MidMed: Towards Mixed-Type Dialogues for Medical Consultation

Xiaoming Shi[1]*, **Zeming Liu**[2]*, **Chuan Wang**[3], **Haitao Leng**[4], **Kui Xue**[1],
Xiaofan Zhang[1], **Shaoting Zhang**[1]†
[1] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[2] Research Center for Social Computing and Information Retrieval, HIT, Harbin, China
[3] State Key Laboratory of Information Security, IIE, CAS, Beijing, China
[4] MMU KuaiShou Inc., Hangzhou, China
`{shixiaoming, xuekui, zhangxiaofan, shaotingzhang}@pjlab.org.cn`
`zmliu@ir.hit.edu.cn; wangchuan@iie.ac.cn; lenghaitao@kuaishou.com`

Abstract

Most medical dialogue systems assume that patients have clear goals (medicine querying, surgical operation querying, etc.) before medical consultation. However, in many real scenarios, due to the lack of medical knowledge, it is usually difficult for patients to determine clear goals with all necessary slots. In this paper, we identify this challenge as how to construct medical consultation dialogue systems to help patients clarify their goals. To mitigate this challenge, we propose a novel task and create a human-to-human mixed-type medical consultation dialogue corpus, termed MidMed [1], covering five dialogue types: task-oriented dialogue for diagnosis, recommendation, knowledge-grounded dialogue, QA, and chitchat. MidMed covers four departments (otorhinolaryngology, ophthalmology, skin, and digestive system), with 8,175 dialogues. Furthermore, we build baselines on MidMed and propose an instruction-guiding medical dialogue generation framework, termed InsMed, to address this task. Experimental results show the effectiveness of InsMed.

1 Introduction

Current medical dialogue systems (Xu et al., 2019; Liao et al., 2020; Zeng et al., 2020; Liu et al., 2022a) mainly focus on diagnosis by obtaining symptoms and then making diagnosis automatically. These dialogue systems have shown significant potential and alluring technological value to simplify diagnostic procedures (Semigran et al., 2015). Previous works assume that patients have explicit goals (medicine querying, surgical operation querying, etc.), and perform in the way of task-oriented dialogue to accomplish patients' goals.

However, explicit patient goals are usually unavailable in real-world scenarios. For example, a patient wants to consult about his itchy skin but lacks medical knowledge. Thus, it is difficult for the patient to decide which slots (e.g. medicine or a surgical operation) are needed. To figure out explicit patient goals, medical consultation services are needed, which provide advice of treatment, medicine, food, etc., as shown in Figure 1. However, those medical consultation services are under explored in previous works.

To facilitate the study of medical consultation, we construct a new human-to-human **mi**xed-type **d**ialogue dataset for **med**ical consultation (MidMed), covering five dialogue types: task-oriented dialogue for diagnosis, knowledge-grounded dialogue, QA, recommendation, and chitchat. MidMed is constructed by revising dialogues of MedDialog (a human-to-human medical diagnosis dialogue dataset) (Zeng et al., 2020). As shown in Figure 1, a patient queries about "sweaty hands", and has no explicit goal for medicine or a surgical operation. In the scenario, the doctor first collects the symptoms and makes a diagnosis. To help clarify the patient's goal, the doctor further recommends medicine and food, replies for foods to avoid, and gives emotional comfort. Through the consultation, the patient determines to apply "dexamethasone cream" and have more "tomatoes". Finally, MidMed is obtained, containing 8,175 dialogues and 98,000 utterances, with at least three dialogue types in each dialogue.

To promote research on medical consultation dialogue systems, we conduct benchmarking experiments on MidMed for end-to-end dialogue generation. Furthermore, to generate informative and relevant responses with dialogue topic sequences, inspired by Schick and Schütze (2021); Wei et al. (2021), we present an **ins**truction-guiding **me**dical **d**ialogue generation framework (InsMed) to handle mixed-type dialogues. InsMed is composed

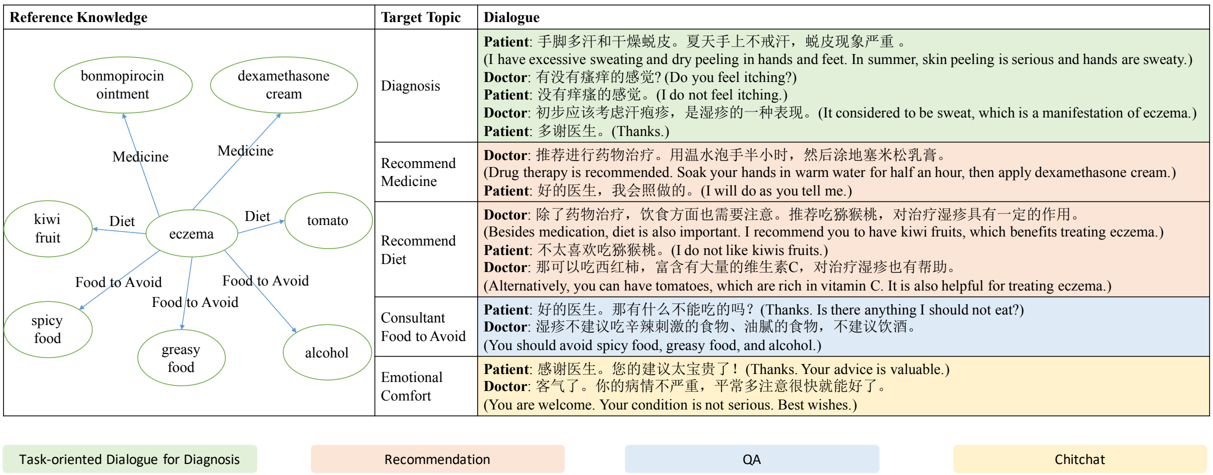
---

* Equal contribution.

† Corresponding author: Shaoting Zhang.

[1] MidMed is publicly available at https://github.com/xmshi-trio/MidMed

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8145–8157
July 9-14, 2023 ©2023 Association for Computational Linguistics



Target Topic	Dialogue
Diagnosis	**Patient**: 手脚多汗和干燥蜕皮。夏天手上不戒汗，蜕皮现象严重 。 (I have excessive sweating and dry peeling in hands and feet. In summer, skin peeling is serious and hands are sweaty.) **Doctor**: 有没有瘙痒的感觉? (Do you feel itching?) **Patient**: 没有痒瘙的感觉。(I do not feel itching.) **Doctor**: 初步应该考虑汗疱疹，是湿疹的一种表现。(It considered to be sweat, which is a manifestation of eczema.) **Patient**: 多谢医生。(Thanks.)
Recommend Medicine	**Doctor**: 推荐进行药物治疗。用温水泡手半小时，然后涂地塞米松乳膏。 (Drug therapy is recommended. Soak your hands in warm water for half an hour, then apply dexamethasone cream.) **Patient**: 好的医生，我会照做的。(I will do as you tell me.)
Recommend Diet	**Doctor**: 除了药物治疗，饮食方面也需要注意。推荐吃猕猴桃，对治疗湿疹具有一定的作用。 (Besides medication, diet is also important. I recommend you to have kiwi fruits, which benefits treating eczema.) **Patient**: 不太喜欢吃猕猴桃。(I do not like kiwis fruits.) **Doctor**: 那可以吃西红柿，富含有大量的维生素C，对治疗湿疹也有帮助。 (Alternatively, you can have tomatoes, which are rich in vitamin C. It is also helpful for treating eczema.)
Consultant Food to Avoid	**Patient**: 好的医生。那有什么不能吃的吗？(Thanks. Is there anything I should not eat?) **Doctor**: 湿疹不建议吃辛辣刺激的食物、油腻的食物，不建议饮酒。 (You should avoid spicy food, greasy food, and alcohol.)
Emotional Comfort	**Patient**: 感谢医生。您的建议太宝贵了！(Thanks. Your advice is valuable.) **Doctor**: 客气了。你的病情不严重，平常多注意很快就能好了。 (You are welcome. Your condition is not serious. Best wishes.)

Figure 1: An example of MidMed.

of a dialogue topic selection, a reference knowledge selection, and an instruction-based response generation module. Specifically, the topic selection module and the reference knowledge selection module are designed to pick suitable dialogue topics and reference knowledge for generating responses, respectively. Then, dialogue topics and reference knowledge are converted to instructions in natural language with well-designed templates. For example, an instruction is "*In the next utterance, the doctor will recommend a diet. The recommended diet is fruits and vegetables*". These instructions are concatenated with dialogue context as the input to generation models.

This work makes the following contributions:

- We identify a new challenge, that is, in many real-world scenarios, it is usually difficult for patients to have clear goals before medical consultations.
- To mitigate this challenge, we propose a novel task, medical consultation over mixed-type dialogue, and collect a new Chinese human-to-human mixed-type dialogue dataset, in which each session has rich variability of dialogue types with natural topic transitions.
- We build baselines on MidMed and propose an instruction-guiding response generation framework InsMed to address this task. Experimental results show the effectiveness of InsMed.

2 Related Work

2.1 Dialogue Systems for Diagnosis

There has been growing research interest in developing dialogue systems for automatic diagnosis. These dialogue systems aim to assist doctors in pre-collecting symptoms and patient information and then give patients diagnoses in time. These works are divided into two categories, the pipeline manner, and the end-to-end manner. Wei et al. (2018); Xu et al. (2019); Lin et al. (2019); Wang et al. (2021); Liu et al. (2022a) break the systems into natural language understanding, dialogue management, and natural language generation, in a pipeline manner. Then, these three modules are trained with respective annotated data and feed their output to the next module. Meanwhile, Zeng et al. (2020) tries to build an end-to-end model on large-scale unannotated medical dialogue data. Compared with the pipeline manner, the end-to-end manner has no requirement for the annotated dataset but has no supervision for the intermediate state.

In addition to methods, many datasets are also publicly available. The medical dialogue datasets are listed in Table 1. Among them, MZ (Wei et al., 2018), DX (Xu et al., 2019), CMDD (Lin et al., 2019), MedDG (Liu et al., 2022a), and DialoACM (Chen et al., 2022) are datasets of pipeline dialogue systems for automatic diagnosis. MedDialog (Zeng et al., 2020) is a large-scale unannotated dataset, utilized for end-to-end training.

These medical dialogue datasets focus on diagnosis, and ignore consultation. Compared with these datasets, MidMed is a medical dialogue dataset for

Datasets	Mixed-type	Medical	Dialogue Types
MZ (Wei et al., 2018)	✗	✓	Task-oriented dialogue for diagnosis
DX (Xu et al., 2019)	✗	✓	Task-oriented dialogue for diagnosis
CMDD (Lin et al., 2019)	✗	✓	Task-oriented dialogue for diagnosis
MedDG (Liu et al., 2022a)	✗	✓	Task-oriented dialogue for diagnosis
MedDialog (Zeng et al., 2020)	✗	✓	Task-oriented dialogue for diagnosis
DialoAMC (Chen et al., 2022)	✗	✓	Task-oriented dialogue for diagnosis
DuRecDial (Liu et al., 2020)	✓	✗	Rec., chitchat, QA, task-oriented dialogue
DodecaDialogue(Shuster et al., 2020)	✓	✗	Know., chitchat, QA, empathetic dialogue, image chat
BlendedSkillTalk(Smith et al., 2020)	✓	✗	Know., empathetic dialogue,chitchat
ACCENTOR(Sun et al., 2021)	✓	✗	Chitchat,task-oriented dialogue
DuRecDial 2.0 (Liu et al., 2021)	✓	✗	Rec., chitchat, QA, task-oriented dialogue
SalesBot(Chiu et al., 2022)	✓	✗	Chitchat,task-oriented dialogue
DuClarifyDial(Liu et al., 2022b)	✓	✗	Rec., know. chitchat, QA, task-oriented dialogue
MidMed (Ours)	✓	✓	Rec., chitchat, know., QA, diagnosis-oriented dialogue

Table 1: Comparison of MidMed with other datasets. "know.", and "rec." stand for knowledge-grounded dialogue, and conversational recommendation, respectively.

consultation, covering mixed-type dialogues.

2.2 Mixed-type Dialogue Systems

Recently, research on the mixed-type dialogue has increased significantly. These researches fall into two categories: (1) train an all-in-one conversation model by using multiple single-skill conversation datasets, such as persona-chat, task-oriented dialogue, to bind multiple dialogue skills (Madotto et al., 2020; Roller et al., 2021; Madotto et al., 2021); (2) collect mixed-type dialog datasets (Shuster et al., 2020; Smith et al., 2020; Liu et al., 2020; Sun et al., 2021; Liu et al., 2021; Chiu et al., 2022; Liu et al., 2022b) to train mixed-type dialog models. Those datasets are intended to mix different dialogue skills to meet specific needs, such as recommending movies and songs, and are unable to solve medical consultations. Compared with them, we collect a mixed-type dialogue corpus, MidMed, to facilitate the study of medical consultations.

3 Dataset Collection

In this section, we describe the three steps for MidMed construction: (1) Selecting basic diagnosis dialogue data; (2) Constructing annotation guidance; (3) Collecting mixed-type dialogue by crowdsourcing.

3.1 Selecting Basic Diagnosis Dialogue

To be close to real-world scenarios, MidMed is constructed based on real diagnosis dialogue dataset MedDialog (Zeng et al., 2020), which is collected from online medical community haodf.com.

MedDialog dataset contains 3.4 million Chinese dialogues (consultations) between patients and doctors, covering 29 broad categories of specialties including internal medicine, pediatrics, dentistry, etc., and 172 fine-grained specialties including cardiology, neurology, gastroenterology, urology, etc.

Basic Dialogue Selection. For MidMed construction, we recruit twenty medical students, who are experts in four departments, otorhinolaryngology, ophthalmology, skin, and the digestive system department. To ensure better data quality and construction efficiency, the dialogues only in these four departments are reserved. Besides, we observe that dialogues with few dialogue utterances are usually of poor quality. Thus, for high data quality and efficiency of data construction, only those conversations with more than four utterances are kept. After the above data processing, there are total 9,000 dialogues obtained.

Coarse-grained Privacy Removing. Furthermore, for ethical concerns, specific regular expressions for coarse-grained filtering are employed to remove privacy. To delete patients' privacy, regular expressions, such as " 我叫... (My name is ...)", are designed to delete sentences containing name, gender, and region. Besides, regular expressions, such as " 陈医生您好... (Hello, doctor Chen, ...)", are utilized to delete doctors' privacy.

3.2 Constructing Annotation Guidance

Annotation guidance is designed to instruct annotators for data annotation, including target dialogue topic sequences and reference knowledge. Specifically, target topic sequences assign topics for each dialogue session. To support the annotation of each topic, reference knowledge is provided.

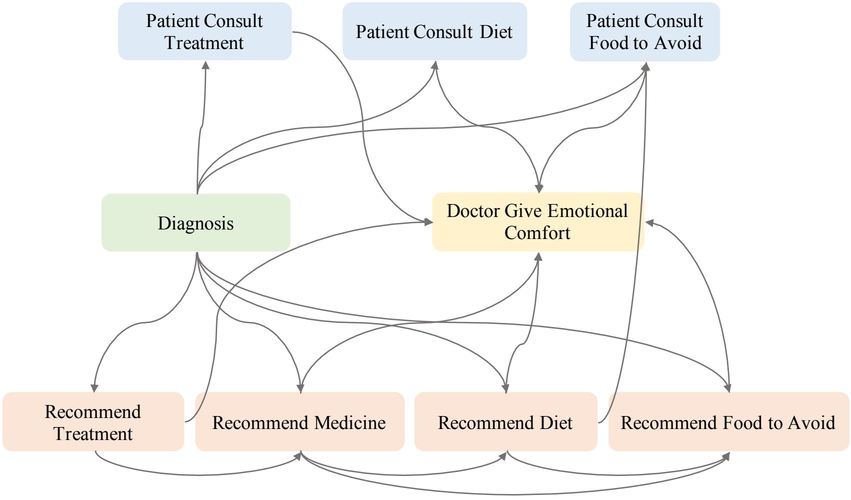


Figure 2: The illustration of dialogue topic sequences.

3.2.1 Target Dialogue Topic Sequence

Due to the complexity of the data annotation, it is of great difficulty to conduct data annotation with only high-level instructions. Inspired by the work of MultiWOZ (Budzianowski et al., 2018), we provide a target dialogue topic sequence for each dialogue construction. The dialogue topic sequences are employed to instruct annotators to annotate the content of specific topics. As shown in Figure 1, the target dialogue topic sequence is composed of dialogue topics, including `Patient Self Report`, `Doctor Inquiry Additional`, `Doctor Recommend Medicine`, etc. The whole dialogue topic sequences are shown in Figure 2. The combination of different topics ensures the diversity of dialogue topic sequences.

3.2.2 Reference Knowledge

The knowledge graph stores large-scale knowledge in the form of easy-to-use triples, and it has various applications in all modules of the human-computer dialogue system (Tuan et al., 2019, 2022; Yang et al., 2020). Therefore, we incorporate knowledge graphs into medical consultation to provide more accurate interactive questions and answers. Specifically, we crawled a large number of web pages from some high-quality medical vertical websites such as 39.net[2] and then obtained a large amount of triplet knowledge by using information extraction techniques such as entity extraction and relation extraction. By using these triples, a large-scale medical knowledge graph is constructed, whose entities include diseases, symptoms, drugs, foods, etc., and relationships include disease-drug relation, disease-food relation, etc.

To provide reference knowledge for dialogue annotation, we extract a knowledge graph subset for each dialogue. Specifically, diseases in the whole knowledge graph are mapped with the dialogue with exact string matching. The disease existing in the medical dialogues are employed as the head entity for select triples from the knowledge graph. Finally, we extract a knowledge graph subset, which covers four types of entities: disease, symptom, diet, and medicine, with a total of 229,570 triples.

3.3 Collecting Mixed-type Dialogue

For data annotation, the trial annotation and the formal annotation are conducted, sequentially. First, the trial annotation aims to select an annotation team and make the annotation team get familiar with the guide. Second, the formal annotation is conducted for collecting the whole dataset.

3.3.1 Trial Annotation

To ensure the high quality of dialogues, trial annotation is conducted. In the trial annotation stage, three crowdsourcing teams (about 20 annotators per team) are selected for trial annotation. There are mainly two advantages. (1) Trial annotation helps select a reliable annotation team. (2) The trial annotation helps the annotation team get familiar with the annotation task. Lastly, the team achieving the best performance in the trial annotation is selected for the formal annotation.

---

[2] http://www.39.net

3.3.2 Formal Annotation

After the trial annotation, the formal annotation is conducted. In the formal annotation, to ensure data quality, the fine-grained privacy removing, skipping option, and quality audit and re-annotating mechanisms are employed. To ensure diversity, the mechanism of annotation without target dialogue topic sequences is applied.

Overall Annotation. In the formal data annotation process, annotators are required to act as doctors and patients in turn. Annotators construct dialogues based on a given basic diagnosis dialogue, a target dialogue topic sequence, and reference knowledge. The annotation progress is conducted as follows. First, the annotator enters the chat interface to start chatting, and the "patient" initiates the conversation. Second, annotators conduct a dialogue based on the dialogue topic sequence. It is important that the information utilized in the dialogue conforms to the reference knowledge. After successfully mentioning all target topics in sequence, the "doctor" ends the conversation.

Furthermore, we introduce the fine-grained privacy removing, the skipping option, quality audit and re-annotating to improve data quality, and introduce the annotation without target dialogue topic sequence mechanism to improve data diversity.

Fine-grained Privacy Removing. In the data annotation process, for better data quality, annotators are also required to delete privacy that cannot be covered by regular expressions, including gender, age, name, institution name, etc.

Skipping Option. We observe that there are many basic diagnosis dialogues with low quality. These bad dialogues may lead to annotated dialogues of low quality. To alleviate the issue, a skip option is provided to annotators. Specifically, annotators can choose whether to annotate the given basic diagnosis dialogue or not to the quality of the given dialogue. If annotators choose "Skip", they then skip the current dialogue directly and conduct the annotation of the next dialogue.

To ensure the option is not being overused, we review all the skipped conversations and select high-quality dialogues from the skipped conversations. Those high-quality dialogues are returned to the annotation process, and the rest low-quality dialogues are abandoned.

Quality Audit and Re-annotating. To deal with low-quality samples, we introduce the quality audit and re-annotation mechanism. Specifically, we review all the annotated samples and pick out low-quality dialogues. These low-quality samples are returned to the annotation team for re-annotation.

# of dialogues	8,175
- Otorhinolaryngology	1,692
- Ophthalmology	1,443
- Skin	2,962
- Digestive System	2,078
# of dialogues w/ goal	7,557
# of dialogues w/o goal	752
Avg. # of utterances in a dialogue	11.79
- Otorhinolaryngology	12.20
- Ophthalmology	11.02
- Skin	12.04
- Digestive System	11.64
Max. # of utterances in a dialogue	46
Min. # of utterances in a dialogue	6
# of tokens	1,887,227
Avg. # of tokens in an utterance	19.26
Max. # of tokens in an utterance	189
Min. # of tokens in an utterance	2

Table 2: Statistics of the MidMed.

Annotation without Target Dialogue Topic Sequence. Though the target dialogue topic sequences lead to good annotation quality, they usually lead to monotonous dialogue structures. To address the issue, annotators are also allowed to construct the dialogues without following the target dialogue topic sequences. This option enables annotators to construct more diverse and flexible dialogues based on the basic diagnosis dialogues. Meanwhile, to prevent this option from being abused, this option is required to be used for no more than ten percent of the whole annotation data.

3.4 Dataset Analysis

Data statistics. Table 2 provides statistics of the MidMed. There are totally 8,175 dialogues with 11.79 utterances in each dialogue on average. The longest dialogue contains 46 utterances. Besides, there are 19.26 tokens in an utterance on average, indicating rich semantic information.

Table 1 lists medical dialogue datasets (MZ (Wei et al., 2018), DX (Xu et al., 2019), CMDD (Lin et al., 2019), MedDG (Liu et al., 2022a), MedDialog (Zeng et al., 2020), DialoAMC (Chen et al., 2022)) and mixed-type dialogue dataset(DuRecDial (Liu et al., 2020), DodecaDialogue (Shuster et al., 2020), BlendedSkillTalk (Smith et al., 2020), ACCENTOR (Sun et al., 2021), DuRecDial 2.0 (Liu et al., 2021), SalesBot (Chiu et al., 2022), DuClarifyDial (Liu et al., 2022b)). MidMed is the first dialogue dataset for consultation, covering five types of dialogues.

Data quality. Following (Liu et al., 2020), for data quality evaluation, we employ human evaluations. Specifically, we assign "1" for dialogues coincident with annotation guidance, and "0" for the others. Then, we conduct a quality evaluation on 100 randomly sampled dialogues. Finally, an average score of "0.90" is achieved. The result indicates that the dialogues in the dataset are with high quality.

4 Method

During training, a dialogue, with a sequence of utterances between a patient and a doctor, is given. Then, the dialogue is processed into a set of samples $\{(\boldsymbol{s}_i, \boldsymbol{t}_i)\} \in \mathcal{D}$, where $\boldsymbol{t}_i$ is i-th target doctor response, $\boldsymbol{s}_i$ is the concatenation of all former utterances before $\boldsymbol{t}_i$, and $\mathcal{D}$ is the training dataset. Dialogue generation is formulated as a sequence-to-sequence generation problem, which aims to generate $\boldsymbol{t}_i$ conditioned on $\boldsymbol{s}_i$.

InsMed has three modules, dialogue topic selecting, reference knowledge selection, and the instruction-guided generation module. The dialogue topic prediction and the reference knowledge selection module aim to obtain dialogue topics and reference knowledge, respectively. Then, for better generation performance, these two types of information are transformed into instructions in natural language. Finally, instructions are concatenated with context, as the input to generation models. Next, the above modules are introduced.

4.1 Dialogue Topic Selection

The dialogue topic selection module is divided into two stages, the dialogue topic prediction, and the dialogue topic converting.

The dialogue topic prediction aims to predict dialogue topics for the next utterance. Formally, this task is regarded as a multi-class classification problem. Specifically, the input of the prediction module is a dialogue context $\boldsymbol{s}_i$, and the output is the predicted dialogue topics. The classification process is formulated,

$$\boldsymbol{p}_i = f(\boldsymbol{s}_i),$$

where f is the classification function BERT (Devlin et al., 2018) and $\boldsymbol{p}_i \in |\mathcal{R}|^{|\mathcal{C}|}$ is the predicted probability value, $\mathcal{C}$ is the predefined category set. The dialogue topic a_i is selected as the predicted dialogue topic if the value of the dimension is the highest probability value in $\boldsymbol{p}_i$.

Then, in the dialogue topic converting stage, a_i is converted into natural language with predefined templates, represented as $\tilde{a}_i$. For example, the predicted topic is `Recommend Medicine`, and the converted instruction is "*In the next utterance, the doctor will recommend medicine*".

4.2 Reference Knowledge Selection

The reference knowledge selection module aims to obtain the reference knowledge for model generation, thus guiding models to generate more informative responses. The module is divided into two parts, knowledge retrieval, and reference knowledge converting.

The knowledge retrieval module aims to retrieve reference knowledge from the whole knowledge graph for response generation. An exact string match is utilized for retrieval. Specifically, the diseases $\boldsymbol{d}_{i=1}^{m}$ in the whole knowledge graph are mapped with medical dialogues with exact string matching, where m is the number of diseases. The disease d_i existing in the medical dialogues are regarded as related diseases of the dialogues. Then, the reference knowledge is obtained by inquiry the knowledge graph with d_i, $\boldsymbol{e} = \{tail|\{head, relation, tail\} \in KG, head = \boldsymbol{d}_i, relation = r\}$, where r is the slot in the predicted dialogue topic a_k. For example, if the dialogue topic is `Doctor Recommend Medicine`, r is `Medicine`, and $\boldsymbol{e} = \{\text{bonmopirocin ointment}, \text{dexamethasone cream}\}$.

Then, in the reference knowledge converting, $\boldsymbol{e}$ is converted into natural language with predefined templates, represented as $\tilde{e}_i$. As the example in Figure 3, the converted knowledge instruction is "*the recommended medicine is bonmopirocin ointment and dexamethasone cream*".

4.3 Instruction-guiding Generation

The Instruction-guiding generation module aims to generate accurate and informative responses with instructions.

The problem of response generation is formulated as a sequence-to-sequence task (Sutskever et al., 2014). The input to the generation model is the concatenation of the dialogue context $\boldsymbol{s}_i$, the predicted dialogue topic instruction $\tilde{a}_i$, and the reference knowledge $\tilde{e}_i$. The output is the doctor's response $\boldsymbol{t}_i$.

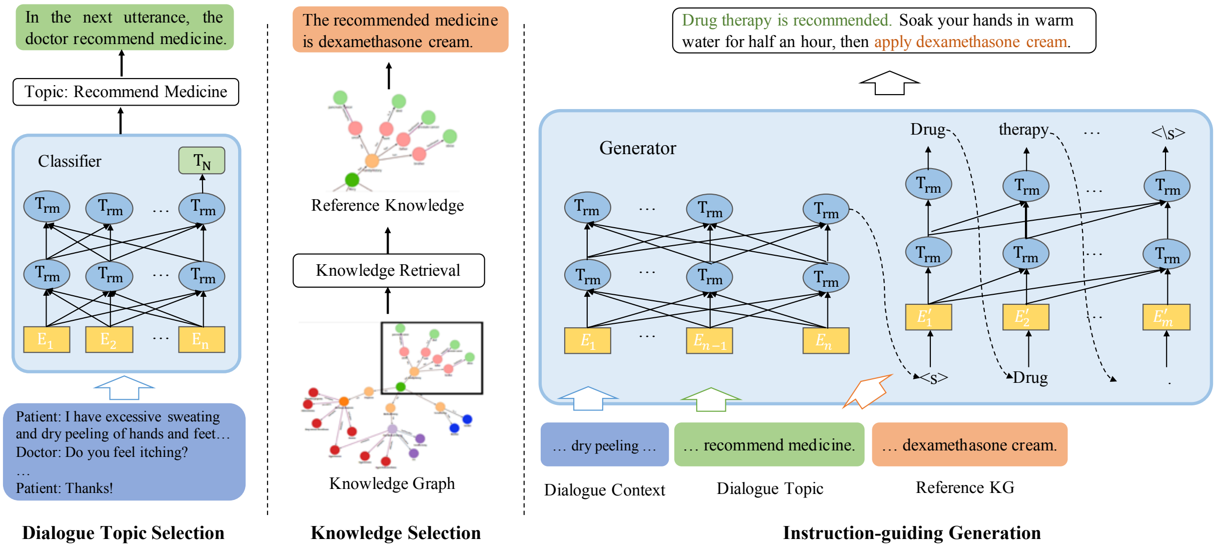

In the next utterance, the doctor recommend medicine.
Topic: Recommend Medicine
Classifier
T_N
Trm
E1
E2
En
Patient: I have excessive sweating and dry peeling of hands and feet…
Doctor: Do you feel itching?
…
Patient: Thanks!
Dialogue Topic Selection
The recommended medicine is dexamethasone cream.
Reference Knowledge
Knowledge Retrieval
Knowledge Graph
Knowledge Selection
Drug therapy is recommended. Soak your hands in warm water for half an hour, then apply dexamethasone cream.
Generator
Drug
therapy
</s>
E'1
E'2
E'm
En−1
<s>
… dry peeling …
… recommend medicine.
… dexamethasone cream.
Dialogue Context
Dialogue Topic
Reference KG
Instruction-guiding Generation

Figure 3: The illustration of the proposed InsMed.

BART (Lewis et al., 2020) is utilized as the generation model. Then, the forward calculation process is formulated,

$$\boldsymbol{t}_i = f_g([\boldsymbol{s}_i; \tilde{a}_i; \tilde{e}_i]),$$

where f_g represents the generation model BART.

5 Experiments and Results

This section introduces experimental setting, data and evaluation metrics, baselines, automatic evaluations, human evaluations, and the ablation study.

5.1 Experimental Setting

Implementation Details. For Transformer, the implementation by HuggingFace [3] is utilized, where the hyperparameters follow the default settings in the original Transformer (Vaswani et al., 2017).

For DialoGPT-small (Zhang et al., 2018), the layer number, the embedding size, and the context size are set as 10, 768, and 300, respectively. In layer normalization, the epsilon hyperparameter is set as 1e-5. In multi-head self-attention, the number of heads is set as 12. The weight parameters are learned with Adam, with the initial learning rate 1.5e-4 and the batch size 32.

For BERT classifier, we use a mini-batch size of 64 and the Adam optimizer with default parameters (a fixed learning rate 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times e^{-8}$) (Kingma and Ba, 2015).

For BART, the large version is employed, with the learning rate $2 \times e^{-5}$. In BART, the BERT encoder and GPT decoder are Transformers with 12 layers and a hidden state size of 768. The dropout rate is set as 0.1. The maximum length of input sequences is truncated to 512 and that of output sequences was truncated to 256.

Computing Platform. Our experiments are conducted on the workstation with an Intel Xeon E5 2.40 GHz CPU, 128 GB memory, an NVIDIA A100 GPU, and CentOS 7.2.

5.2 Data and Evaluation Metrics

We split MidMed into the training set, the validation set, and the test set by randomly sampling 70%, 10%, and 20% data.

5.2.1 Automatic Evaluation Metrics

Following Zeng et al. (2020), four basic automatic evaluation metrics for generation tasks are utilized in this work, including ROUGE (Lin, 2004), NIST-4 (Doddington, 2002), BLEU-n (Papineni et al., 2002) (where n is the size of n-gram), and METEOR (Agarwal and Lavie, 2007). These metrics all measure the similarity between the generated responses and the ground truth via n-gram matching.

5.2.2 Human Evaluation Metrics

Following Liu et al. (2020), three human evaluation metrics are utilized in this work, including relevance, informativeness, and human-likeness.

Relevance measures fluency, relevancy and logical consistency of each response when given the current goal and global context:

- score 0 (bad): more than two-thirds responses irrelevant or logical contradictory to the given current goal and global context.

---

[3] https://github.com/huggingface/transformers

- score 1 (fair): more than one-third responses irrelevant or logical contradictory to the given current goal and global context.
- score 2 (good): otherwise.

Informativeness examines how much knowledge (goal topics and topic attributes) is provided in responses:

- score 0 (bad): no knowledge is mentioned at all.
- score 1 (fair): only one knowledge triple is mentioned in the response.
- score 2 (good): more than one knowledge triple is mentioned in the response.

Human-likeness examines similarity between each generated response with corresponding human response from the perspectives of appropriateness, fluency, and proactivity:

- score 0 (bad): not like human responses.
- score 1 (fair): like human responses, but some parts still have deficiencies.
- score 2 (good): otherwise.

5.3 Baselines

We carefully select a few strong baselines for comparison. Specifically, two baselines for mixed-type dialogue generation (BST (Smith et al., 2020), MGCG (Liu et al., 2020)), a baselines for medical dialogue generation (VRbot (Li et al., 2021)), two common baselines for medical dialogue (Seq2Seq (Sutskever et al., 2014), DialoGPT (Zhang et al., 2020)), and a baseline for general dialogue generation (BART (Lewis et al., 2020)) are used in this experiment. Besides, the proposed model utilizes the same data as these baselines, with domain-specific knowledge.

BST (Smith et al., 2020) is a mixed-type dialogue model that can display many skills, and blend them in a seamless and engaging way.

MGCG (Liu et al., 2020) consists of a goal-planning module and a goal-guided responding module. The goal-planning module conducts dialog management to control the dialog flow. The responding module generates responses for completing each goal.

VRbot (Li et al., 2021) introduces both patient state and physician action as latent variables with categorical priors for explicit patient state tracking and physician policy learning, respectively. A variational Bayesian generative approach is utilized to approximate posterior distributions over patient states and physician actions.

Seq2Seq (Sutskever et al., 2014) Sutskever et al. (2014) uses a multilayered Long Short-Term Memory (LSTM) to map the input sequence to a vector of fixed dimensionality, and then another LSTM to decode the target sequence from the vector.

DialoGPT (Zhang et al., 2020) is a large, tunable neural conversational response generation model based on GPT. DialoGPT is trained on 147M conversation-like exchanges extracted from Reddit comment chains over a period spanning from 2005 through 2017.

BART (Lewis et al., 2020) is a denoising autoencoder for pretraining sequence-to-sequence models. It is composed of a BERT encoder (a bidirectional encoder) and a GPT decoder (a left-to-right decoder).

5.4 Automatic Evaluation

The results on automatic evaluation metrics are shown in Table 3. InsMed is compared with the other five generation models on various evaluation metrics. The results show the following conclusions.

First, BART (large) is much better than other baseline generation models. The reason may be that BART (large) is much more powerful than other generation models, with more parameters and more training data.

Second, InsMed achieves state-of-the-art performance on almost all metrics. This demonstrates that instructions help BART to generate more accurate responses.

5.5 Human Evaluation

Table 4 shows the human evaluation results on the test set of MidMed.

First, comparing BART, InsMed with other baselines, the results demonstrate that pre-training on large-scale data improves relevance, informativeness, and human-likeness. The reason may be that pre-training on large-scale data provides a large amount of common language knowledge.

	ROUGE	NIST-4	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
BST (Smith et al., 2020)	13.64	0.81	2.88	14.01	4.89	2.15	1.02	13.81
MGCG (Liu et al., 2020)	14.37	0.98	3.36	15.88	5.39	2.61	1.06	15.24
VRbot (Li et al., 2021)	23.01	1.41	4.84	22.67	8.07	3.55	1.31	18.66
Seq2Seq (Sutskever et al., 2014)	12.21	0.77	2.93	14.25	4.92	2.08	1.01	13.66
DialoGPT (Zhang et al., 2020)	19.58	1.14	4.62	17.64	5.97	2.84	1.53	17.36
BART (Lewis et al., 2020)	31.63	3.12	21.95	41.36	27.26	22.04	18.87	34.74
InsMed (Ours)	**40.59**	**3.30**	**23.13**	**42.61**	**28.46**	**23.00**	**19.73**	**36.45**
w/o Topic	34.99	3.17	22.41	42.03	27.97	22.60	19.26	35.26
w/o KG	32.22	3.14	22.19	41.13	27.21	22.12	18.91	34.80

Table 3: Automatic evaluation results of five baseline models and InsMed, on eight evaluation metrics. Values of ROUGE, BLEU, and METEOR are expressed as percentages (%).

	BST	MGCG	VRbot	Seq2Seq	DialoGPT	BART	InsMed (Ours)	Groundtruth
Relevance	0.33	0.39	0.42	0.27	0.50	1.32	1.42	1.98
Informativeness	0.29	0.31	0.36	0.24	0.46	1.30	1.54	2.00
Human-likeness	0.37	0.41	0.48	0.32	0.70	1.68	1.88	2.00

Table 4: Human evaluation results of six baseline models, InsMed, and Groundtruth, on three aspects, including relevance, informativeness, and human-likeness. Scores of “0”, “1”, and “2” are assigned to each dialogue, where “0” represents bad samples and “2” represents good samples. The average scores are reported.

Second, comparing InsMed with BART, the results show that InsMed performs better than BART, especially in relevance and informativeness. The reason may be that instructions in InsMed provide specific targets for generation, leading to a more relevant and informative response generation.

5.6 Ablation Study

Table 3 shows the ablation results, where “w/o Topic” means removing dialogue topic instructions from the InsMed and “w/o KG” means removing reference knowledge instructions from the InsMed. Results show that reducing any module of MidMed leads to poor results. This illustrates the effectiveness of each module of the InsMed.

6 Conclusion

This work identified the challenge of helping patients clarify their goals through medical consultations. To address this challenge, this work proposed a novel task, medical consultation over mixed-type dialogue, and collected a new Chinese human-to-human mixed-type dialogue dataset, in which each session has rich variability of dialogue types with natural topic transitions. To facilitate further research, we conducted benchmarking experiments on MidMed for end-to-end dialogue generation and proposed an instruction-guiding medical dialogue generation framework InsMed. Experimental results show the effectiveness of InsMed. In the future, we will investigate the possibility of cross-departments (e.g. dermatology and endocrinology) medical consultation at low cost.

7 Limitation

InsMed is built based on the large-scale pre-training model BART, which requires high computing resources. Besides, the data currently only covers four departments, limiting the usage scenarios of the data.

8 Ethical Statement

We make sure that MidMed is collected in a manner that is consistent with the terms of use of any sources and the intellectual property and privacy rights of the original authors of the texts. And crowd workers were treated fairly. This includes, but is not limited to, compensating them fairly, ensuring that they were able to give informed consent, and ensuring that they were voluntary participants who were aware of any risks of harm associated with their participation.

9 Acknowledgements

Thanks for the insightful comments from reviewers. This work is supported by the Shanghai Artificial Intelligence Laboratory.

References

Abhaya Agarwal and Alon Lavie. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of WMT-08*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Zhongyu Wei, et al. 2022. A benchmark for automatic medical consultation system: Frameworks, tasks and datasets. *arXiv preprint arXiv:2204.08997*.

Ssu Chiu, Maolin Li, Yen-Ting Lin, and Yun-Nung Chen. 2022. SalesBot: Transitioning from chit-chat to task-oriented dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6143–6158, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *In Proc. of NAACL-HLT*, page 4171–4186.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.

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*, pages 7871–7880, Online. 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*.

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 preprint arXiv: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.

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.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022a. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Zeming Liu, Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, and Hua Wu. 2022b. Where to go for the holidays: Towards mixed-type dialogs for clarification of user goals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1024–1034, Dublin, Ireland. Association for Computational Linguistics.

Andrea Madotto, Zhaojiang Lin, Yejin Bang, and Pascale Fung. 2021. The adapter-bot: All-in-one controllable conversational model. In *AAAI*.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, Jamin Shin, and Pascale Fung. 2020. Attention over parameters for dialogue systems.

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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and et al. 2021.

Recipes for building an open-domain chatbot. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze questions for few shot text classification and natural language inference. *Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–269.

Hannah L Semigran, Jeffrey A Linder, Courtney Gidengil, and Ateev Mehrotra. 2015. Evaluation of symptom checkers for self diagnosis and triage: audit study. *BMJ*, 351:h3480.

Kurt Shuster, Da JU, Stephen Roller, Emily Dinan, Y-Lan Boureau, and Jason Weston. 2020. The dialogue dodecathlon: Open-domain knowledge and image grounded conversational agents. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Kai Sun, Seungwhan Moon, Paul A. Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *NAACL*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Yi-Lin Tuan, Sajjad Beygi, Maryam Fazel-Zarandi, Qiaozi Gao, Alessandra Cervone, and William Yang Wang. 2022. Towards large-scale interpretable knowledge graph reasoning for dialogue systems. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 383–395, Dublin, Ireland. Association for Computational Linguistics.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. 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 1855–1865, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zifeng Wang, Rui Wen, Xi Chen, Shilei Cao, Shao-Lun Huang, Buyue Qian, and Yefeng Zheng. 2021. Online disease diagnosis with inductive heterogeneous graph convolutional networks. In *Proceedings of the Web Conference 2021*, pages 3349–3358.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. *In Proc. of ACL*, pages 201–207.

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 Proc. of AAAI*.

Shiquan Yang, Rui Zhang, and Sarah Erfani. 2020. GraphDialog: Integrating graph knowledge into end-to-end task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1878–1888, Online. Association for Computational Linguistics.

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.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, 31.

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

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1 and 2

☑ A4. Have you used AI writing assistants when working on this paper?
Grammarly

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
1 and 3

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ 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)?
Not applicable. Left blank.

☐ 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?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
3

☑ 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.
3

C ☑ Did you run computational experiments?

5

☑ 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

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

☑ 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?
5

☑ 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.)?
5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☑ 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.?
3

☑ 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)?
3

☑ 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?
3

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
3

☐ 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.