ReMeDi: Resources for Multi-domain, Multi-service, Medical Dialogues

Guojun Yan
Shandong University
Qingdao, China
yan_gj@mail.sdu.edu.cn

Jiahuan Pei
University of Amsterdam
Amsterdam, The Netherlands
j.pei@uva.nl

Pengjie Ren
Shandong University
Qingdao, China
renpengjie@sdu.edu.cn

Zhaochun Ren
Shandong University
Qingdao, China
zhaochun.ren@sdu.edu.cn

Xin Xin
Shandong University
Qingdao, China
xinxin@sdu.edu.cn

Huasheng Liang
WeChat Tencent
Shenzhen, China
watsonliang@tencent.com

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Zhumin Chen*
Shandong University
Qingdao, China
chenzhumin@sdu.edu.cn

ABSTRACT

Medical dialogue systems (MDSs) aim to assist doctors and patients with a range of professional medical services, i.e., diagnosis, treatment and consultation. The development of MDSs is hindered because of a lack of resources. In particular. (1) there is no dataset with large-scale medical dialogues that covers multiple medical services and contains fine-grained medical labels (i.e., intents, actions, slots, values), and (2) there is no set of established benchmarks for MDSs for multi-domain, multi-service medical dialogues.

In this paper, we present ReMeDi, a set of resources for Chinese medical dialogues. ReMeDi consists of two parts, the ReMeDi dataset and the ReMeDi benchmarks. The ReMeDi dataset contains 96,965 conversations between doctors and patients, including 1,557 conversations with fine-gained labels. It covers 843 types of diseases, 5,228 medical entities, and 3 specialties of medical services across 40 domains. To the best of our knowledge, the ReMeDi dataset is the only medical dialogue dataset that covers multiple domains and services, and has fine-grained medical labels.

The second part of the ReMeDi resources consists of a set of state-of-the-art models for (medical) dialogue generation. The ReMeDi benchmark has the following methods: (1) pretrained models (i.e., BERT-WWM, BERT-MED, GPT2, and MT5) trained, validated, and tested on the ReMeDi dataset, and (2) a self-supervised contrastive learning (SCL) method to expand the ReMeDi dataset and enhance the training of the state-of-the-art pretrained models.

We describe the creation of the ReMeDi dataset, the ReMeDi benchmarking methods, and establish experimental results using the ReMeDi benchmarking methods on the ReMeDi dataset for future research to compare against. With this paper, we share the dataset, implementations of the benchmarks, and evaluation scripts.

*Corresponding author.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/3477495.3531809

CCS CONCEPTS

• **Applied computing** → **Health care information systems**.

KEYWORDS

Dialogue dataset, Dialogue benchmarks, Medical dialogues

ACM Reference Format:
Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke and Zhumin Chen. 2022. ReMeDi: Resources for Multi-domain, Multi-service, Medical Dialogues. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3477495.3531809

1 INTRODUCTION

Medical research with AI-based techniques is growing rapidly [10, 63, 66, 79]. Medical dialogue systems (MDSs) promise to increase access to healthcare services and to reduce medical costs [31, 76, 78]. MDSs are more challenging than common task-oriented dialogue systems (TDSs) for, e.g., ticket or restaurant booking [33, 50, 70] in that they require a great deal of expertise. For instance, there are much more professional terms, which are often expressed in colloquial language [61].

Recently, extensive efforts have been made towards building data for MDS research [34, 61]. Despite these important advances, limitations persist: (1) In currently available datasets, there is a lack of a complete diagnosis and treatment procedure. A practical medical dialogue is usually a combination of consultation, diagnosis and treatment, as shown in Figure 1. To the best of our knowledge, no previous study considers all three medical services simultaneously [39, 68, 74, 76]. (2) In currently available datasets, labels are not comprehensive enough. Most datasets only provide the slot-value pairs for each utterance. Intent labels and medical knowledge triples related to each utterance are rarely provided.

Figure 1: A realistic medical dialogue involving diagnosis, treatment and consultation. They are all dependent. Combined with the knowledge triple in the upper right corner, we can better infer the related diseases. The lower right part is our annotation example, including intent/action, slot and value.

For example, there is one utterance in [80]: "Patient: Doctor, could you please tell me is it premature beat?" It only has the slot-value label "Symptom: Cardiopalmus", without the intent label "Inquire" and the required knowledge triple "<premature beat, symptom, cardiopalmus>". (3) In currently available datasets, labels are not fine-grained enough. Composite utterances, which contain more than one intent/action, are common in practice. For example, for the third utterance in Figure 1, the patient says "Ten days. Yes. What is the disease?", there are three kinds of intents: informing time, informing symptom status, and inquiring diseases. Previous studies usually provide a single coarse-grained label for the whole composite utterance, which might mislead the training of models and/or lead to inaccurate evaluation. Second, we find that the values defined in previous work can hardly accurately convey complex information. Instead, we provide main-subordinate values, each of which includes a main value and a subordinate value. For example, for the labeling "Value=duration, ten days" of the second user utterance in Figure 1, the main value is "duration" and the subordinate value is "ten days". The main-subordinate values have a stronger capacity to convey complex information: (a) Negation status of an entity, e.g., without experiencing symptom sore throat. (b) The specific value of an entity, e.g., the specific number of blood pressure. (c) Relationship between entities, e.g., the side effect of a medicine. (4) Besides the limitations above, some datasets only involve limited medical entities. For example, MedDG [78], a very recent medical dialogue dataset, only contains 12 diseases.

To address the lack of a suitable dataset, our first contribution in this paper is the introduction of the *resources for Chinese medical dialogues* (ReMeDi) dataset. The ReMeDi dataset has the following features: (1) medical dialogues for consultation, diagnosis and treatment, as well as their mixture; (2) comprehensive and fine-grained labels, e.g., intent-slot-value triples for sub-utterances; and (3) more than 843 diseases, 20 slots and 5,228 medical entities are covered. Moreover, we ground the dialogues with medical knowledge triples by mapping utterances to medical entities.

Our second contribution in this paper is a set of medical dialogue models for benchmarking against the ReMeDi dataset. Recent work considers MDSs as a kind of TDS [34, 68, 74] by decomposing a MDS system into well-known sub-tasks, e.g., natural language understanding (NLU) [61], dialogue policy learning (DPL) [68], and natural language generation (NLG). There is, however, no comprehensive analysis on the performance of all the above tasks when achieved and/or evaluated simultaneously. To establish a shared benchmark that addresses all three NLU, DPL, and NLG tasks in a MDS setting, we adopt causal language modeling, use several pre-trained language models (i.e., BER-WWM, BERT-MED, MT5 and GPT2) and fine-tune them with the ReMeDi dataset. In addition, we provide a pseudo labeling algorithm and a natural perturbation method to expand the proposed dataset, and enhance the training of state-of-the-art pretrained models based on self-supervised contrastive learning.

Below, we detail the construction of the ReMeDi dataset and the definition of the ReMeDi benchmarks, and evaluate the ReMeDi benchmarks against the ReMeDi dataset on the NLU, DPL, and NLG tasks, thereby establishing a rich set of resources for medical dialogue system to facilitate future research. Details on obtaining the resources are included in the appendix of this paper.

2 RELATED WORK

We survey related work in terms of datasets, models and contrastive learning.

2.1 Medical dialogue datasets

Most medical dialogue datasets contain only one domain, e.g., Pediatrics [34, 68, 74], COVID-19 [76], Cardiology [80], Gastroenterology [39] and/or one medical service, e.g., Diagnosis [36, 37], Consultation [61]. However, context information from other services and/or domains is often overlooked in a complete medical aid procedure. For example, in Figure 1, the symptom "sore throat" mentioned in the diagnosis service has the long-term effect on the suggestion "talk less" in the follow-up consultation service. To this end, we provide medical dialogues for consultation, diagnosis and treatment, as well as their mixture in the ReMeDi dataset. Although a few datasets [31, 78] contain multiple medical services in multiple domains, they target the NLG only without considering the NLU and DPL. Differently, ReMeDi contains necessary labels for NLU, DPL and NLG. Another challenge of existing datasets is the medical label insufficiency problem. The majority of datasets only provide a spot of medical labels for slots or actions, e.g., one slot [14, 37, 61], single-value [31, 36, 39]. Moreover, their labels are too coarse to distinguish multiple intents or actions in one utterance. Unlike all datasets above, our dataset provides comprehensive and fine-grained intent/action labels for constituents of an utterance.

To sum up, ReMeDi is the first multiple-domain multiple-service medical dialogue dataset with fine-grained medical labels and large-scale entities, which is more competitive compared with the datasets mentioned above in terms of 9 aspects (i.e., domain, service, task,

intent, slot, action, entity, disease, dialogue). A summary can be found in Table 1.

2.2 Pretrained models for task-oriented dialogue systems

Large language models have achieved state-of-the-art performance for TDSs [1, 81]. BERT [13] is widely used as a benchmark for TDSs [41, 82] and has been shown to be effective for understanding and tracking dialogue states [5, 26]. In terms of dialogue generation, BERT is usually used in a selective way (e.g., TOD-BERT [71]). The GPT family of language models [52, 53] serves as a competitive and common benchmark in recent work on TDSs [3, 21, 77]. GPT is used as a promising backbone of recent research on generating dialogue states [4, 42, 71] and actions [32]. MT5 [54] is the current benchmark for TDSs, because it inherits T5 [55]'s powerful capabilities of text generation and provides with multilingual settings [38, 40, 83]. Large neural language models are data hungry and data acquisition for TDSs is expensive [56]. An effective method for alleviating this issue is contrastive learning (CL). CL compares similarity and dissimilarity by positive/negative data sampling, and defining contrastive training objectives [23, 27]. Most studies work on re-optimizing the representation space based on contrastive word [20, 22, 43, 49] or sentence [16, 18, 25, 60, 67] pairs. Some also focus on sampling negative data pairs [28, 64, 65, 73]. Research has also explored different contrastive training objectives based on single [11, 19] or multiple [9, 59, 62] positive and negative pairs, along with their complex relations [17, 45, 58].

In this work, we share several pretrained language models based on BERT, GPT2, MT5 as benchmarks of ReMeDi. To alleviate the data hungry problem, we enhance the pretrained language models with contrastive learning.

2.3 Medical dialogue systems

Similar to TDSs [7], a MDS system can be divided into several sub-tasks, e.g., NLU, DPL, and NLG.

NLU aims to understand user utterances by intent detection [68] and slots filling [8, 51, 69]. Du et al. [14, 15] formulate NLU as a sequence labeling task and use Bi-LSTM to capture contextual representation for filling entities and their relations into slots. Lin et al. [37] improve filling entities with global attention and symptom graph. Shi et al. [61] propose the label-embedding attentive multi-label classifier and improve the model by weak supervision from responses. dialogue state tracking (DST) tracks the change of user intent [44]. Zhang et al. [80] employ a deep matching network, which uses a matching-aggregate module to model turn-interaction among utterances encoded by Bi-LSTM. In this work, we integrate DST into vanilla NLU to generate intents and updated slot values simultaneously.

DPL decides system actions given a set of slot-value dialogue states and/or a dialogue context [7]. Wei et al. [68] first adopt reinforcement learning (RL) to extract symptoms as actions for disease diagnosis. Xu et al. [74] apply deep Q-network based on a medical knowledge graph to track topic transitions. Xia et al. [72] improve RL based DPL using generative adversarial learning with regularized mutual information. Liao et al. [34] use a hierarchical RL model to alleviate the large action space problem. We generate system actions as general tokens to fully avoid action space exploration in these RL models.

NLG generates system responses given the outputs from NLU and DPL [47]. Yang et al. [76] apply several pretrained language models (i.e., Transformer, GPT, and BERT-GPT) to generate doctors' responses for COVID-19 medical services. Liu et al. [39] provide several NLG baselines based on sequence-to-sequence models (i.e., Seq2Seq, HRED) and pretrained language models (i.e., GPT2 and DialoGPT). Li et al. [30] use pretrained language models to predict entities and generate responses. Recently, meta-learning [36] and semi-supervised variational Bayesian inference [31] are adopted for low-resource medical response generation.

3 THE ReMeDi DATASET

The ReMeDi dataset is built following the pipeline shown in Figure 2: (1) We collect raw medical dialogues and knowledge base from online websites; (2) We clean dialogues by a set of reasonable rules, and sample dialogues by considering the proportions of disease categories; (3) We define annotation guidelines and incrementally improve them by dry-run annotation feedbacks until standard annotation guidelines are agreed by annotators; (4) We conduct human annotation with standard annotation guidelines.

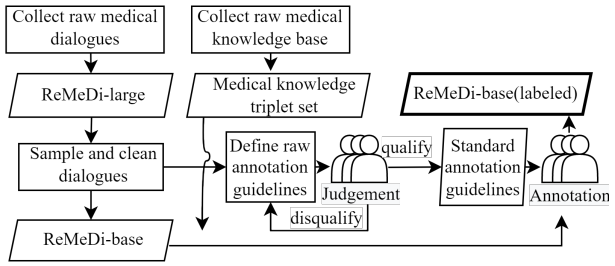

Figure 2: Process of the ReMeDi dataset construction.

Note that we provided two versions of the dataset: a labeled ReMeDi-base (1,557 dialogues) and an unlabeled ReMeDi-large (95,408 dialogues). The former is for evaluating the performance of the benchmark models and the latter is for improving the training of large models (see details in §4.4.3).

3.1 Collecting raw dialogues and knowledge base

We collect 95,408 natural multiple-turn conversations between doctors and patients from ChunYuYiSheng,[1] a Chinese online medical community. All information from the website is open to the public and has been processed with due care for privacy concerns, e.g., the sensitive information of patients, such as names, has been anonymized. To further ensure data privacy, we anonymize more potentially sensitive information, e.g., the name of doctors, the address of hospitals, etc. These raw dialogues constitute a large-scaled unlabeled dataset, called ReMeDi-large. It covers 40 domains (e.g., pediatrics), 3 services (i.e., diagnosis, consultation, and treatment), 843 diseases (e.g., upper respiratory tract infection), and 5,228 medical entities. We crawled 2.6M medical <entity1, relation, entity2>

[1] https://www.chunyuyisheng.com/

Table 1: Comparison between the proposed corpora and other medical dialogue corpora.

Dataset	(#)Domains	(#)Services	(#)Tasks	#Intents/Slots/Actions	#Entities	#Diseases	#Dialogues
CMDD [37]	Pediatrics	Diagnosis	NLU	- / 1 / -	162	4	2,067
SAT [14]	14	**3**	NLU	- / 1 / -	186	-	2,950
MSL [61]	Pediatrics	Consultation	NLU	- / 1 / -	29	5	1,652
MIE [80]	Cardiology	2	NLU	- / 4 / -	71	6	1,120
MZ [68]	Pediatrics	Diagnosis	DPL	- / 2 / 6	67	4	710
DX [74]	Pediatrics	Diagnosis	DPL	- / 2 / 5	41	5	527
RD [34]	Pediatrics	Diagnosis	DPL	- / 2 / 2	90	4	1,490
SD [34]	9	Diagnosis	DPL	- / 2 / 2	266	90	30,000
COVID-EN [76]	COVID-19	**3**	NLG	- / - / -	-	1	603
COVID-CN [76]	COVID-19	**3**	NLG	- / - / -	-	1	1,088
MedDG [39]	Gastroenterology	2	NLG	- / 5 / -	160	12	17,864
MedDialog-EN [78]	51	**3**	NLG	- / - / -	-	96	257,332
MedDialog-CN [78]	29	**3**	NLG	- / - / -	-	172	**3,407,494**
Chunyu [36]	-	Diagnosis	NLG	- / 2 / -	-	15	12,842
KaMed [31]	**100**	**3**	NLG	- / 4 / -	**5,682**	-	63,754
ReMeDi-base	30	**3**	**3**	**5/20/7**	4,825	491	1,557
ReMeDi-large	40	**3**	**3**	**5/20/7**	5,228	**843**	95,408

triplets from CMeKG2.0,[2] a Chinese medical knowledge base. For example, the triplet <paracetamol, indication, headache> denotes paracetamol can relieve headache. The entities involve about 901 diseases, 920 drugs, 688 symptoms, and 200 diagnosis and treatment technologies. The number of relation types is 125.

3.2 Cleaning and sampling dialogues

We conduct the following steps to obtain a set of dialogues for human annotation: (1) Filtering out noise dialogues. First, we filter out short-turn dialogues with less than 8 utterances, because we find these short dialogues usually do not contain much information. Next, we filter out inaccurate dialogues with images or audios and keep dialogues with literal utterances only. Finally, we filter out dialogues in which too few medical entities emerged in the crawled knowledge triplet set. (2) Anonymizing sensitive information. We use special tokens to replace sensitive information in raw dialogues, e.g., "[HOSPITAL]" is used to anonymize the specific name of a hospital. (3) Sampling dialogues by disease categories. In order to balance the distribution of diseases, we extract the same proportion of dialogues from each disease to form ReMeDi-base for annotation.

3.3 Incremental definition of annotation guidelines

We hire 15 annotators with the relevant medical background to work on the annotation process. We define 5 intents, 7 actions, and 20 slots and design a set of primer annotation guidelines. First, each annotator is asked to annotate 5 dialogues and then to report unreasonable, confusing, and ambiguous guidelines with corresponding utterances. Second, we summarize the confusing issues and improve the guidelines by a high agreement among annotators. We repeat the above two steps in three rounds and obtain a set of standard annotation guidelines.

3.4 Human annotation and quality assurance

To make the annotation more convenient, we build a web-based labeling system similar to [57], which is available online.[3] In the system, each annotator is assigned with 5 dialogues each round and is asked to label all utterances following the standard annotation guidelines. To assure annotation quality, we provide: (1) Detailed guidelines. For each data sample, we introduce the format of the data, the specific labeling task, the examples of various types of labels, and detailed system operations. (2) A real-time feedback paradigm. We maintain a shared document to track problems and solutions in real-time. All annotators can write questions on it; some dialogues with ambiguous labels will be managed: we discussed them with experts and gave the final decision. (3) A semi-automatic quality judgment paradigm. We adopt a rule-based quality judgment model to assist annotators in re-labeling the untrusted annotations. (4) An entity standardization paradigm. We use Levinstein distance ratio [29] to compute the similarity between an annotation and an entity in medical knowledge triplet. If a max similarity score is in [0.9,1], we ask the annotator to replace the annotation with a standard entity from the medical knowledge triplet.

3.5 Dataset statistics

Table 2 shows the data statistics. ReMeDi contains 95,408 unlabeled dialogues and 1,557 dialogues with sub-utterance-level semantic labels in the format of intent-slot-value or action-slot-value. ReMeDi-large is used for training, and ReMeDi-base is randomly divided into 657/100/800 dialogues for fine-tuning, validation, testing, respectively. ReMeDi-large has 40 domains and 843 diseases. ReMeDi-base has 30 domains and 491 diseases. In ReMeDi, about 70% of the dialogues involve multiple services.

Figure 3 shows the number of utterances in ReMeDi-base distributed over different types of intents/actions and slots. In the left chart, there are 5 patient intents (i.e., "Informing", "Inquiring",

[2] http://cmekg.pcl.ac.cn/

[3] https://github.com/yanguojun123/Medical-Dialogue

Table 2: Statistics of the ReMeDi dataset.

	Train	Fine-tune	Valid.	Test	Total
#Dialogue	95,408	657	100	800	96,965
#Utterance	1,753,624	10,642	1,718	13,086	1,779,070
#Utterance/dialogue	18.38	16.50	17.18	16.36	18.35
#Character/dialogue	302.29	311.80	332.63	302.52	316.50
#Character/utterance	16.18	19.25	19.36	19.46	16.23
#Label/dialogue	18.68	29.85	31.01	29.85	18.86
#Label/utterance	1.0	1.84	1.81	1.82	1.01

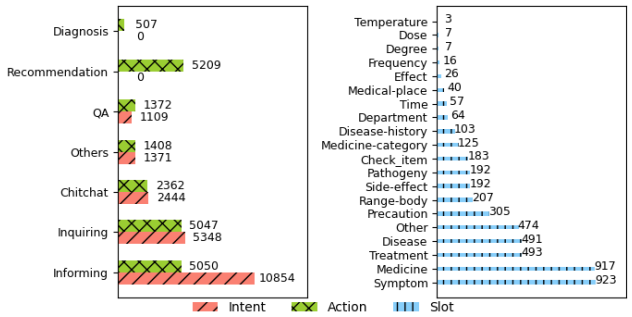

Figure 3: Distribution of utterances of ReMeDi-base containing different types of intents/actions (left) and slots (right), respectively.

"Chitchat", "QA" and "Others") and 7 doctor actions (including 5 intent types plus "Recommendation" and "Diagnosis"). These cover 25,446 utterances in total, and an utterance might contain multiple intents/actions. "Informing" account for the largest proportion (63%), while "Diagnosis" takes up the minimal proportion (2%). It shows that patients have a huge demand for online medical consultations, while doctors are very cautious to make online diagnosis. The right chart contains 20 types of slots covering 4,825 entities in total. "Symptom" (19%) has the largest proportion of entities, followed by "Medicine" (19%), "Treatment" (10%) and "Disease" (10%). In addition, 16% of label have subordinate value.

4 THE ReMeDi BENCHMARKS

In this section, we unify all tasks as a context-to-text generation task (§4.1). Then we introduce two types of benchmarks, i.e., causal language model (§4.2) and conditional causal language model (§4.3). Last, we introduce how to enhance models with CL to build the state-of-the-art benchmarks (§4.4).

4.1 Unified MDS framework

We view a MDS as a context-to-text generation problem [21, 48] and deploy a unified framework called SeqMDS. Formally, given a sequence of dialogue context X, a MDS aims to generate a system response Y which maximizes the generation probability $P(Y|X)$. Specifically, all sub-tasks are defined by the following formation.

The NLU part of SeqMDS aims to generate a list of intent-slot-value triplets I_t:

$$I_t = \text{SeqMDS}(C_t), \tag{1}$$

where dialogue history $C_t = [U_1, S_1, \ldots, U_t]$ consists of all previous utterances. And I_t can be used to retrieve a set of related knowledge triplets K from the knowledge base.

The DPL part of SeqMDS generates the action-slot-value pairs A_t given C_t, I_t, and K_t as an input:

$$A_t = \text{SeqMDS}([C_t; I_t; K_t]). \tag{2}$$

The NLG part of SeqMDS generates a response based on all previous information:

$$S_t = \text{SeqMDS}([C_t; I_t; K_t; A_t]). \tag{3}$$

SeqMDS in the above equations can be implemented by a causal language model (§4.2) or a conditional causal language model (§4.3).

4.2 Causal language model

We consider the concatenation $[C; I; K; A; S]$ as a sequence of tokens $X_{1:n} = (x_1, x_2, \ldots, x_n)$. The j-th element x_j can be an intent token (in intent-slot-value pairs), an action token (in action-slot-value pairs), or a general token (in utterances from patients or doctors). For the i-th sequence $X^i_{1:n}$, the goal is to learn the joint probability $p_\theta(X^i_{1:n})$ as:

$$p_\theta(X^i_{1:n}) = \prod_{j=1}^{n} (x^i_j | X^i_{0:j-1}). \tag{4}$$

The cross-entropy loss is employed to learn parameters θ:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{N}\sum_{j=1}^{n_i} x^i_j \log p_\theta(x^i_j | X^i_{0:j-1}), \tag{5}$$

where N denotes batch size and n_i denotes length of i-th utterance. In this work, we implement the causal language model based on GPT2 [53].

4.3 Conditional casual language model

We consider the concatenation $[C]$, $[C; I; K]$, $[C; I; K; A]$ as the input sequence $X_{1:n}$ and I, A, S as the generated sequence $Y_{1:m}$ in NLU, DPL and NLG, respectively.

For each input sequence, a transformer encoder is used to convert $X_{1:n} = (x_1, x_2, \ldots, x_n)$ to the corresponding hidden states $H_{1:n} = (h_1, h_1, \ldots, h_n)$, together with the current decoded tokens $Y_{1:j-1}$. A transformer decoder is used to learn the probability $p_\theta(Y_{1:m}|H_{1:n})$ over the vocabulary V at the j-th timestamp by:

$$p_\theta(Y_{1:m}|H_{1:n}) = \prod_{j=1}^{m} p_\theta(y_j | Y_{0:j-1}, H_{1:n}). \tag{6}$$

Similarly, the model can be learned by minimizing the cross-entropy loss as follows:

$$\mathcal{L}_{ce} = -\sum_{i=1}^{N}\sum_{j=1}^{n_i} y^i_j \log p_\theta(y^i_j | Y^i_{0:j-1}, H^i_{1:n}). \tag{7}$$

In this work, we implement the conditional causal language model based on MT5 [75].

4.4 Self-supervised contrastive learning

To extend upon the ReMeDi benchmark approaches introduced so far and enhance model training based on augmented data, we describe a self-supervised contrastive learning (SCL) approach. First, we generate data by two heuristic data augmentation approaches, i.e., pseudo labeling (§4.4.1) followed by natural perturbation (§4.4.2). Then, we adopt contrastive learning (§4.4.3) to assure the models are aware that the augmented data is similar to the original data.

4.4.1 Pseudo labeling. We propose a pseudo labeling algorithm to extend the unlabeled dialogues. As shown in Algorithm 1, we

Algorithm 1: Pseudo labeling.

Input : $D_L = \{(T_L^i, E_L^i)\}|_{i=1}^{|D_L|}$; $D - D_L$; R;
Output: $D_P = \{(T_P^i, E_P^i)\}|_{i=1}^{|D_P|}$;

foreach $T_P^i \in T_P$ **do**
  $\eta, e = \text{MaxSimilariy}(T_P^i, D_L)$
  if $\eta > \delta$ **then**
    $E_P^i \longleftarrow e$;
  else
    foreach $R^i \in R$ **do** Update E_P^i;
Function $\text{MaxSimilariy}(T_P^i, D_L)$**:**
  $\eta = 0; e = null; x = len(T_P^i); y = len(T_L^k)$;
  foreach $T_L^k \in T_L$ **do**
    $\hat{\eta} = 1 - \text{LevenshteinDistance}(x, y)/(x + y)$;
    if $\hat{\eta} > \delta$ **then**
      $\eta \longleftarrow \hat{\eta}; e \longleftarrow E_L^k$
  return η, e;

decompose the labeled dialogues and unlabeled dialogues into utterance sets D_L and D_P, respectively. Each element of D_L contains an utterance T_L (from user or system) and its corresponding semantic label E_L (in the format of intent-slot-value or action-slot-value). R is a set of predefined rules, e.g., if "take orally" is mentioned in some utterance, then the action is "Recommendation" and the slot is "Medicine". The output is D_P with pseudo labels E_P. The main procedure is as follows. For each utterance in D_P, we calculate the similarities between the current utterance T_P^i and all labeled utterances in D_L to get the maximum similarity η and the corresponding label e. If $\eta > \delta(\delta = 0.8)$, e is assigned as the pseudo label of T_P^i. Otherwise, each rule in R is applied to T_P^i to update E_P^i gradually. The similarity is deployed based on Levenshtein distance [29], which considers both the overlap rate and the order of characters.

4.4.2 Natural perturbation. We use three natural perturbation strategies to extend the labeled dialogues: (1) Alias substitution. If an utterance contains a drug with an alias, then the drug will be replaced with its alias to obtain a new data. For example, people from different regions may have different names for the same drug. (2) Back-translation. Chinese utterances are first translated into English and then back into Chinese to form a new data. Patients often use colloquial expressions, which motivates us to adopt back-translation to produce formal utterances from the informal ones. (3) Random modification. We randomly add, delete and replace a character of several medical entities in utterances. This simulates the common situation: typographical errors in online medical communities.

4.4.3 Contrastive Learning. We adopt an effective contrastive learning method to further increase the gains of natural perturbation. Following the CL framework [9], we efficiently learn an enhanced representation by contrasting the positive pairs with the negative pairs within a mini-batch.

Given an observed data (X^i, Y^i), we randomly sample one natural perturbation strategy to get the augmented data (X^j, Y^i). Let $z \in \mathbb{R}^d$ denote the sentence representation with d dimension. We construct the representation of observed and augmented data as a positive pair (z^i, z^j), and the representation of other data within the mini-batch as negative pairs $\{(z^i, z^k)\}_{k=1,k\neq i,k\neq j}^{2N}$. We compute the pairwise contrastive loss between the observed and augmented data:

$$l(i,j) = -\log \frac{\exp(sim(z^i, z^j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}^{[k\neq i]} \exp(sim(z^i, z^k)/\tau)}, \quad (8)$$
$$z^i = h_1^i,\ h_t^i = g(y_{t-1}^i, M^i; \theta),\ M^i = f(X^i; \theta),$$

where τ is temperature coefficient and $\mathbb{1}^{[k\neq i]} \in \{0, 1\}$ denotes an indicator function evaluating to 1 iff $k \neq i$. f, g denote the encoder and decoder respectively and θ are the model parameters. The function $sim(u, v) = u^\top v/\|u\|\|v\|$ computes cosine similarity.

For one batch, we minimize contrastive loss across positive pairs, for both (z^i, z^j) and (z^j, z^i):

$$\mathcal{L}_{cl} = \frac{1}{2N} \sum_{k=1}^{N} [l(2k-1, 2k) + l(2k, 2k-1)]. \quad (9)$$

We jointly learn CL loss with task-specific cross-entropy loss, and the final loss function is defined as:

$$\mathcal{L} = \lambda \mathcal{L}_{ce} + (1 - \lambda)\mathcal{L}_{cl}, \quad (10)$$

where λ is the coefficient to balance the two training losses.

5 EVALUATING THE ReMeDi BENCHMARKS AGAINST THE ReMeDi DATASET

In this section, we first list our evaluation settings, which includes 3 dialogue tasks, 5 benchmark models, 8 automatical metrics and 2 human evaluation metrics. Then we report on the results and detailed analysis of the ReMeDi benchmarks.

5.1 Tasks

The ReMeDi benchmarks address three tasks, NLU, DPL and NLG:

NLU aims to generate a list of intent-slot-value triplets given a dialogue context.

DPL aims to generate a list of action-slot-value triplets given a dialogue context and a list of intent-slot-value triplets.

NLG aims to generate a response given a dialogue context, intent-slot-value triplets and action-slot-value triplets.

5.2 ReMeDi benchmark models

We employ several pretrained models as benchmarks:

BERT-WWM is a BERT [13] pre-trained on a Chinese Wikipedia corpus [12].

BERT-MED is a BERT pre-trained on Chinese medical corpus.[4]
GPT2 is used as a transformer decoder for causal language modeling; we use one pre-trained on Chinese chitchat dialogues [53].[5]
MT5 is used as a transformer encoder-decoder model for conditional causality modeling. We use the one pre-trained on multilingual C4 dataset [75].[6]
MT5+CL is an extension of **MT5** with contrastive learning.

5.3 Evaluation setup

We consider two types of evaluation: automatic (for the NLU and DPL tasks) and human (for the NLG task). For the *automatic evaluation*, we use 4 metrics to evaluate the NLU and DPL tasks:

Micro-F1 is the intent/action/slot F1 regardless of categories.
Macro-F1 denotes the weighted average of F1 scores of all categories. In this work, we use the proportion of data in each category as the weight.
BLEU indicates how similar the generated values of intent/action slots are to the golden ones [6].
Combination is defined as $0.5 * \text{Micro-F1} + 0.5 * \text{BLEU}$. This measures the overall performance in terms of both intent/action/slot and the generated value.

We use 4 metrics to evaluate the NLG task:

BLEU1 and BLEU4 denote the uni-gram and 4-gram precision, indicating the fraction of the overlapping n-grams out of all n-grams for the responses [6].
ROUGE1 refers to the uni-grams recall, indicating the fraction of the overlapping uni-grams out of all uni-grams for the responses [2].
METEOR measures the overall performance, i.e., harmonic mean of the uni-gram precision and recall [35].

For the NLG task, we sample 300 context-response pairs to conduct the *human evaluation*. We ask annotators to evaluate each response by choosing a score from 0, 1, 2, which denotes bad, neutral, good, respectively. Each data sample is labeled by 3 annotators. We define 2 human evaluation metrics:

Fluency measures to what extent the evaluated responses are fluent.
Specialty measures to what extent the evaluated responses provide complete and accurate entities compared with the reference responses.

5.4 Outcomes

In this section, we report the results of the ReMeDi benchmark models (§5.2) on the NLU, DPL, NLG tasks, respectively. Please note that BERT treats NLU and DPL as a classification task, however, it is inapplicable to NLG task.

5.4.1 Natural language understanding. Table 3 shows the performance of all models, and the ablation study of MT5 (oracle), on the NLU task. First, for intent label identification, MT5 achieves the best Micro-F1 of 75.32%, followed by GPT2 of 73.32%. MT5 outperforms BERT-WWM/BERT-MED by 3.56%/3.85% and GPT2 wins by 1.56%/1.85%. So, MT5 and GPT2 can generate more accurate intent labels compared with BERT models. Second, for intent-slot label identification, BERT models outperform others by large margins in terms of both Micro-F1 and Macro-F1. BERT-MED achieves 2.01%/8.41% higher Micro-F1 and 5.65%/12.45% higher Macro-F1 than MT5 and GPT2. We believe one of the reasons is that BERT predicts over the label space rather than the whole vocabulary (like GPT2 and MT5), which makes the task easier. But BERT models are not able to predict the slot-values for the same reason. Another reason is that unlike intent identification, the training samples of intent-slot identification are inefficient and imbalanced (See Figure 3), so the generation models (e.g., MT5 and GPT2) can hardly beat the classification models (e.g., BERT-WWM and BERT-MED). Third, for value generation, MT5 significantly outperforms GT2 by 10.04% in terms of BLEU and BERT models are unable to generate values. It shows that conditional casual language model is more conducive for value generation. Fourth, MT5 outperforms others in terms of overall performance, i.e., Combination. We conducted an ablation study, and find that pseudo labeling, natural perturbation, and historical utterances all have positive effect on the overall performance. Specifically, historical utterances have the largest influence (−1.02%), followed by natural perturbation (−0.62%) and pseudo labeling (−0.40%). All scores decrease except the BLEU score of MT5 without natural perturbation. This is because that the meaning of entities might be ambiguous after modification, e.g., "azithromycin" is replaced by its common name as "泰力特 (tylett)", which is hard to be distinguished from "力比泰 (alimta)" in Chinese. CL improves the performance of MT5 in terms of most metrics. Especially, for NLU, it increases 2.76% of Macro-F1, although it slightly decreases Micro-F1. CL performs better on types of slots that account for a larger proportion of the data (e.g. "Medicine" and "Symptom" in Figure 3).

Table 3: Performance on the NLU task.

	Micro-F1/Macro-F1(%)		BLEU(%)	Combi.
	Intent	Intent-Slot	Value	
BERT-WWM	71.76/71.79	57.38/58.21	-	-
BERT-MED	71.47/71.79	**57.64/58.72**	-	-
GPT2	73.32/69.23	49.23/46.27	20.23	34.73
MT5	75.32/**72.67**	55.63/53.07	30.27	42.95
-Pseudo labeling	74.33/71.12	54.84/52.01	30.25	42.55
-Natural perturbation	73.90/70.77	53.97/50.99	30.68	42.33
-Historical utterances	74.43/71.62	54.10/51.19	29.75	41.93
MT5 + CL	**75.76**/72.65	55.31/55.83	**30.72**	**43.02**

5.4.2 Dialogue policy learning. Table 4 shows the performance of all models, and the ablation study of MT5 (oracle), on the DPL task. First, MT5 (oracle) outperforms all the other models on all metrics. Specifically, it outperforms BERT-WWM by 0.59% and 1.35% on Micro-F1 for action and action-slot label identification, respectively. This reveals that MT5 can beat BERT models when more given more information in the input, especially the result from NLU. Besides, it achieves 2.86% higher BLEU and 7.11% higher Combination compared with GPT2 (oracle), which indicates that conditional casual

[4] https://code.ihub.org.cn/projects/1775
[5] https://github.com/yangjianxin1/GPT2-chitchat
[6] https://github.com/google-research/multilingual-t5

Table 4: Performance on the DPL task. The label "(oracle)" indicates that the ground truth from NLU is used instead of the prediction.

	Micro/Macro-F1(%)		BLEU(%)	Combi.
	Action	Action-Slot	Value	
BERT-WWM	52.48/51.98	37.23/35.12	-	-
BERT-MED	49.83/49.60	35.76/34.19	-	-
GPT2	43.79/38.80	22.37/19.55	7.58	14.98
GPT2 (oracle)	45.79/41.63	27.22/24.35	9.58	18.40
MT5	46.78/41.37	26.49/22.58	9.41	17.95
MT5 (oracle)	53.07/52.07	38.58/36.51	12.44	25.51
-Pseudo labeling	52.04/50.53	38.00/36.24	11.48	24.74
-Natural perturbation	52.40/49.82	37.64/35.63	12.61	25.13
-Historical utterances	50.73/47.97	35.98/33.63	11.88	23.93
-External knowledge	51.06/48.20	31.86/28.74	10.90	21.38
MT5 (oracle) + CL	**57.78/55.66**	**40.49/38.49**	**12.63**	**26.56**

language modeling is more effective in this case. Second, we explore the joint learning performance for MT5 and GPT2, where the prediction from NLU is used as an input of DPL. MT5 still outperforms GPT2 by 2.97% for the Combination performance, specifically 2.99% for the action label identification, 4.12% for the action-slot label identification, and 1.83% for the value generation. Third, we conducted an ablation study on MT5 and find that pseudo labeling, natural perturbation, historical utterances, and external knowledge are still helpful. Specifically, external knowledge has the largest influence (−4.13%), followed by historical utterances (−1.58%), pseudo labeling (−0.97%) and natural perturbation (−0.38%). All scores decrease generally. One exception is that BLEU increases by 0.17% without natural perturbation. Similar to the case in NLU, some modified entities may cause ambiguity. CL is beneficial in terms of all evaluation metrics. Specifically, CL increases 1.05% in Combination, while improving the generation of actions by 4.71%/3.59% and action-slots by 1.91%/1.98% in terms of Micro/Macro-F1. Thus CL helps the DPL task more than it helps the NLU task.

5.4.3 Natural language generation. Table 5 shows the automatic evaluation of GPT2 and MT5, and the ablation study of MT5 (oracle), on NLG. First, MT5 (oracle) outperforms GPT2 (oracle) on METEOR. Specifically, MT5 (oracle) is 0.56% and 1.03% superior on BLEU1 and BLEU4 but 0.93% inferior on ROUGE1. It shows that although GPT2 can generate relevant tokens, the generation of MT5 is more precise. Second, we explore the joint learning performance, where the prediction of NLU and DPL is used as the input of NLG. We find that MT5 is inferior to GPT2 as METEOR, BLEU1, BLEU4 and ROUGE1 drop by 3.43%, 2.65%, 0.52% and 2.69%. It is because that the predictive quality of upstream tasks have more influence on MT5 than GPT2. Third, we conduct an ablation study for MT5 (oracle). We find that pseudo labeling, natural perturbation, historical utterances, and external knowledge are still helpful. Natural perturbation (−3.16%) is the most influential, followed by historical utterances (−1.50%), pseudo labeling (−0.95%), and external knowledge (−0.05%). CL improves MT5 (oracle) the most on BLUE1 (+0.99%), followed by ROUGE1 (0.48%), METEOR (+0.29%) and BLEU4 (0.09%). Unlike the NLU and DPL tasks, the gains of CL for the NLG task are limited.

Table 5: Automatic evaluation on the NLG task. The label "(oracle)" indicates that the ground truth from NLU and DPL is used instead of the prediction.

	Word-in-Utterance (%)			
	BLEU1	BLEU4	ROUGE1	METEOR
GPT2	17.29	2.11	68.15	19.55
GPT2 (oracle)	29.15	5.89	**74.43**	32.62
MT5	14.61	1.59	65.46	16.12
MT5 (oracle)	29.71	6.92	73.49	33.11
-Pseudo labeling	28.76	6.59	72.89	32.41
-Natural perturbation	26.55	6.49	71.73	30.18
-Historical utterances	28.21	6.62	72.69	31.98
-External knowledge	29.66	6.97	73.41	33.06
MT5 (oracle) + CL	**30.70**	**7.01**	73.97	**33.40**

Table 6: Human evaluation of the NLG task. κ is the average pairwise Cohen's kappa coefficient between annotators.

	Fluency	Specialty
GPT2 (oracle)	1.72	1.04
MT5 (oracle)	1.82	1.22
Ground truth	1.91	2.00
κ	0.64	0.62

Table 6 shows the human evaluation on the NLG task. We did not consider the joint-learned GPT2 and MT5, as they contain the accumulated error from the upstream tasks, which will influence the evaluation of NLG. First, MT5 (oracle) performs better than GPT2 (oracle) on Fluency and Specialty. This indicates that MT5 can generate more fluent responses that provide more accurate medical knowledge compared with GPT2. This is consistent with the results of automatic evaluation. Second, the Fluency score is higher than Specialty for both GPT2 and MT5. This is because Specialty is more difficult, as generating responses with massive and accurate expertise is more challenging. Third, the average pairwise Cohen's kappa coefficient is larger than 0.6 for all metrics, which indicates a good annotation agreement.

5.5 Analysis

In this section, we analyze one of the strongest performing ReMeDi benchmarks, MT5, and reflect on the ReMeDi dataset creation process in terms of dataset size and data acquisition types.

5.5.1 Dataset size. The ReMeDi benchmarks achieve a solid performance for future research to compare against. Would an increase in the ReMeDi dataset size have helped to make the benchmarks even more challenging? To answer this question, we simulate the situation of feeding MT5 with more data by pseudo labeling. We investigate the performance on the NLU, DPL, NLG tasks with increasing volumes of training dialogues, as shown in Figure 4. We see that feeding more simulated data has a positive effect on all

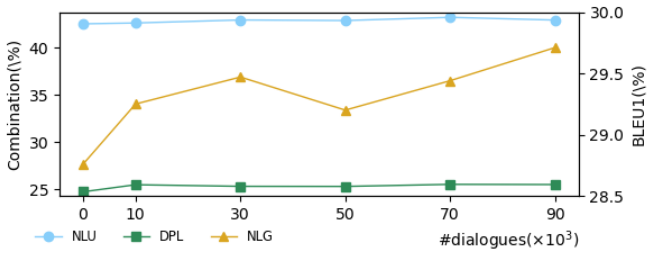

Figure 4: Analysis of data hungry tolerance on the NLU, DPL, and NLG tasks w.r.t. different size of training dialogues. The x-axis is the number of training dialogues, and the 0 point denotes no pseudo labeled dialogues. The left y-axis is the Combination score on the NLU and DPL tasks, and the right y-axis is the BLEU1 on the NLG task.

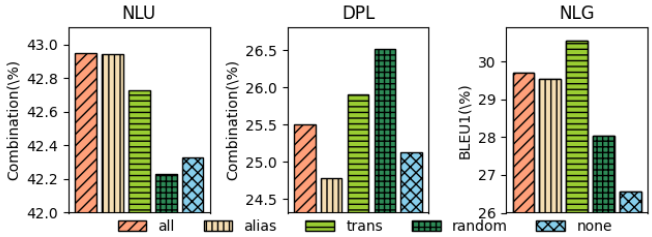

Figure 5: Analysis of data acquisition types on the NLU, DPL, and NLG tasks. We study five settings: (1) all: use all three strategies of natural perturbation; (2) alias: alias substitution; (3) trans: back-translation; (4) random: random modification; (5) none: without any natural perturbations.

three tasks, as the overall trends of the lines are upward. Specifically, NLG is increased by 0.95% of BLUE1, followed by DPL (+0.79% of Combination) and NLU(+0.69% of Combination). However, the improvement has an upper bound. For example, adding dialogues to 90K, NLU and DPL do not improve and even slightly decrease compared with the performance on 70K. It shows to what extent the current volume of dialogues suffices to approach the upper bound performance. This helps with the pains-gains trade-off of data acquisition.

5.5.2 Data acquisition types. What types of data should we expand to enlarge the gains of data acquisition? As shown in Figure 5, we compare the influence of different natural perturbation strategies on all three tasks. The overall influence of diverse data with lots of perturbation is positive. The mixture of "all" strategies significantly outperforms the "none" of strategies on all tasks. NLG is developed most by 3.16%, followed by NLU (+0.62%) and DPL (+0.38%). Besides, different strategies have different influences on different tasks. The alias strategy improves NLU most. This might be because adding alias entities to data samples helps with entity recognition. The random strategy has the largest effect on DPL, as it can improve the robustness with more input information. The trans strategy archives the best on NLG, as it can generate large-scale dialogues compared with the other two strategies. Therefore, adding data that can improve the diversity of ReMeDiis more than welcomed.

6 CONCLUSION AND FUTURE WORK

In this paper, we have introduced key *resources for Chinese medical dialogues* (ReMeDi): a dataset and benchmarks. The ReMeDi dataset is a multiple-domain, multiple-service dataset with fine-grained labels for medical dialogue systems. We focus on providing the community with a new test set for evaluation and provide a small fine-tuning set to encourage low-resource generalization without large, monolithic, labeled training sets. We consider NLU, DPL and NLG in a unified SeqMDS framework, based on which, we deploy several state-of-the-art pretrained language models, with contrastive learning, as benchmarks, the ReMeDi benchmarks. We have also evaluated the ReMeDi benchmarks against the ReMeDi dataset. Both the ReMeDi datasets and benchmarks are available online; please see the Appendix for details.

The resources released with this work have broader implications in that: (1) The fine-gained labels provided with ReMeDi can help research on the interpretability of medical dialogue systems. (2) The performance of the baseline models are far from satisfactory; therefore, we hope that the ReMeDi resources facilitate and encourage research in low-resource medical dialogue systems.

One limitation of the ReMeDi dataset is that we do not provide explicit boundaries between dialogue sessions with different service types. This makes it challenging to explicitly model relationships among multiple services.

As to future work, on the one hand, we will extend ReMeDi with service boundary labels to facilitate research on dialogue context modeling among multiple services. On the other hand, we will extend ReMeDi with more languages to help study multilingual MDSs. Last but not least, we call for studies to improve the benchmark performance, as well as conduct underexplored research, e.g., dialogue tasks for rare diseases under extremely low-resource settings.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of China (61902219, 61972234, 62072279, 62102234), the Natural Science Foundation of Shandong Province (ZR2021QF129), the Key Scientific and Technological Innovation Program of Shandong Province (2019JZZY010129), Shandong University multidisciplinary research and innovation team of young scholars (No. 2020QNQT017), the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, the Tencent WeChat Rhino-Bird Focused Research Program (JR-WXG-2021411), Meituan. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

A APPENDIX

A.1 Resources

All resources presented in this paper, the dataset, code for the baselines, and evaluation scripts are shared at https://github.com/yanguojun123/Medical-Dialogue. The shared resources are organized in multiple folders: (1) The folder "annotation_guide" contains the code and guidelines for the labeling system. (2) The folder "data" contains the ReMeDi-large and ReMeDi-base datasets. Each

dialogue consists of multiple turns of utterances from doctors or patients, identified with a unique dialogue id. Each utterance in a dialogue turn is provided with: turn id, dialogue role, utterance text, and label list. Each label consists of sub-sentence text, the start/end position of sub-sentences, and the intent-slot-value or action-slot-value labels. (3) The folder "data_process" contains the code for processing the crawled raw data, the pseudo labeling and natural perturbation. (4) The folder "model" contains the code of the benchmark models based on BERT, GPT2, and MT5. (5) The folder "evaluate" contains the code for automatic evaluation in terms of all metrics. All resources are licensed under the MIT license.

A.2 Implementation details

BERT-WWM and BERT-MED use 12 transformer blocks with 12 attention heads and the hidden size is 768. The maximum length of input tokens is 512 and the learning rate is 2e-5.

GPT2 uses 10 transformer decoder blocks with 12 attention heads and the hidden size is 768. MT5 uses 8 transformer encoder blocks followed by 8 decoder blocks with 12 attention heads and the hidden size is 512. For GPT2 and MT5, the maximum length of input tokens is 800 and the learning rate is 1.5e-4. We set the coefficient λ as 0.8 and the temperature τ as 0.5 for the contrastive learning.

We fine-tune the models on three training sets produced by pseudo labeling, natural perturbation, and human annotation, respectively. We use AdamW [24] as the optimization algorithm. The maximum training epochs are set to 30. We implemented benchmark models by PyTorch [46]. Our model is trained with 4 Nvidia TITAN RTX GPUs with 20 GB of memory. The results reported in this work can be reproduced with the random seed fixed.

REFERENCES

[1] Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. 2021. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 239–251.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 65–72.

[3] Pawel Budzianowski and Ivan Vulic. 2019. Hello, It's GPT-2 - how can I help you? Towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*. 15–22.

[4] Ernie Chang, Vera Demberg, and Alex Marin. 2021. Jointly improving language understanding and generation with quality-weighted weak supervision of automatic labeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. 818–829.

[5] Guan-Lin Chao and Ian Lane. 2019. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. *arXiv:1907.03040* (2019).

[6] Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the 9th Workshop on Statistical Machine Translation*. 362–367.

[7] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter* (2017), 25–35.

[8] Sixuan Chen and Shuai Yu. 2019. WAIS: word attention for joint intent detection and slot filling. In *Proceedings of the 33rd Association for the Advancement of Artificial Intelligence*. 9927–9928.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*. 1597–1607.

[10] Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the 2nd Workshop on Natural Language Processing for Medical Conversations*. 66–76.

[11] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 539–546.

[12] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for Chinese bert. *arXiv:1906.08101* (2019).

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4171–4186.

[14] Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. Extracting symptoms and their status from clinical conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 915–925.

[15] Nan Du, Mingqiu Wang, Linh Tran, Gang Li, and Izhak Shafran. 2019. Learning to infer entities, properties and their relations from clinical conversations. In *Proceedings of the 16th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 4978–4989.

[16] Hongchao Fang and Pengtao Xie. 2020. CERT: Contrastive self-supervised learning for language understanding. *arXiv:2005.12766* (2020).

[17] Nicholas Frosst, Nicolas Papernot, and Geoffrey Hinton. 2019. Analyzing and improving representations with the soft nearest neighbor loss. In *International conference on machine learning*. 2012–2020.

[18] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *arXiv:2104.08821* (2021).

[19] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the 13th international conference on artificial intelligence and statistics*. 297–304.

[20] Michael Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.* (2012), 307–361.

[21] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Proceedings of the 34th Conference on Neural Information Processing Systems*. 3104–3112.

[22] Degen Huang, Jiahuan Pei, Cong Zhang, Kaiyu Huang, and Jianjun Ma. 2018. Incorporating prior knowledge into word embedding for Chinese word similarity measurement. *ACM Transactions on Asian and Low-Resource Language Information Processing* (2018), 1–21.

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*.

[24] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*.

[25] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems*. 3294–3302.

[26] Tuan Manh Lai, Quan Hung Tran, Trung Bui, and Daisuke Kihara. 2020. A simple but effective bert model for dialog state tracking on resource-limited systems. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. 8034–8038.

[27] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access* (2020), 193907–193934.

[28] Seanie Lee, Dong Bok Lee, and Sung Ju Hwang. 2021. Contrastive learning with adversarial perturbations for conditional text generation. In *9th International Conference on Learning Representations*.

[29] Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Proceedings of the Soviet Physics Doklady*. 707–710.

[30] Bin Li, Encheng Chen, Hongru Liu, Yixuan Weng, Bin Sun, Shutao Li, Yongping Bai, and Meiling Hu. 2021. More but correct: Generating diversified and entity-revised medical response. *arXiv:2108.01266* (2021).

[31] 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. *Proceedings of the 44th International Conference on Research and Development in Information Retrieval* (2021), 544–554.

[32] Shuang Li, Xavier Puig, Yilun Du, Ekin Akyürek, Antonio Torralba, Jacob Andreas, and Igor Mordatch. 2021. Language model pre-training improves generalization in policy learning. (2021).

[33] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. *arXiv:1703.01008* (2017).

[34] 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:2004.14254* (2020).

[35] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out*. 74–81.

[36] Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2021. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Proceedings of the 35th Association for the Advancement of Artificial Intelligence*. 13362–13370.

[37] 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 16th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 5033–5042.

[38] Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. *arXiv:2106.02787* (2021).

[39] Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020. MedDG: A large-scale medical consultation dataset for building medical dialogue system. *arXiv:2010.07497* (2020).

[40] Olga Majewska, Evgeniia Razumovskaia, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2022. Cross-lingual dialogue dataset creation via outline-based generation. *arXiv:2201.13405* (2022).

[41] Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv:2009.13570* (2020).

[42] Luke Melas-Kyriazi, George Han, and Celine Liang. 2019. Generation-distillation for efficient natural language understanding in low-data settings. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. 124–131.

[43] Nikola Mrkšic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašic, Lina Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 142–148.

[44] Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 1777–1788.

[45] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4004–4012.

[46] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*. 8024–8035.

[47] Jiahuan Pei, Pengjie Ren, and Maarten de Rijke. 2019. A modular task-oriented dialogue system using a neural mixture-of-experts. In *Proceedings of the 42nd SIGIR Workshop on Conversational Interaction Systems*.

[48] Jiahuan Pei, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2020. Retrospective and prospective mixture-of-generators for task-oriented dialogue response generation. In *Proceedings of the 24th European Association of Computational Linguistics*. 2148–2155.

[49] Jiahuan Pei, Cong Zhang, Degen Huang, and Jianjun Ma. 2016. Combining word embedding and semantic lexicon for Chinese word similarity computation. In *Natural Language Understanding and Intelligent Applications*. 766–777.

[50] Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Yun-Nung Chen, and Kam-Fai Wong. 2018. Adversarial advantage actor-critic model for task-completion dialogue policy learning. In *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech and Signal*. IEEE, 6149–6153.

[51] Libo Qin, Wanxiang Che, Yangming Li, Haoyang Wen, and Ting Liu. 2019. A stack-propagation framework with token-level intent detection for spoken language understanding. In *Proceedings of the 16th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 2078–2087.

[52] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI blog* (2018).

[53] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog* (2019).

[54] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683* (2019).

[55] 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.* (2020), 140:1–140:67.

[56] Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Anna Korhonen, and Ivan Vulic. 2021. Crossing the conversational chasm: A primer on multilingual task-oriented dialogue systems. *arXiv:2104.08570* (2021).

[57] Pengjie Ren, Zhongkun Liu, Xiaomeng Song, Hongtao Tian, Zhumin Chen, Zhaochun Ren, and Maarten de Rijke. 2021. Wizard of search engine: Access to information through conversations with search engines. In *Proceedings of the 44th International Conference on Research and Development in Information Retrieval*. 533–543.

[58] Ruslan Salakhutdinov and Geoff Hinton. 2007. Learning a nonlinear embedding by preserving class neighbourhood structure. In *Artificial Intelligence and Statistics*. 412–419.

[59] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 815–823.

[60] Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv:2009.13818* (2020).

[61] Xiaoming Shi, Haifeng Hu, Wanxiang Che, Zhongqian Sun, Ting Liu, and Junzhou Huang. 2020. Understanding medical conversations with scattered keyword attention and weak supervision from responses. In *Proceedings of the 34th Association for the Advancement of Artificial Intelligence*. 8838–8845.

[62] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 1857–1865.

[63] Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, and Jie Fu. 2021. Pre-trained language models in biomedical domain: A systematic survey. *arXiv:2110.05006* (2021).

[64] Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. CLINE: Contrastive learning with semantic negative examples for natural language understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing,*. 2332–2342.

[65] Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. 2022. SNCSE: Contrastive learning for unsupervised sentence embedding with soft negative samples. *arXiv:2201.05979* (2022).

[66] Shanshan Wang, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jian-Yun Nie, Jun Ma, and Maarten de Rijke. 2020. Coding electronic health records with adversarial reinforcement path generation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 801–810.

[67] Jason W. Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv:1901.11196* (2019).

[68] 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 201–207.

[69] Henry Weld, Xiaoqi Huang, Siqi Long, Josiah Poon, and Soyeon Caren Han. 2021. A survey of joint intent detection and slot-filling models in natural language understanding. *arXiv:2101.08091* (2021).

[70] Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve J. Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 16th European Chapter of the Association for Computational Linguistics*. 438–449.

[71] Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 17th Conference on Empirical Methods in Natural Language Processing*. 917–929.

[72] Yuan Xia, Jingbo Zhou, Zhenhui Shi, Chao Lu, and Haifeng Huang. 2020. Generative adversarial regularized mutual information policy gradient framework for automatic diagnosis. In *Proceedings of the 34th Association for the Advancement of Artificial Intelligence*. 1062–1069.

[73] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations*.

[74] 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 *Proceedings of the 33rd Association for the Advancement of Artificial Intelligence*. 7346–7353.

[75] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. MT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 19th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 483–498.

[76] Wenmian Yang, Guangtao Zeng, Bowen Tan, Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, Qingyang Wu, Zhou Yu, et al. 2020. On the generation of medical dialogues for COVID-19. *arXiv:2005.05442* (2020).

[77] Yunyi Yang, Yunhao Li, and Xiaojun Quan. 2020. Ubar: Towards fully end-to-end task-oriented dialog systems with gpt-2. *arXiv:2012.03539* (2020).

[78] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog:

A large-scale medical dialogue dataset. In *Proceedings of the 17th Conference on Empirical Methods in Natural Language Processing*. 9241–9250.

[79] Taolin Zhang, Zerui Cai, Chengyu Wang, Minghui Qiu, Bite Yang, and Xiaofeng He. 2021. SMedBERT: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 5882–5893.

[80] Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. MIE: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6460–6469.

[81] Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences* (2020), 2011–2027.

[82] Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Journal of the Transactions of the Association for Computational Linguistics* (2020), 281–295.

[83] Lei Zuo, Kun Qian, Bowen Yang, and Zhou Yu. 2021. AllWOZ: Towards multilingual task-oriented dialog systems for all. *arXiv:2112.08333* (2021).