towards interactivity and interpretability: A rationale-based legal judgment prediction framework. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4787–4799. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Zeqiu Wu, Michel Galley, Chris Brockett, Yizhe Zhang, Xiang Gao, Chris Quirk, Rik Koncel-Kedziorski, Jianfeng Gao, Hannaneh Hajishirzi, Mari Ostendorf, and Bill Dolan. 2021. A controllable model of grounded response 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 14085–14093. AAAI Press.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3351–3357. AAAI Press.

Ze Yang, Wei Wu, Jian Yang, Can Xu, and Zhoujun Li. 2019. Low-resource response generation with template prior. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 1886–1897. Association for Computational Linguistics.

Yan Zeng and Jian-Yun Nie. 2021. An investigation of suitability of pre-trained language models for dialogue generation - avoiding discrepancies. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4481–4494. Association for Computational Linguistics.

Rongzhi Zhang, Jiaming Shen, Tianqi Liu, Jialu Liu, Michael Bendersky, Marc Najork, and Chao Zhang. 2023. Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. *arXiv preprint arXiv:2305.05010*.

Rongzhi Zhang, Rebecca West, Xiquan Cui, and Chao Zhang. 2022a. Adaptive multi-view rule discovery

for weakly-supervised compatible products prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4521–4529.

Rongzhi Zhang, Yue Yu, Pranav Shetty, Le Song, and Chao Zhang. 2022b. Prboost: Prompt-based rule discovery and boosting for interactive weakly-supervised learning. *arXiv preprint arXiv:2203.09735*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. 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, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 654–664. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4623–4629. ijcai.org.

A Appendices

A.1 The Settings of Parameters

All models are trained on 2 V100 GPU(16GB). The settings of parameters of our model are shown in Tab. 4.The train/eval/decode step is the same as `https://github.com/becxer/pointer-generator`.

Name	value	Note
hidden_dim	128	dimension of RNN hidden states
emb_dim	300	dimension of word embeddings
batch_size	16	minibatch size
max_sen_num	20	max rounds in history
max_enc_steps	200	max timesteps of encoder (max source text tokens)
max_dec_steps	20	max timesteps of decoder (max generated text tokens)
beam_size	4	beam size for beam search decoding
min_dec_steps	10	Minimum sequence length of generated text.
vocab_size	50,000	Size of vocabulary
lr	0.10	learning rate
keep_prob	0.5	keep prob
adagrad_init_acc	0.1	initial accumulator value for Adagrad
rand_unif_init_mag	0.02	magnitude for lstm cells random uniform inititalization
trunc_norm_init_std	0.1	std of trunc norm init, used for initializing everything else
max_grad_norm	2.0	for gradient clipping

Table 4: The settings of parameters of FRG.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

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

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ 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)?
No response.

☐ 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?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ 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.
No response.

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4

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?
4

☑ 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?
4

☑ 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.)?
4

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
4

☑ 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.?
4

☑ 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)?
4

☑ 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?
4

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
4

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.