DIALMED: A Dataset for Dialogue-based Medication Recommendation

Zhenfeng He[1]*, Yuqiang Han[1]*, Zhenqiu Ouyang[3], Wei Gao[4], Hongxu Chen[5], Guandong Xu[5], Jian Wu[2]†

[1]College of Computer Science and Technology, Zhejiang University
[2]School of Public Health, Zhejiang University [3]Polytechnic Institute, Zhejiang University
[4]Ningbo Institute of Technology, Zhejiang University
[5]University of Technology Sydney
{hezf, hyq2015, oyzq, gw, wujian2000}@zju.edu.cn
{hongxu.chen, guandong.xu}@uts.edu.au

Abstract

Medication recommendation is a crucial task for intelligent healthcare systems. Previous studies mainly recommend medications with electronic health records (EHRs). However, some details of interactions between doctors and patients may be ignored or omitted in EHRs, which are essential for automatic medication recommendation. Therefore, we make the first attempt to recommend medications with the conversations between doctors and patients. In this work, we construct DIALMED, the first high-quality dataset for medical dialogue-based medication recommendation task. It contains 11,996 medical dialogues related to 16 common diseases from 3 departments and 70 corresponding common medications. Furthermore, we propose a **D**ialogue structure and **D**isease knowledge aware **N**etwork (**DDN**), where a QA Dialogue Graph mechanism is designed to model the dialogue structure and the knowledge graph is used to introduce external disease knowledge. The extensive experimental results demonstrate that the proposed method is a promising solution to recommend medications with medical dialogues. The dataset and code are available at https://github.com/f-window/DialMed.

1 Introduction

The outbreak of COVID-19 has challenged the healthcare systems and led to millions of patients facing delays in diagnosis and treatment. As an essential complement to the traditional face-to-face medicine, telemedicine relieved the therapeutic stress caused by the diversion of medical resources. According to the report of WeDoctor[1], an online health consultation platform in China, about 1.2 million patients conducted online medical consultations during the COVID-19 Pandemic.

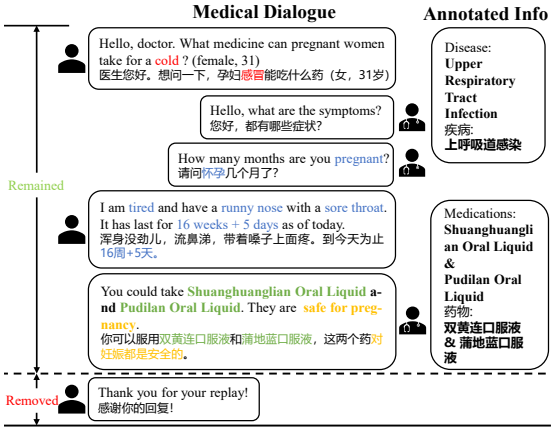


Figure 1: A typical medication consultation dialogue. Here, the disease is ***Upper Respiratory Tract Infection***, and the medication is ***Shuanghuanglian Oral Liquid*** and ***Pudilan Oral Liquid***.

Telemedicine can increase the availability of medical treatment, reduce healthcare costs, and improve the quality of care. Consequently, it has attracted increasing attention due to its vast application potential.

Our study found that around 31% of online consultations are about what medications the patients should take based on their current conditions[2]. Figure 1 demonstrates a typical medication consultation dialogue. The patient reported the health issues initially, with some personal information, such as gender and age. Then the doctor asked for further information (*e.g.*, symptoms and disease history) about the patient. Finally, the doctor provided medication advice based on the gathered information and clinical experience.

Existing studies on medication recommendation are primarily based on EHRs (Zhang et al., 2017; Shang et al., 2019b; An et al., 2021), accumulatively collected according to a diagnostic procedure in clinics. However, the doctors will omit some details of interactions with patients in EHRs,

*Both authors contributed equally to this research.
†Corresponding author.
[1]https://www.guahao.com/
[2]Refer to Appendix D.1 for details of statistic.

Proceedings of the 29th International Conference on Computational Linguistics, pages 721–733
October 12–17, 2022.

which are essential for the automatic medication recommendation. Compared to EHRs, medical dialogues retain original interactions between doctors and patients, containing more rich information. To this end, medical dialogue-based medication recommendation is a promising and challenging task.

Therefore, in this work, we study the new task, namely dialogue-based medication recommendation. Due to the lack of available datasets, we firstly construct a high-quality online medical dialogues dataset (DIALMED) for this task. It contains 11,996 consultation dialogues, 16 diseases from 3 different departments, and 70 related common medications.

Then, to further advance the research of this task, we propose a **D**ialogue structure and **D**isease knowledge aware **N**etwork (**DDN**). In DDN, for the input dialogue, we first utilize a pre-trained language model to extract the semantic information of each utterance. A mechanism named QA Dialogue Graph is designed to understand the questions&answers implied in utterances, and then we apply graph attention network on this QA graph to get the dialogue embedding. Meanwhile, for the input disease, we use its identity to query the entity in a knowledge graph CMeKG[3], and input the dialogue embedding to a graph attention network to get contextual disease embedding. The two embeddings are fused to make the medication prediction. Moreover, we conduct extensive experiments to show that the proposed method can effectively recommend medications with medical dialogues.

Our contributions can be summarized as follows:

- We construct the first high-quality human-annotated dialogue dataset for dialogue-based medication recommendation task.
- We propose a novel medication recommendation framework which models dialogue structure with QA Dialogue Graph and introduces external disease knowledge.
- We conduct extensive experiments to demonstrate DDN can extract the essential information to make medication recommendation effectively.

2 Related Work

Medication Recommendation. Existing medication recommendations are mainly based on EHRs. It could be categorized into instance-based and longitudinal-based recommendation methods (Shang et al., 2019b). Instance-based methods are based on the current health conditions extracted from recent visit (Zhang et al., 2017; Wang et al., 2019a). For example, (Zhang et al., 2017) proposed a multi-instance multi-label learning framework to predict medication combination based on patient's current diagnoses. Longitudinal-based methods leverage the temporal dependencies among clinical events (Choi et al., 2016; Le et al., 2018; Shang et al., 2019b,a; Wang, 2020; He et al., 2020; Wang et al., 2021; Yang et al., 2021). Among them, (Shang et al., 2019a) combined the power of graph neural networks and BERT for medication recommendation. (Yang et al., 2021) proposed a drug-drug interactions (DDI)-controllable drug recommendation model to leverage drugs' molecule structures and model DDIs explicitly.

Unlike the work mentioned above, dialogue-based medication recommendation task is more challenging in practice due to the noisy and sparse data. Because of the privacy issue, it is difficult to get historical dialogues of a patient on online consultation platforms. So we perform the medication recommendation solely based on the current medical dialogues.

Graph Neural Networks. Graph neural networks have attracted a lot of attention for processing data with graph structures in various domains (Zhou et al., 2020). For example, (Kipf and Welling, 2017) proposed the graph convolutional networks (GCN) . With integration of attention mechanisms, graph attention networks(GAT) (Veličković et al., 2018) has become one of the most popular methods in graph neural networks.

Recently, some works have applied GAT to the dialogue modeling. (Chen et al., 2020) used Graph attention and recurrent GAT to fully encode dialogue utterances, schema graphs, and previous dialogue states for dialogue state tracking. (Qin et al., 2020) proposed a co-interactive GAT layer to simultaneously solve both dialog act recognition and sentiment classification task. In this work, we utilize GAT to model the intra- and inter-speaker correlations to propagate semantic on the QA Dialogue Graph and extend the GAT on knowledge graph to introduce external knowledge.

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

3 Corpus Description

In this section, we introduce the construction details and statistics of DIALMED, and its comparison with other studies.

3.1 Construction Details

Our dataset is collected from Chunyu-Doctor[4], which is a popular Chinese medical consultation website for doctors and patients. The conversations between doctors and patients contain rich but complex information, mainly related to the patients' current conditions. The diagnosed diseases and symptoms both are indispensable for accurate medication recommendation. Considering the complexity of the symptoms, we decide to utilize information from *explicit disease* and *implicit symptoms* in this paper. So we annotate the diagnosed diseases and recommended medications (replaced with a mask token to keep the original dialogue structure). For the example in Figure 1, we annotate the disease *Upper Respiratory Tract Infection*, and replace the medications *Shuanghuanglian Oral Liquid* and *Pudilan Oral Liquid* with special token *[MASK]*. Moreover, the future utterances after the point of recommendation are removed to make DIALMED more realistic, as the decision of doctors should not be influenced by future contexts.

The procedure of annotation consists of two parts, labeling and normalization of medications and diseases. First, we select 16 common diseases and the corresponding common medications from 3 departments (i.e., respiratory, gastroenterology, and dermatology) with the guidance of a doctor. These diseases have abundant medication consultations online. Then three annotators with relevant medical backgrounds are involved. Each dialogue is annotated by two annotators and will be further judged by another one if there is any inconsistency. The annotation consistency, i.e., the Cohen's kappa coefficient (Fleiss and Cohen, 1973) of the labelled dialogues is 88.4%. For the quality of dataset, conversations containing unsuitable medications for patients would be discarded.

Secondly, we normalize the medications since there are many generic names, trade names, or colloquial expressions for the same drug in dialogues. Specifically, different brands of the same drug are grouped into one cluster and normalized as a common name from DXY Drugs Database[5]. For example, **Omeprazole enteric-coated tablet** and **Omeprazole tablet** are normalized to **Omeprazole**. Similarly, we normalize the different names of diseases into ICD-10 standard names. The dialogues, hard to give diagnosed diseases or given diseases out of our scope, would be marked as a special placeholder, **None or Others**.

3.2 Dataset Statistics

Top of Table 2 summarizes the statistics of DIALMED. The scenario of dialogues in the dataset is similar to outpatient procedure, so the number of medicines per dialogue is relatively small. Then, the frequency of medications and diseases are shown in Figure 2(a) and Figure 2(b) respectively. The distributions of quantity demonstrate that DIALMED aligns with the real-world case.

Compared to the other medical dialogue datasets in Table 1, our dataset has three advantages: (1) DIALMED has the largest volume among the manual annotation datasets, as unlabeled datasets are mainly constructed for the task of dialogue generation. (2) Though the future contexts after recommendation are removed, the average number of dialogue turns in DIALMED still remains high compared to other datasets. It is mainly benefited from our evaluation for inclusion of short dialogues in DIALMED during the labeling process. (3) We carefully choose the fields suitable for medication recommendation and avoid coarsely expanding the scope of medical domains, which makes DIALMED have a higher quality.

The panoramas of medications & diseases' frequency could be found in Appendix D.2.

3.3 The comparison with other studies

To our best knowledge, DIALMED is the first dataset for the medication recommendation based on medical dialogues. It has the following differences with the existing work.

Dataset Medical dialogue has attracted increasing attention in recent years. Although there are medication mentions in many medical dialogue datasets, the distributions are fragmentary and the authors do not categorize and normalize these drug mentions which would lead to label explosion. For instance, medication mentions, **Omeprazole entericcoated tablet**, **Omeprazole tablet** and **Omeprazole**, which may occur in dialogues would be three classes without normalization. In fact, they are essentially equivalent in the eyes of doctors. By

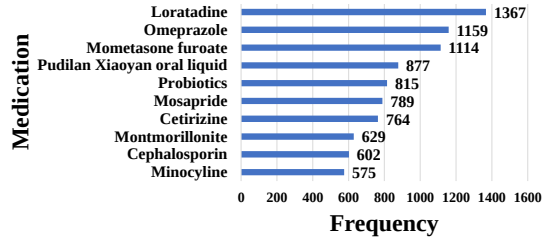
[4] https://www.chunyuyisheng.com/
[5] http://drugs.dxy.cn/

Dataset	#Task	#Domain	#Disease	#Dialogue	#Avg. Turn	#Annotation
MZ(Wei et al., 2018)	Diagnosis	Pediatrics	4	710	-	Man.
DX(Xu et al., 2019)	Diagnosis	Pediatrics	5	527	5.34	Man.
CMDD(Lin et al., 2019)	Diagnosis	Pediatrics	4	2,067	42.09	Man.
SAT(Du et al., 2019)	Extraction	14	-	2,950	-	Man.
MIE(Zhang et al., 2020)	Extraction	Cardiology	6	1,120	16.19	Man.
MSL(Shi et al., 2020)	Extraction	Pediatrics	5	2,652	-	Man.
MedDG(Liu et al., 2020)	Extraction	Gastroenterology	12	17,864	21.60	Man.& Semi-Auto.
COVID-EN(Yang et al., 2020)	Generation	COVID-19	1	603	8.7	None
COVID-CN(Yang et al., 2020)	Generation	COVID-19	1	1088	2.0	None
MedDialog-EN(Zeng et al., 2020)	Generation	51	96	257,332	2	None
MedDialog-CN(Zeng et al., 2020)	Generation	29	172	3,407,494	3.3	None
Chunyu(Lin et al., 2021)	Generation	-	15	12,842	24.7	Rule
KaMed(Li et al., 2021)	Generation	100	-	63,754	11.62	None
ReMeDi(Yan et al., 2022)	Diag.&Ext.&Gene.	30	491	1,557	16.34	Man.
DIALMED(ours)	Medication	R&G&D	16	11,996	10.94	Man.

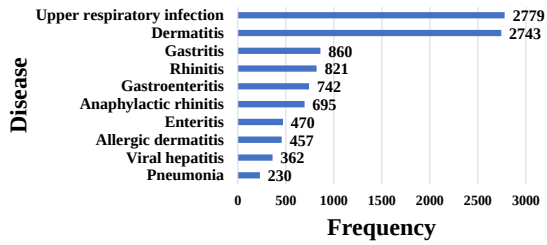
Table 1: Comparison between our dataset and other related medical dialogue datasets. Extraction, Generation and Medication mean information extraction, dialogue generation and medication recommendation separately. R&G&D, Man. and Semi-Auto are abbreviations of Respiratory&Gastroenterology&Dermatology, Manual and Semi-Automated respectively.

	#Dial.	#Dise.	#Med.	Avg.M	Avg.T	Max.T	Avg.U	Max.U
Resp.	4,859	4	45	2.06	10.76	52	18.18	374
Gastro.	3,818	9	39	1.88	13.05	58	16.70	463
Derma.	3,319	3	27	1.62	8.77	44	18.82	453
Total	11,996	16	70	1.88	10.94	58	17.76	463
Train.	9,605	16	70	1.88	10.95	58	17.74	463
Dev.	1,192	16	70	1.89	11.25	49	17.45	298
Test.	1,199	16	70	1.89	10.58	42	18.27	293

Table 2: Data statistics of DIALMED. M, T, and U represent medicine, dialogue turns, and utterance.



(a) The frequency of medications.



(b) The frequency of diseases.

Figure 2: The frequencies of medications and diseases. Top 10 are exhibited for the constraint of space.

contrast, we reduce the complexity caused by the doctors' preferences for different brands through categorization and normalization. DIALMED is developed for drug recommendation.

Task Drug recommendation is a sub-task of medical diagnosis. According to patients' questions, the objectives of current diagnosis systems are to generate the optimal clinical responses which may be intended as one of greeting, inquiry or diagnosis. Even if there contains drug mentions in responses, it is just one of the system's options. Drug recommendation is a key task and requires specialized dataset. DIALMED goes a step forward.

Scenario There are remarkable distinctions between DIALMED and MIMIC-III (Johnson et al., 2016), an EHR database which is relied on in current medication recommendation study. The scenario of the former is outpatient procedure while the data from the latter is generated from Intensive Care Units (**ICU**). In MIMIC-III, for example, the number of medications is 145, the average number of medications in each visit is 8.80, and the average number of diagnosis in each visit is 10.51. In contrast, the labels in medical dialogues are relatively sparse, leading to a more challenging task.

4 Our Approach

In this section, we first introduce the dialogue-based medication recommendation task, and then describe the proposed DDN in detail.

4.1 Problem Formulation

In the online medical dialogue setting, each dialogue consists of a sequence of utterances from the patient and the doctor. Formally, each dialogue can be represented as $\mathcal{D}_n = \{u_1, u_2, ..., u_{|\mathcal{D}_n|}\}$, where $n \in \{1, 2, ..., N\}$, N denotes the total number of dialogues in the dataset, and $|\mathcal{D}_n|$ represents the number of utterances in a dialogue $\mathcal{D}_n$. Each utterance can be represented as $u_i = \{w_i^1, ..., w_i^j, ..., w_i^{|u_i|}\}$, where w_i^j is the j-th word in u_i and $|u_i|$ denotes the number of words in u_i. We collect all the diseases and medications mentioned in the dataset to construct a disease corpus $\mathcal{S}$ and medication corpus $\mathcal{M}$. To avoid notation clutter, we hereinafter remove the subscript n as we only consider a single dialogue instance. Formally, given the consultation dialogue $\mathcal{D}$ and the diagnosed disease d as inputs, dialogue-based medication recommendation aims to recommend potential treatment medications $\mathbf{y}$ in $\mathcal{M}$, where $\mathbf{y} \in \{0, 1\}^{|\mathcal{M}|}$.

4.2 Model Overview

The proposed end-to-end framework is presented in Figure 3, consisting of two parts: (1) Dialogue Encoder, encoding the medical dialogues between patient and doctor by comprehensively capturing the semantic information and dialogue structure. (2) Disease Encoder, incorporating external medical knowledge based on the disease information from the dialogue and knowledge graph.

4.3 Dialogue Encoder

Dialogues contain two types of important information: (1) the rich semantic information, (2) strong structural correlations between utterances.

Utterance Encoding Pre-trained language models (*e.g.*, RoBERTa) are utilized to capture the semantic information in utterances. First, special tokens [CLS] (capturing utterance representation) and [SEP] (separating different utterances) are inserted at the beginning and end of each utterance token sequence u_i. Then the position embedding of each token in a utterance is calculated. In addition, two types of speaker embeddings (*i.e.*, *Doctor* and *Patient*) are proposed to make model aware of the speaker role of the utterance. The model takes the sum of three embeddings as input and outputs the representation of [CLS] as the utterance embedding $\mathbf{h}$. So a dialogue $\mathcal{D}$ can be represented as $\mathbf{h}_D = \{\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{|\mathcal{D}|}\}$.

QA Dialogue Graph In medical conversations, the interactions between doctors and patients tend to be in the form of questions and answers. For example, in Figure 3, the doctor asked two questions in u_2 and u_3, and the patient gave the answers in u_4. So it's important to capture the structure of QA pairs in conversation in order to understand the whole medical dialogue. We propose a new method to model the dialogue based on the observation that there is a high possibility of question-and-answer relations between adjacent utterances.

Specifically, we design a mechanism named QA Dialogue Graph, where each utterance is represented as a node in graph, and consecutive utterances spoken by the same speaker is represented as a block, e.g., u_2 and u_3 constitute a block with two nodes, and u_4 is another block with one node. Then the constructions of edges between nodes can be defined as follows:

- Within a block, each node connects with all other nodes in the block. This represents the intra-speaker correlation and ensures the information from the same speaker propagates among utterances within a local context.

- For two adjacent blocks, each node in a block connects with all nodes in the other block. This represents the inter-speaker correlation and ensures the information flow between doctors and patients within consecutive contexts.

An example of the adjacency matrix of the dialogue is shown in Figure 3. In general, when compared to previous works on dialogue modeling, QA Dialogue Graph has two advantages. Firstly, the construction of graphs does not require additional supervised information (Joshi et al., 2021; Feng et al., 2021). Secondly, our method comprehensively captures the structural and semantic information of QA pairs, which is key to understanding conversations (Qin et al., 2020; Shen et al., 2021b).

Dialogue Encoding GAT is employed to automatically aggregate semantic and structure features on QA Dialogue Graph. In particular, the l-th layer representation of a vertex can be computed as:

$$\mathbf{h}_i^{(l)} = \sigma(\sum_{j \in \mathcal{N}_i} \alpha_{ij} W_h \mathbf{h}_j^{(l-1)}) \quad (1)$$

where $\mathcal{N}_i$ is the first-order neighbors of vertex i, $W_h \in \mathbb{R}^{d_l \times d_{l-1}}$ is a trainable weight matrix, and σ is a nonlinear activation function. The weight

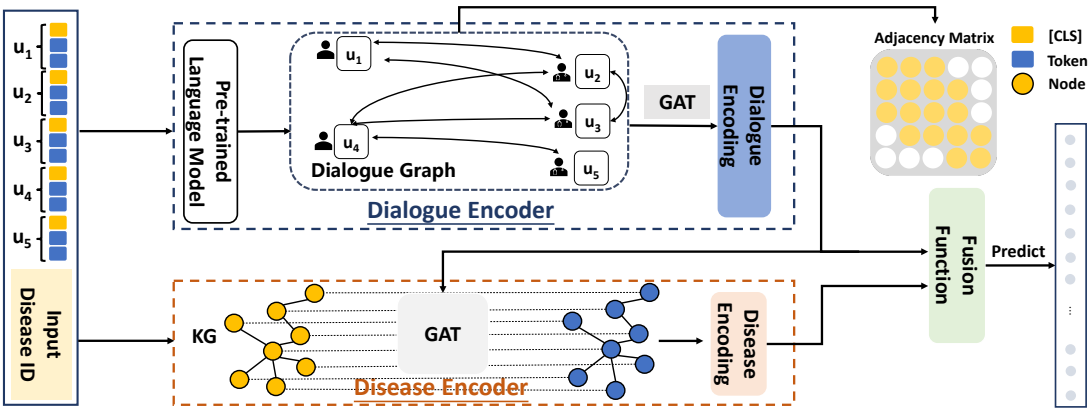


Figure 3: The framework of the proposed DDN for dialogue-based medication recommendation.

α_{ij} which determines the relatedness between two vertices can be calculated following (Veličković et al., 2018):

$$\alpha_{ij} = \frac{\exp\left(\sigma(\mathbf{a}^T W_h[\mathbf{h}_i \,||\, \mathbf{h}_j])\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\sigma(\mathbf{a}^T W_h[\mathbf{h}_i \,||\, \mathbf{h}_k])\right)} \tag{2}$$

where $\mathbf{a} \in \mathbb{R}^{2d_l}$ is a trainable weight matrix, and σ is the LeakyReLU activation function. Finally, we apply the attention pooling on nodes embedding to obtain the dialogue representation $\mathbf{h}_\mathcal{D}$, where $\mathbf{W}_a$ is a learnable parameter and $\mathbf{h}^{(l)}$ is the representation of utterances after l^{th} layer.

$$\widehat{\alpha} = softmax(\mathbf{W}_a\mathbf{h}^{(l)}) \tag{3}$$

$$\mathbf{h}_\mathcal{D} = \sum_i \widehat{\alpha}_i \mathbf{h}_i^{(l)} \tag{4}$$

4.4 Disease Encoder

Disease knowledge is crucial for delivering accurate medication recommendation. In this paper, we incorporate knowledge from CMeKG, a high-quality Chinese medical knowledge graph. TransR (Wang et al., 2019b) is utilized to get the initial entities embedding. Given a disease d, we first identify the corresponding entity in CMeKG, and then a KG subset with K hops starting from the disease entity is sampled randomly, finally the GAT network is used to get the disease embedding under the dialogue context.

Here, we fuse the entity, relation and dialogue information to get the attention scores:

$$\beta_{ij} = \frac{\exp\left(\sigma(\mathbf{a}^T[W[\mathbf{h}_i, \mathbf{h}_j] \,||\, W_r\mathbf{r}_\varphi \,||\, W_D\mathbf{h}_\mathcal{D}])\right)}{\sum_{j\in\mathcal{N}_i}\exp\left(\sigma(\mathbf{a}^T[W[\mathbf{h}_i, \mathbf{h}_j] \,||\, W_r\mathbf{r}_\varphi \,||\, W_D\mathbf{h}_\mathcal{D}])\right)} \tag{5}$$

where σ is the LeakyReLU function, $\mathbf{h}_i$, $\mathbf{h}_j$ and $\mathbf{r}_\varphi$ are the embeddings of node i, j and their relation separately. And W, W_r, and W_D are learnable weights to transform node, relation and dialogue embeddings, respectively. Then the l-th layer of disease embedding can be obtained as follows:

$$\mathbf{s}_i^{(l)} = \sigma\Big(\sum_{j\in\mathcal{N}_i} \beta_{ij} W_k \mathbf{h}_j^{(l-1)}\Big) \tag{6}$$

The contextual embedding of last layer is the disease d's representation, denoted by $\mathbf{s}_d$.

For dialogues with None or Others placeholder rather than a disease label, a learnable vector $\hat{\mathbf{s}}_d$ would be assigned to $\mathbf{s}_d$.

4.5 Model Inference and Optimization

The dialogue $\mathbf{h}_\mathcal{D}$ and disease $\mathbf{s}_d$ are fused by the *fusion function* to make prediction. In this work, we concatenate them and then fed it into decoder to make the medication prediction as follows:

$$\mathbf{y} = \sigma(W_o[\mathbf{h}_\mathcal{D}; \mathbf{s}_d] + \mathbf{b}_o) \tag{7}$$

where $W_o \in \mathbb{R}^{|\mathcal{M}|\times 2d}$ and $\mathbf{b}_o \in \mathbb{R}^{|\mathcal{M}|}$ are trainable weight matrices for the decoder, σ is the sigmoid activation function. Here, we reserve all the candidates whose probability is higher than the threshold of 0.5 as the recommended treatment medication combination.

Since medication combination recommendation is treated as a multi-label classification task (Shang et al., 2019b; Yang et al., 2021), we utilize the binary cross-entropy loss as the objective function, which can be formulated as:

$$\mathcal{L} = -\sum_{i=1}^{|\mathcal{D}|}\sum_{j=1}^{|\mathcal{M}|}\left(y_j^{(i)} \log \widehat{y}_j^{(i)} + (1-y_j^{(i)})\log(1-\widehat{y}_j^{(i)})\right) \tag{8}$$

where $|\mathcal{D}|$ is the number of dialogues in the training set, $|\mathcal{M}|$ is the number of medications. $y_j^{(i)}$ is the ground truth label which equals 1 if medication

j is prescribed by the doctor in dialogue i, and 0 otherwise. $\widehat{y}_j^{(i)}$ is the predicted probability of recommending medication j.

5 Experiments

5.1 Experimental Setup

Dataset In our experiments, we divide the data into train/development/test dialogue sets as shown in Table 2. The average number of medications in each dialogue is approximately the same, as well as the the average length of utterances and dialogues, meaning the distribution of the data is relatively consistent among three sets.

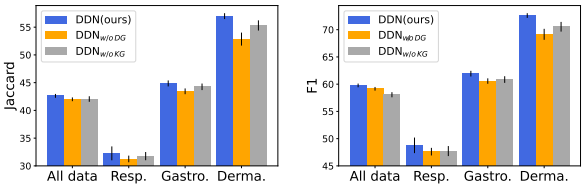
Implementation Details The pretained model we use is Chinese RoBERTa-base model. The learning rate and the batch size are set as 2×10^{-5} and 8, respectively. Adam optimizer is utilized to optimize the model. All methods are implemented and trained using Pytorch on GeForce RTX 3090 GPUs. The results are the mean of five trainings.

Baselines Since there is no standard baselines for this task, we implement several methods of related tasks, including statistics-based (i.e., **TF-IDF (Salton and Buckley, 1988)**), RNN-based (i.e., **LSTM-*flat***, **LSTM-*hier***, **RETAIN** (Choi et al., 2016) and **DAG-ERC (Shen et al., 2021b)**), and transformer-based methods (i.e., **HiTANet** (Luo et al., 2020), **LSAN** (Ye et al., 2020)) and **DialogXL (Shen et al., 2021a)**. The RETAIN, HiTANet and LSAN are strong baselines for EHR-based medication recommendation or risk prediction. DAG-ERC and DialogXL are the SOTA methods at Emotion Recognition in Conversation (**ERC**). *Among them, LSTM-hier takes the dialogue structure into consideration, and LSAN and DialogXL are modified to incorporate disease knowledge*. Refer to Appendix B.1 for more details.

Evaluation Metrics We adopt two commonly used metrics, namely **Jaccard** and **F1** scores, to evaluate the model performance.

5.2 Main Results

Table 3 shows performances of all methods under the metric of Jaccard and F1 on four datasets. The results clearly indicate that DDN has achieved the best performances among all baselines. Particularly, DDN improves 6.35%, 5.14%, 3.95%, and 8.31% compared with the second best method (i.e., **DialogXL**) at Jaccard, respectively. Further,



(a) Jaccard on four datasets (b) F1 on four datasets

Figure 4: Performance comparison of DDN and its variants.

RETAIN and LSTM-*hier* outperform LSTM-*flat*, demonstrating the dialogue structure is important for the dialogue understanding. And LSAN, DialogXL outperforms HiTANet, indicating that disease knowledge is also essential for the dialogue modeling. Our well-designed model DDN considers both of the above and achieves the best performance. In addition, it is worth noting that the performance varies over three departments, which may attribute to the considerable difference of medication and disease frequencies between different departments.

5.3 Ablation Study

Figure 4 summarizes the contributions of QA Dialogue Graph and disease knowledge of our model. We notice that by removing the QA Dialogue Graph, the variant DDN$_{w/o\ DG}$ shows considerable performance decrease at both Jaccard and F1 compared with DDN, especially on three departments datasets. It demonstrates that dialogue graph structure is important for the medical information extraction in dialogue-based medication recommendation task. Similarly, by removing the Knowledge Graph module, DDN$_{w/o\ KG}$ also shows similar performance decrease trends, indicating that disease knowledge can improve the medication recommendation performance. This is reasonable and accords with the actual medication consultation situations.

5.4 Task Feasibility Analysis

To prove the feasibility of dialogue-based medication recommendation, we provide incomplete discourses to DDN during the inference process to explore whether the dialogue can provide necessary medical information. Figure 5 shows the model performances under different portions of discourses. We can see that with the increasing of dialogue discourse percentage, the performance gets better, especially within the first 20% and the last 20%.

		All Data		Respiratory		Gastroenterology		Dermatology	
Type of Model	Model	Jaccard	F1	Jaccard	F1	Jaccard	F1	Jaccard	F1
Statistics	TF-IDF(Salton and Buckley, 1988)	21.25±0.41	35.05±0.56	16.06±0.44	27.68±0.66	23.85±0.40	38.52±0.52	28.84±0.14	44.77±0.17
RNN-Based	LSTM-*flat*	27.50±1.09	42.54±1.22	18.07±0.44	30.18±0.64	31.31±1.33	47.18±1.59	32.69±1.71	48.55±1.18
	LSTM-*hier*	30.20±0.47	46.39±0.56	22.86±0.42	37.21±0.56	32.90±0.93	49.51±1.05	36.00±0.50	52.94±0.54
	RETAIN(Choi et al., 2016)	31.16±0.82	42.16±0.99	21.13±0.64	30.49±0.96	36.70±0.86	48.54±0.73	43.19±1.06	54.14±1.20
	DAG-ERC(Shen et al., 2021b)	29.08±0.56	44.05±0.70	23.74±0.76	35.71±1.02	36.16±0.46	53.80±0.52	31.18±1.05	47.52±1.23
Transformer	HiTANet(Luo et al., 2020)	30.75±0.69	44.57±0.83	22.01±1.04	33.62±1.44	33.95±1.26	48.39±1.26	39.17±1.93	53.41±2.21
	LSAN(Ye et al., 2020)	34.33±0.58	46.14±0.45	26.11±1.06	38.89±1.01	39.28±0.22	52.49±0.62	50.29±1.24	57.90±1.09
	DialogXL(Shen et al., 2021a)	36.27±0.34	53.23±0.40	27.12±0.24	42.67±0.36	40.91±0.14	58.06±0.15	48.68±0.81	65.48±0.66
	DDN(Ours)	**42.62±0.35**	**59.77±0.34**	**32.26±1.25**	**48.77±1.43**	**44.86±0.54**	**61.93±0.52**	**56.99±0.53**	**72.60±0.43**

Table 3: Performance (%) comparison of DDN with baseline methods over the overall and three departments datasets. The best result in each column is highlighted in boldface. The performance gain of our method over all baselines is statistically significant with $p < 0.05$ under t-test.

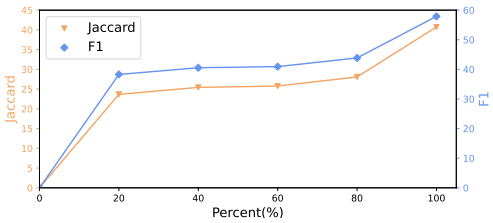


Figure 5: Average Jaccard scores on different percents(%) of dialogue discourse. In this setting, we choose dialogues with more than four turns in test set.

This may be because that the first and last parts of dialogue contain much patient complaints and symptoms that are closely related to the medications. The results demonstrate that recommending medication based on medical dialogues is feasible.

5.5 Error Analysis

Although we have elaborately designed a model for the task, the results are not so well satisfactory. So we make detailed analysis of the error cases in the test set. Table 4 summarizes the statistics of our defined five type of errors. We can see that (1) 86.38% of the cases (#3, #4, #5) predict wrong medications, which is mainly caused by DDN failing to distinguish the medications with similar effect. (2) 7.20% of the cases predict none labels, which can be attributed to that these dialogues provide a little disease-related information.

No.	Type of error	# Cases
#1	$P \subseteq \emptyset$	65(7.20%)
#2	$P \subset T$ & $P \not\subseteq \emptyset$	58(6.42%)
#3	$T \subset P$	182(20.16%)
#4	$T \not\subseteq P$ & $P \not\subseteq T$ & $P \cap T \not\subseteq \emptyset$	299(33.11%)
#5	$T \not\subseteq P$ & $P \not\subseteq T$ & $P \cap T \subseteq \emptyset$	299(33.11%)
Total	-	903

Table 4: The statistics of errors on test set. P and T are the predicted and golden label set, respectively.

Sample	Medications
Disease: Duodenitis **P :** Hello, doctor. In March this year, I had a duodenal ulcer, bleeding, and was hospitalized. Stomach rises a bit uncomfortable and bloating in the night a week recently. Is it recrudescent? **D :** Duodenal ulcers are indeed prone to recurrence or inflammation. **P :** Can you prescribe some medicine for me? I don't have time to go to the hospital right now. **D :** Besides what you said, do you have any other complaints? Like acid reflux, heartburn. **P :** No. What does heartburn mean? I don't have this feeling at ordinary times. At present, I wake up uncomfortably in some nights. **P :** Almost no symptoms during the day. **D :** I suggest you take [MASK], [MASK], [MASK].	**TF-IDF:** Omeprazole, Digestive enzymes, 2 Missed **LSTM-*flat*:** Omeprazole, 2 Missed **LSTM-*hier*:** Omeprazole, Mosapride, Digestive enzymes, 1 Missed **RETAIN:** Mosapride, 2 Missed **DAG-ERC:** Omeprazole, 2 Missed **HiTANet:** Mosapride, 2 Missed **LSAN:** Omeprazole, 2 Missed **DialogXL:** Omeprazole, 2 Missed **DDN(Ours):** Omeprazole, Mosapride, Glutamine

Figure 6: The sample is extracted from the DIALMED test set. Golden labels of this case are Omeprazole, Mosapride and Glutamine. The "Missed" means the medication is in golden labels but not be predicted, and the underlined drugs in red represent the predicted medications that are not in ground truth.

5.6 Case Study

We further provide a case study to illustrate the superiority of DDN. Figure 6 shows the medical dialogue and the medications recommended by all baselines and our method. The baselines either miss some medications, e.g., LSTM-*flat*, RETAIN, HiTANet, LSAN, or give the wrong drugs, e.g., TF-IDF, LSTM-*hier*. DDN takes full account of *Duodenitis*-related information from the dialogue (e.g., the symptoms in chief complaint and past medical history) and the external knowledge graph. It recommends *Omeprazole* (inhibiting gastric acid secretion) and *Mosapride* (promoting gastric dynamics), as well as *Glutamine* which is omitted by all baselines.

6 Conclusions

In this paper, we studied a new task, namely dialogue-based medication recommendation. First, we presented the first high-quality medical dialogue dataset DIALMED for this task. And then we implemented several baselines, as well as designed a dialogue structure and external disease knowledge aware model. Experimental results show that medication recommendation quality can be enhanced with the help of dialogue structure and external disease knowledge.

Ethical considerations

Data in DIALMED is publicly collected from Chunyuyisheng, and personal information (e.g., usernames) is preprocessed. The annotating process is as described in Section 3. Furthermore, to ensure the quality of dataset, we paid the annotators 1 yuan ($0.16 USD) per label. The applications of machine learning in medical treatment would inevitably raise ethical issues. But the research on AI medicine should not be stopped by this, since the purpose of such research is how to make machines better serve human beings. We have seen many advanced achievements (Lin et al., 2021; Li et al., 2021; Zhang et al., 2020; Liu et al., 2020; Lin et al., 2019; Xu et al., 2019; Wei et al., 2018) in this field. For this study, the ethical issue is that there may cause bad cases in practical application. However, individual errors could be reduced by making doctors responsible for decisions while machines are used as assistants.

Acknowledgements

This research was partially supported by National Key R&D Program of China under grant No. 2018AAA0102102, National Natural Science Foundation of China under grants No. 62176231 and 62106218, Zhejiang public welfare technology research project under grant No. LGF20F020013. The authors would thank Ruochen Yan for her help in data processing.

References

Yang An, Liang Zhang, Haoyu Yang, Leilei Sun, Bo Jin, Chuanren Liu, Ruiyun Yu, and Xiaopeng Wei. 2021. Prediction of treatment medicines with dual adaptive sequential networks. *IEEE Transactions on Knowledge and Data Engineering*.

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *AAAI*, volume 34, pages 7521–7528.

Edward Choi, Mohammad Taha Bahadori, Joshua A Kulas, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *NIPS*.

Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. 2019. Extracting symptoms and their status from clinical conversations. *arXiv preprint arXiv:1906.02239*.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue discourse-aware graph model and data augmentation for meeting summarization. *IJCAI*, pages 3808–3814.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Yong He, Cheng Wang, Nan Li, and Zhenyu Zeng. 2020. Attention and memory-augmented networks for dual-view sequential learning. In *SIGKDD*, pages 125–134.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan Black, and Yulia Tsvetkov. 2021. Dialograph: Incorporating interpretable strategy-graph networks into negotiation dialogues. *arXiv preprint arXiv:2106.00920*.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Hung Le, Truyen Tran, and Svetha Venkatesh. 2018. Dual memory neural computer for asynchronous two-view sequential learning. In *SIGKDD*, pages 1637–1645.

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*, page 544–554.

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 *AAAI*, volume 35, pages 13362–13370.

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 *EMNLP-IJCNLP*, pages 5036–5045.

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 preprint arXiv:2010.07497*.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *SIGKDD*.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2020. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. *arXiv preprint arXiv:2012.13260*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Junyuan Shang, Tengfei Ma, Cao Xiao, and Jimeng Sun. 2019a. Pre-training of graph augmented transformers for medication recommendation.

Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. 2019b. Gamenet: Graph augmented memory networks for recommending medication combination. In *AAAI*, volume 33, pages 1126–1133.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *AAAI*, volume 35, pages 13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

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 AAAI Conference on Artificial Intelligence*, volume 34, pages 8838–8845.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*.

Shanshan Wang, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2019a. Order-free medicine combination prediction with graph convolutional reinforcement learning. In *CIKM*, pages 1623–1632.

Shuai Wang. 2020. Seqmed: Recommending medication combination with sequence generative adversarial nets. In *BIBM*, pages 2664–2671. IEEE.

Xiang Wang, Xiangnan He, Yixin Cao, Meng Liu, and Tat-Seng Chua. 2019b. Kgat: Knowledge graph attention network for recommendation. In *SIGKDD*, pages 950–958.

Yanda Wang, Weitong Chen, Dechang Pi, and Lin Yue. 2021. Adversarially regularized medication recommendation model with multi-hop memory network. *Knowledge and Information Systems*, 63(1):125–142.

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 *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 *AAAI*, volume 33, pages 7346–7353.

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.

Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. 2021. Safedrug: Dual molecular graph encoders for safe drug recommendations. In *IJCAI*.

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 preprint arXiv:2005.05442*.

Muchao Ye, Junyu Luo, Cao Xiao, and Fenglong Ma. 2020. Lsan: Modeling long-term dependencies and short-term correlations with hierarchical attention for risk prediction. In *CIKM*, pages 1753–1762.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: Large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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 *ACL*, pages 6460–6469.

Yutao Zhang, Robert Chen, Jie Tang, Walter F Stewart, and Jimeng Sun. 2017. Leap: learning to prescribe effective and safe treatment combinations for multimorbidity. In *SIGKDD*, pages 1315–1324.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.

A Corpus

A.1 Details of corpus construction

First of all, diseases and related medications were identified in a dialogue. Secondly, we selected and annotated those dialogues containing drugs in our medication list. To speed up tagging process, we built an annotation tool based on this task. For each raw medical dialogue, the annotators need to annotate the disease of patients and medications recommended by doctors. We believe that the context after the doctor recommending the drug is not meaningful for drug inference. Due to the emergence of new medications in the labeling process and existence of ambiguity on recommendation, two additional annotation processes were carried out. Next we will focus on the processing of diseases and medications.

Disease Processing. With the guidance of a doctor, we select 16 diseases from 3 departments (i.e., respiratory, gastroenterology and dermatology) with following reasons: (1) they are common diseases and research on them have more practical value. (2) they could be consulted online and there are abundant medication consultations. As described by Section **Corpus Description**, we normalize the diseases to improve the quality of DIALMED, e.g., **chronic gastritis** and **acute gastritis** are mapped to **gastritis**. The dialogues without explicit disease information or diseases in our scope were marked as **None or Others**. We mark one disease according to the chief complaint of patients who have more than one disease, because patients have only one complaint in most diagnostic scenarios.

Medication Processing. As for medications, the ones we choose are commonly prescribed by doctors. Considering the differences between traditional Chinese medicines and Western medicine, both are included to achieve complementary advantages. Since there are many generic names, trade names and colloquial expressions for the same drug in conversations, it is significant to normalize the drug to a single label. For example, **Omeprazole enteric-coated tablet** and **Omeprazole enteric-coated capsule** could be mapped to **Omeprazole**. For compound medicines, we combine drugs that have the same ingredients into one, e.g., Tylenol represents all medicines that contain acetaminophen, pseudoephedrine hydrochloride, dextromethorphan hydrobromide and chlorpheniramine maleate. Due to space constraints, more normalization of diseases and medications could be found in our repository[6].

B Experiments

B.1 Baselines

- **TF-IDF**. This is a traditional bag-of-word model for text classification. We view each dialogue as text and the corresponding medication as label, and train a classification model based on TF-IDF features of words.
- **LSTM-*flat***. This is a LSTM-based method. It concatenates all the sentences in a dialogue as a long sentence and feeds the long sentence into the BiLSTM to get the dialogue embedding for medication prediction.
- **LSTM-*hier***. This is also a LSTM-based method. Different from LSTM-*flat*, it uses a hierarchical BiLSTM where each word in an utterance are fed into BiLSTM to get the utterance embedding and then the utterances are fed into another BiLSTM to get the final dialogue embedding. It captures both word-level and utterance-level dependencies.
- **RETAIN**. This is a RNN-based EHR medication recommendation method using on a two-level neural attention network that detects influential past visits. In the current scenario, it is used to model the dialogues.
- **DAG-ERC**. This method designed a directed acyclic neural network to model the information flow between long-distance conversation background and nearby context. Following the implementation in (Shen et al., 2021b), the features of utterances extracted from fine-tuning RoBERTa are inputted in model while the model structure is RNN based, so DAG-ERC is regraded as a RNN-based model.
- **HiTANet**. This is a Transformer-based risk prediction approach on EHR, which model time information in local and global stages. We transform this method to model the hidden temporal information in medical dialogues.
- **LSAN**. This is also a Transformer-based risk prediction approach, to model the hierarchical structure of EHR data. We modified this

[6] https://github.com/Hhhhhhhzf

method to model the hierarchical structure in medical dialogues and add disease module of DDN to encoder the external knowledge.

- **DialogXL**. This method improves XLNet with enhanced memory and dialog-aware self-attention. We modify the softmax layer to sigmoid layer in this model to fit the multi-label task in medication recommendation and add the disease module of DDN.

- **DDN**. This is our proposed model. It utilizes the dialogue structure and external disease knowledge to enhance the dialogue-based medication recommendation performance.

B.2 Evaluation Metrics

$$\text{Jaccard} = \frac{1}{|D|}\sum_{k=1}^{|D|}\frac{|Y^{(k)} \cap \hat{Y}^{(k)}|}{|Y^{(k)} \cup \hat{Y}^{(k)}|} \tag{9}$$

$$\text{F1} = \frac{1}{|D|}\sum_{k=1}^{|D|}\frac{2 \cdot \text{P}^{(k)} \cdot \text{R}^{(k)}}{\text{P}^{(k)} + \text{R}^{(k)}} \tag{10}$$

where $|D|$ is the number of dialogues in the test set. $Y^{(k)}$ represents the ground truth medication set of the kth dialogue, and $\hat{Y}^{(k)}$ represents the predicted medication set of the kth dialogue by the model. $\text{P}^{(k)}$, $\text{R}^{(k)}$ represents the Precision and Recall of the kth dialogue, respectively.

B.3 Additional Experiment on DDI

Medication combination recommendation would trigger the Drug-Drug Interaction (**DDI**) inevitably, which might lead to adverse outcomes. To this end, we explore the DDI in DIALMED. And we follow the previous work (Shang et al., 2019b) to give the DDI rate definition (smaller value means better).

$$DDIRate = \frac{\sum_{k}^{N}\sum_{i,j}|\{(c_i, c_j) \in \hat{Y}^{(k)}|(c_i, c_j) \in \mathcal{E}_d\}|}{\sum_{k}^{N}\sum_{i,j}1} \tag{11}$$

where the set will count each medication pair (c_i, c_j) in recommendation set $\hat{Y}$ if the pair belongs to edge set $\mathcal{E}_d$ of the DDI graph. Here N is the size of test dataset. In addition, DDI relationships among medications in DIALMED are collected from YAOZH [7], a medical data retrieval system.

The evaluation results are shown in Table 5. We could find that ground truth DDI rate is very small (compared to the 8.08% in MIMIC-III (Yang et al., 2021)), which may lead to the low rate on models. In view of this situation, we think it is no need for additional efforts to control the DDI rate at the current stage. Considering for the future research, we open source our DDI relationship graph in our repository.

	DDI Rate			
Model	All Data	Respiratory	Gastroenterology	Dermatology
G.T.	1.12	0.78	2.06	0.74
TF-IDF	1.10	0.46	2.01	0.51
LSTM-*flat*	0.58	1.36	0.93	0.00
LSTM-*hier*	1.02	0.11	0.91	0.65
RETAIN	1.92	1.12	1.89	0.00
DAG-ERC	0.81	1.01	1.53	0.48
HiTANet	0.45	1.49	1.09	0.50
LSAN	1.57	0.00	1.62	0.48
DialogXL	1.34	1.09	1.59	0.40
DDN	1.90	0.20	1.54	0.47

Table 5: DDI Rate (%) comparison on DIALMED. G.T. represents the Ground Truth.

C Task

C.1 Medical Utility

Medical treatment includes a number of steps: registration, examination, image reading, report interpretation, diagnosis, prescription and so on. AI medicine could help optimize resource allocation and improve efficiency in all aspects of health care. To this end, there are two kinds of computer aided diagnosis system, image diagnosis and text diagnosis. Due to the higher threshold of diagnosis, current researches are more inclined to image analysis, and there is still a lot of room for development in text diagnosis. Conversations in outpatient clinics are not reserved and involved many severe data privacy implications, leading to dialogue-based drug recommendation mainly oriented to telemedicine. The medical dialogue system, as a assistant of doctors, could give auxiliary medication suggestions based on the contexts when doctors and patients are communicating with each other.

D Statistics

D.1 Ratio of consulting for medications

The ratio of the patients to consult for medications is calculated with regular expressions. In the first place, 10,0000 different medical conversations from our dialogue corpus based on random sampling are fetched. For every dialogue, we apply the regular expression (e.g., "[Ww]hat (medication|drug|medicine) should I (take|eat)") on the utterances spoken by the patient and assume

[7] https://db.yaozh.com/interaction

that it is a case of consulting for drugs if the regular expression matches. The regular expressions are collected based on our observation and understanding of data. More regular expressions could be found in our repository.

D.2 Complete Corpus Statistics

The frequency of all diseases and medications is shown in Figure 7 & 8.

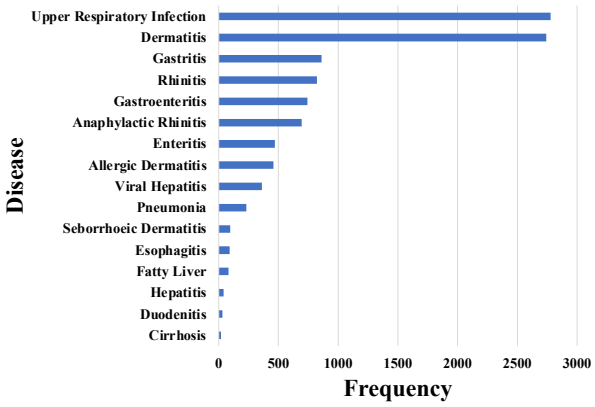


Figure 7: The frequency of all diseases.

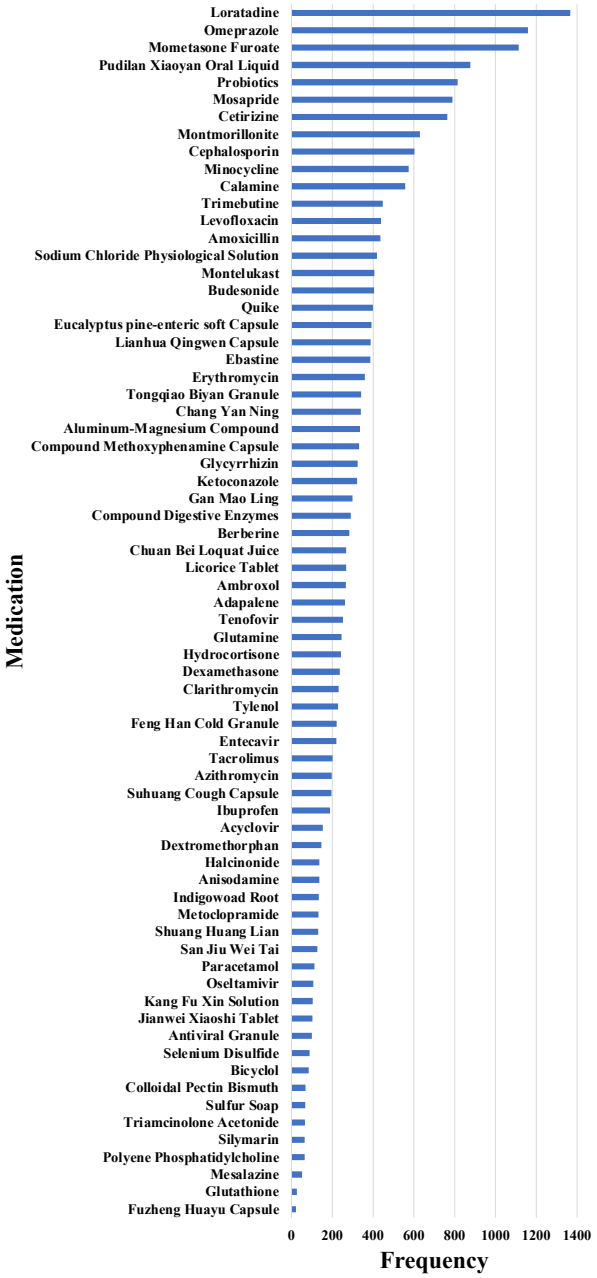


Figure 8: The frequency of all medications. The names are translated from Chinese.