Knowledge-driven Chinese Medical Dialogue Text Generation with Corpus Construction

Yupeng Liu*, Gang Liu, Yuhao Zhang, Jiang Zhang, Han Zhang, Yi Qin

Harbin University of Science and Technology; No. 52, Xuefu Road, Nangang District, Harbin 150080, China

Keywords: Gastroenterology; Intention; Entity; Slot-value; Knowledge-driven

Citation: Liu Y.P., Liu G., Zhang Y.H., et al.: Knowledge-driven Chinese Medical Dialogue Text Generation with Corpus Construction. Data Intelligence, 2026, Vol. 8, Art. No.: 20250101. DOI: https://doi.org/10.3724/2096-7004.di.2025.0101

ABSTRACT

This study proposes a multi-task multi-label doctor-patient dialogue dataset (TLMedDialog) that focuses on gastrointestinal medicine. First, the study describes the labelling specification of the doctor-patient dialogue dataset and then discusses the dataset's labelling process and statistics. Based on this dataset, the study then develops a knowledge-based medical dialogue model. The current medical dialogue text generation process has many problems, such as insufficient medical entities in the responses, ambiguity in medical entity relationships, and a lack of role information etc. This study proposes using entity prediction, dialogue history, and medical knowledge graphs to assist in the generation of medical dialogue text in order to improve the credibility and fluency of the process. The study also performs corresponding performance comparisons for natural language understanding (including intent classification and medical entity recognition (MER)) and dialog policy learning.

1. INTRODUCTION

The medical dialogue system has attracted wide attention from both academic and medical circles. The construction of a complete medical dialogue system involves many aspects of medical language analysis technology, such as intention understanding and named entity recognition. Medical dialogue text generation is based on analysis of multiple rounds of doctor-patient dialogue generation tasks. Specifically, it automatically generates natural and reasonable responses according to historical information on doctor-patient dialogue to assist users in completing specific tasks in the information interaction process. The end-to-end framework has achieved good results in the field of machine translation, and researchers have thus applied it to doctor-patient dialogue generation tasks. They have used historical doctor-patient dialogue

*Corresponding author: Yupeng Liu (E-mail: flyeagle99@126.com).

© 2026 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

information as the input of the encoder and conducted encoder and decoder mapping to obtain the doctor's reply. However, some meaningless replies are frequently generated, and these may be considered intentions and entities in conversational replies. In principle, the identification of intent and medical entities can enable us to better understand medical texts [1].

To enable the provision of more effective medical consultation services, this study provides a multi-task multi-label medical dialogue dataset as an important basis for the construction of smart medicine [1-6]. This study first describes the relevant research on the medical dialogue dataset. The dataset specifications, annotation process, and statistics are then discussed. Subsequently, a knowledge-driven Chinese medical dialogue text generation model is proposed. Finally, the validity of the dataset is verified via experiments, and the results of the WeChat mini program display are presented. The contributions of the dataset and knowledge-driven medical dialogue text generation model proposed in this study are as follows:

1) This study constructs a dataset with abundant labels, including multiple interrelated labels for medical tasks (intention, slot, slot value, and entity). Provide data reference for building interpretable and training data for building more accurate and intelligent medical dialogue systems. The dataset proposed in this study is applicable to a variety of medical dialogue tasks including natural-language understanding (NLU), Dialog Policy Learning (DPL), and natural language generation (NLG).
2) This study introduces medical knowledge (dialogue history, entity prediction, and knowledge graph) to assist in the generation of medical dialogue to solve insufficiencies of knowledge modelling ability. The multi-task medical dialogue model integrates knowledge models (doctor and patient intentions, important entities appearing in the dialogue such as symptoms, diseases, drugs etc.) into dialogue generation to generate high-quality responses.

2. RELATED WORK

2.1 NLU, DPL, And NLG Medical Datasets

NLU plays a crucial role in text understanding systems as well as in recommendation, question-answer, text generation, and medical inquiry systems.

Terminology. In this paper, a dialogue denotes one complete doctor-patient consultation session. An utterance denotes one continuous speaking segment produced by a single speaker (doctor or patient) and is the basic unit for annotation and statistics. To avoid ambiguity, we use utterance consistently instead of turn throughout the paper.

The CMDD medical dataset [1], first proposed by Lin et al. for NLU in the medical field, contains 2,067 doctor-patient utterances. Owing to the arbitrariness of spoken language, symptom entities in dialogue are generally non-standard medical expressions, which are standardized during the annotation process. In addition, the symptom entity in a doctor–patient dialogue may not be a real symptom; therefore, it is also necessary to mark whether the symptom entity is a real symptom and use 'True' to indicate when it is. This dataset promotes research into conversation-based symptom diagnosis and lays

the foundation for subsequent development. The MSL medical dataset [2] proposed by Shi focuses on the slot-filling task, which aims to classify sentences into predefined categories (slot-value pairs). This dataset consists of 1,652 doctor-patient utterances. To construct the medical slot-filling model, the doctor-patient dialogue is labelled by marking the disease entities appearing in each sentence with slot values. Owing to the randomness of spoken language, the marked slot values need to be converted into standardized expressions. The MIE medical dataset [3] proposed by Zhang extracts information from a doctor-patient dialogue in order to transform it into electronic medical records. This dataset contains 1,120 doctor-patient utterances. In the past, medical entities were often labelled in the form of BIO; however, such labelling could not mark entities with unclear or discontinuous expressions, which is not rare in oral expression. Therefore, the labelling method for this dataset was performed using a sliding window instead of sequential labelling. It has four categories: 'symptoms', 'surgery', 'tests', and 'others'. In addition, the status of symptoms mentioned in the doctor-patient conversation can be marked as 'present' or 'absent' according to whether the patient actually has them.

The DPL can enable a dialogue system to generate more meaningful responses. The MZ medical dataset [4] proposed by Liu is used for dialogue strategy learning aimed at doctor-patient dialogue generation based on a template. The dataset consists of 710 doctor-patient utterances, consisting of self-reports from patients and doctor-patient conversation data. Symptoms in the data are labelled, and symptoms mentioned in patients' self-reports are classified as 'explicit' whereas symptoms mentioned in doctor-patient conversations are classified as 'implicit'.

Natural language generation, which can generate a response based on a preceding dialog, is often used in task-based dialog systems using pipeline and end-to-end methods. The COVID-CN dataset [5] proposed by Zeng et al. uses an end-to-end approach for NLG. This dataset contains 1,088 utterances from COVID-19-related patient-doctor dialogues. This dataset is not specially labelled, but the corresponding doctor-patient conversations are collected, and pre-training models such as GPT are used to generate the conversations. The MedDG medical dataset [6] proposed by Liu is also used for dialog generation. Unlike traditional dialog generation, this dataset is used for two types of tasks: entity prediction and reply generation using predicted entities. When the data are labelled, five entity categories are defined: 'disease', 'symptom', 'drug', 'test', and 'attribute'. A multi-label classification of entities is performed using the model to enable entity prediction through labelled entity information, so that the conversation generation contains more medical entities.

The datasets provided in [1-4] are primarily focused on diagnosis and symptom extraction, while [5-6] concentrate on specific diseases (such as COVID-19) or entity-driven dialogue generation. This paper includes more comprehensive intents and slot values, and explores gastrointestinal diseases. It has greater domain adaptability and annotation accuracy compared to existing datasets.

2.2 End-to-end Conversation Model

Google's Vinyals et al. [7] were the first to apply the sequence-to-sequence model of machine translation to the dialogue generation task; it treats the reply generation problem as a translation problem. It is unreasonable to encode all contexts indiscriminately. Tian et al. [8] proposed the WSEQ model, in which cosine similarity is used to measure the importance of context clauses. Xing et al. [9] proposed the HRAN model, which introduced the traditional attention mechanism to capture related clauses and words in historical information at both sentence and word levels, respectively. Chen et al. [10] proposed the HVMN model, which uses a memory network to enable the model to retrieve the relevant context during decoding. Zhang et al. [11] proposed the ReCoSa model, which applies a self-attention mechanism to solve the long-distance dependence problem and capture relevant historical information. The role of external information (including common sense information and the speaker's past experiences) is the biggest difference between people and models. Kumar et al. [12] proposed a DMN model that combines common sense information to generate more human responses. The topic constraint feature ensures consistency between the generated replies and dialogue semantics. Mou et al. [13] proposed the seq2BF model, which uses point-by-point mutual information to predict a noun as a topic word and then generates a reply containing the topic word. Subject terms do not necessarily appear in the generated replies and can be used as implicit constraints for them. Yao et al. [14] built Seq2Seq by introducing a network unit that controls themes, which are reflected in status updates at every moment. Xing et al. [15] proposed a TA-Seq2Seq model and introduced an LDA topic model to increase the occurrence probability of subject words using the attention mechanism in decoding. Dziri et al. [16] proposed the THRED model, which combines historical and thematic information to make responses more diversified.

Intention plays a key role in conversation, and determining the speaker's motivation is the only way to generate a reasonable response. Jo et al. [17] proposed an unsupervised model based on a conversational behavior sequence, while Zhang et al. [18] proposed an HRG model that uses a layered response mechanism to capture intent in a conversation.

End-to-end conversation models have strong single-task modeling capability, but they don't model knowledge well in complex tasks such as medical dialogue. In terms of task, only some medical tasks are considered, and it is not possible to model the overall medical dialogue tasks that are interrelated.

2.3 Dialogue Model Based On Knowledge Base

Wen et al. [19] modularized each component of the model and modelled it using neural networks, thus constituting a differentiable end-to-end model. Subsequently, Bordes et al. [20] constructed an end-to-end memory network model, treated task-based dialogue as a reading comprehension task, and reasoned through multiple rounds of reasoning by considering the user's discourse, conversation history, and system response as questions, contexts, and answers, respectively. Maddotto et al. [21] also used an end-to-end memory network model and input knowledge base information into the memory network to make the model more intelligent. Eric et al. [22] extracted information from a knowledge base and constructed a

key-value storage network model. Lei et al. [23] proposed an end-to-end generation model that uses only the conversation state rather than the conversation behaviour representation. First, the model encodes the conversation history. Then, the conversation state is generated via LSTM and CopyNet. Finally, the model uses the conversation state to generate natural and smooth responses.

There are many types of external knowledge, including structured knowledge maps, unstructured background information, information about people, and topic information. Wu et al. [24] constructed the DuConv dataset and used knowledge graphs and conversation objectives to guide the dialogue. Dinan et al. [25] created the WoW dataset; the model needs to use Wikipedia to generate replies to ensure that they include relevant information. Zhou et al. [26] constructed a Chinese multi-domain KdConv dataset, which is mainly used in the study of knowledge-driven multi-round dialogue. Xing et al. [15] studied several topics highlighted in multiple rounds of dialogue and proposed a Twitter LDA model to integrate topic information into the dialogue system. Zhou et al. [27] focused on topic information in dialogues and created a recommendation dialogue system based on topic clue words. Qian et al. [28] found that dialogue systems contain different types of role information and therefore proposed building a dialogue system that is consistent with a given role. Zhang et al. [29] manually collected a role-based dialog dataset using Mechanical Turk; to ensure a consistent conversation state, the role information contained in a conversation can be used as external knowledge to be integrated into the dialogue model. Mazare et al. [30] built a large-scale role-based dialogue dataset to enrich dialogue scenes. To improve the quality of the generated responses, Dinan et al. [25] used a two-stage framework of knowledge selection and dialogue generation to complete the dialogue. Kim et al. [31] proposed an SKT with knowledge prior and posterior distributions, which can improve knowledge-selection performance. Lian et al. [32] adopted BOW Loss to enhance the correlation between knowledge and true responses.

Unlike general medical dialogue data, this paper focuses on medical dialogue data of specific gastrointestinal diseases. Analyze and construct data for the specific domain. Unlike the traditional large language models, the model proposed in this paper can perform NLU and DPL in medical dialogue processes and provide feature embedding for the final NLG task.

3. PROBLEM FORMULATION

3.1 Labelling Specification

The raw data for this study were collected from the online consultation platform Haodaifu (www.haodf.com). The dataset was marked with a fine granularity. The intention, slot value, and entity scope of each utterance spoken by patients and doctors are annotated at the utterance level. Intents are considered to determine the speaker's motivation, which helps us to analyze the semantics of the generated statements. Patient and doctor intents are specified by role; common intents include inquiries and chitchats. Slots are designed to give each sentence a domain equivalent to an entity meaning and specify several slots for conversation generation tasks. The slot values and entity ranges comprise specific tagging information for each sentence and can be used for tasks such as named entity recognition, disease, and symptom prediction.

A doctor-patient dialogue consists of multiple utterances; these do not necessarily alternate strictly, since one speaker may produce several consecutive utterances. An utterance can have multiple intentions, and an intention can have multiple slot values. Patient and doctor intentions are shown in Tables 1 and 2, respectively.

Table 1. Definition of patient intent.

Intention	Role	Explanation	Example
Inform	Patient	The patient informs the doctor	da bian zheng chang, ye bu e xin 大便正常, 也不恶心
Inquire	Patient	The patient asks the doctor	wo de le ji xing fu xie,gai ru he zhi liao? 我得了急性腹泻, 该如何治疗？
Want confirm	Patient	Want to get the doctor's confirmation	shi chang dao xi jun bu ping heng ma? 是肠道细菌不平衡吗？
Confirm	Patient	Reply to the doctor for confirmation	you 有
Deny	Patient	Deny the doctor's confirmation	fa shao mei you 发烧没有
Chitchat	Patient	Chitchat between patient and doctor	hao de, xie xie 好的, 谢谢

Table 2. Definition of doctor intent.

Intention	Role	Explanation	Example
Diagnosis	Doctor	Doctor diagnoses patient	zhe shi shi guan fan liu 这是食管反流
Recommendation	Doctor	Advice given by a doctor to a patient	jian yi chi dian ao mei la zuo 建议吃点奥美拉唑
Not recommendation	Doctor	Things that doctors don't advise patients to do	bi mian shi yong ci ji xing shi wu 避免使用刺激性食物
Want confirm	Doctor	Want confirmation	bian mi you me 便秘有吗
Confirm	Doctor	Reply to the patient for confirmation	shi de 是的
Deny	Doctor	Deny the patient's confirmation	bu ke yi 不可以
Inquire	Doctor	The doctor questions the patient	hai you qi ta zheng Zhuang ma? 还有其他症状吗？
Chitchat	Doctor	Chitchat between doctor and patient	hao de, xie xie 好的, 谢谢

The slot-value pairs are divided into slot and slot values. In a medical dialogue, slot–value pairs limit the content communicated by doctors or patients, and convey the semantic information being expressed. Slots can be regarded as a category, such as 'diseases', 'symptoms', and 'drugs'. Slots represent specific manifestations, and each slot position can have multiple slot values. Slot definitions are listed in Table 3. A slot value can be obtained only through an analysis of a specific sentence.

Table 3. Slot definition.

Slot	Slot value	Explanation	Example
Symptom	Multiple values can be identified from statements or with reference to standard medical terms	Statement for symptoms that may occur	du qi zhou wei【yin yin zuo tong】 肚脐周围【隐隐作痛】
Disease	Multiple values can be identified from statements or with reference to standard medical terms	Statements that refer to diseases that may not appear explicitly	shu yu 【gao zhi xing yi xian yan】 属于【高脂性胰腺炎】。
Drug	Identified from the statement; there can be multiple values	The name of the drug mentioned in the statement	ci shu duo jiu chi 【si mi da】 次数多就吃【思密达】
Check	Identified from the statement; there can be multiple values	The statement mentions the check that was done	You cha【xue chang gui】ma 有查【血常规】吗
Symptomatic site	Identified from the statement; there can be multiple values	Statement where symptoms occur	【zuo ce lei gu bu wei】hen teng 【左侧肋骨部位】很疼
Food name	Identified from the statement; there can be multiple values	The food mentioned in the statement	man xing chang wei yan hui fu qi ke yi chi【zhi shi】ma 慢性肠胃炎恢复期可以吃【芝士】吗
Treatment site	Identified from the statement	The treatment site mentioned in the statement	【san jia yi yuan】wu tong wei jing yi ban qi bai duo 【三甲医院】无痛胃镜一般七百多
Department	Identified from the statement	The department mentioned in the statement	dao【pi fu ke】kan yi xia 到【皮肤科】看一下
Body part	Identified from the statement	The part of the body mentioned in the statement	bu hui qian che dao【zuo bian jian bu】 不会牵扯到【左边肩部】
Dosage of drug	Identified from the statement	The dosage recommended by the doctor	pi wei xiu an pian【yi ci yi pian】 匹维溴铵片【一次一片】

Table 3. *Continued.*

Slot	Slot value	Explanation	Example
Frequency of medication	Identified from the statement	The frequency of medication recommended by the doctor	pei fei kang jiao nang, 【yi tian liang ci】 培菲康胶囊, 【一天两次】
Preventive measure	Identified from the statement	The advice provided in the statement for alleviating the condition	【shao chi dian rou zhi pin, he dou zhi pin】, 【duo chi shu cai shui guo】 【少吃点肉制品, 和豆制品】, 【多吃蔬菜水果】
Drug class	Identified from the statement	Class of drug mentioned in the statement	【a mo xi lin】shi xiao yan yao 【阿莫西林】是消炎药
State	Positive, negative, other	Whether the proposed disease or symptom is real	-
Time	Identified from the statement	The time and duration of symptoms described by the patient	【liang san tian】 le. 【两三天】了。
Behaviour	Identified from the statement	An expression that refers to an action in a statement	yin yin 【zuo tong】 隐隐【作痛】 shi liang 【shao chi】 适量【少吃】

3.2 Data Annotation Process

Doctor-patient dialogue data were annotated according to the information defined in the annotation specification.

Two versions of the dataset were used: an unlabelled TLMedDialog (3,640 dialogues) and a labelled TLMedDialog (2,347 dialogues). The former was used to fine-tune the model, and the latter was used to evaluate the performance of the benchmark model. Figure 1 presents a flowchart of the labelling.

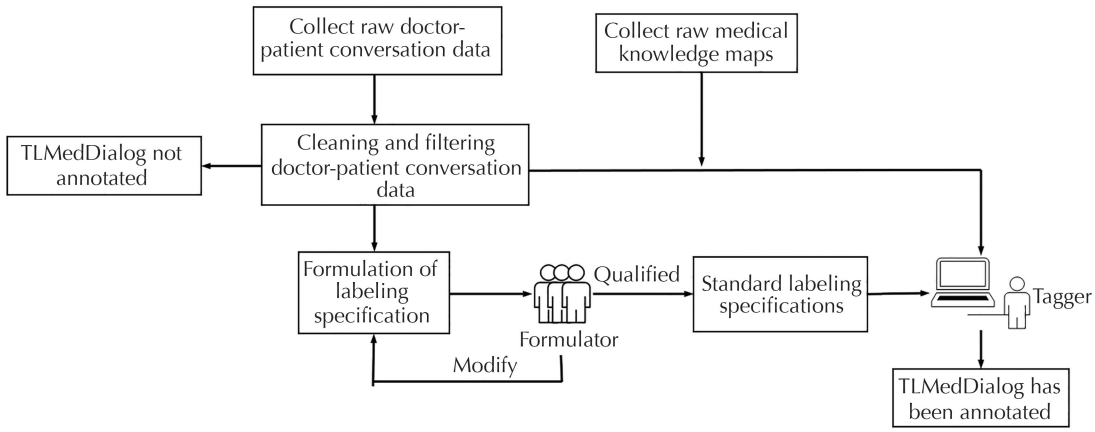


Figure 1. Annotation flowchart.

3.3 Data Statistics

Statistical analysis was performed on the labelled doctor-patient dialogue data to illustrate the characteristics of the dialogue dataset. Table 4 presents the statistics. Figure 2 shows the number of utterances distributed across different intent and slot types. For consistency, all average statistics are recomputed using unified denominators, and label-related totals are reported on the labeled subset only.

Table 4. Dataset statistics.

Dataset	Fine-tune	Train	Valid	Test	Total
Dialogue	3640	1463	487	487	6077
Utterance	42239	14863	4783	5125	67010
Utterance/dialogue	11.60	10.16	9.82	10.52	11.03
Character/dialogue	301.33	310.80	305.45	309.74	308.85
Character/utterance	20.46	19.58	19.34	19.44	19.60
Label/dialogue	-	30.58	31.78	30.15	30.73
Label/utterance	-	2.00	1.94	1.98	1.98

Note: Utterance denotes one continuous speaking segment by a single speaker. Total Dialogue and Total Utterance are computed over all subsets (Fine-tune, Train, Valid, and Test). Label-related statistics (Label/dialogue and Label/utterance) are computed on the labeled subsets only (Train, Valid, and Test), because the Fine-tune subset is unlabeled.

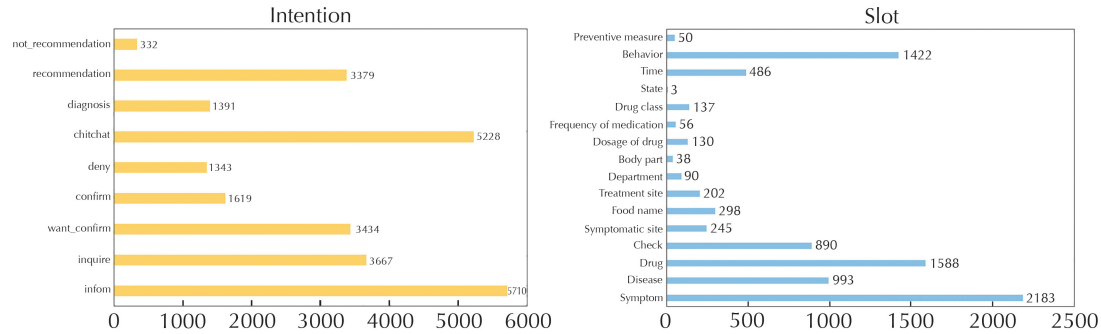


Figure 2. Intent and slot distribution diagram.

4. THE PROPOSED APPROACH

The overall model framework of this paper is shown in Figure 3, including the Shared embedding module, Entity prediction module, and Coding fusion module. The pipeline model sequentially connects the output of each module to the input of the next module. Considering the essential impact of medical knowledge and the difficulty of reproducing code, a simple and open-source GPT-2 is adopted as the base model, which can also be quickly migrated to other open-source models. The medical dialogue generation

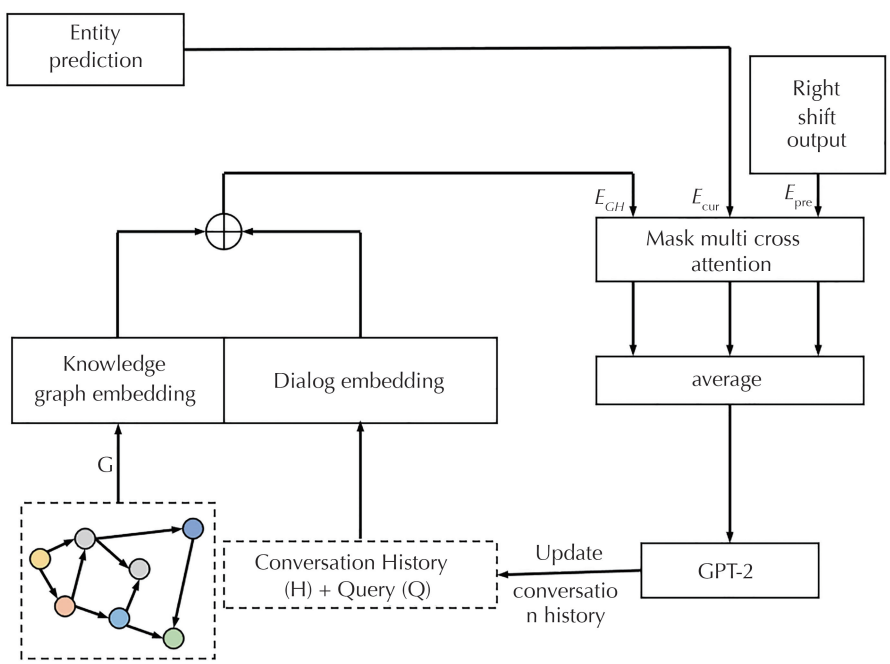


Figure 3. Entity-aware medical dialogue model based on knowledge graph.

task needs to provide the history of the doctor-patient dialogue (represented by H); this consists of multiple utterances and can be represented as $H = \{U_1, S_1, \ldots, U_i, S_i, \ldots, U_{|H|}, S_{|H|}\}$, where U_i and S_i represent the utterances from the patient and the doctor, respectively.Given the external knowledge graph, the set of medical entities is represented by $E = \{e_1, e_2, \ldots, e_i, \ldots, e_{|E|}\}$. The knowledge graph can be represented by triples containing head entity h, relation r, and tail entity t. Each knowledge is formalized as a triple $< h, r, t>$, defined as $\beta_{i,j}$, and triples with e_i as head entity are represented as $G(e_i) = \{\beta_{i,1}, \beta_{i,2}, \ldots, \beta_{i,j}, \ldots, \beta_{i,|G(e_i)|}\}$. Each piece of knowledge is composed of a series of words after word segmentation, which can be expressed as $\beta_{i,j} = \{E_{i,j,1}, E_{i,j,2}, \ldots, E_{i,j,k}, \ldots, E_{i,j,|\beta_{i,j}|}\}$. According to the dialogue history H and the current patient query, denoted as Q, the model must generate a reply Y, denoted as $Y = \{y_1, y_2, \ldots, y_i, \ldots, y_{|Y|}\}$, where y_i is the i-th word in reply Y.

Shared embedded module: The knowledge graph G, the conversation history H, and the current patient query Q are taken together as inputs, and the input representation is constructed as illustrated in Figure 4, incorporating token, position, entity, and relation embeddings. The first part of the sequence contains specific conversation information (conversation history and current patient query), and the second part contains information related to the knowledge graph. These embeddings are used to encode both the dialogue information and the structural information of the knowledge graph, enabling their effective integration.

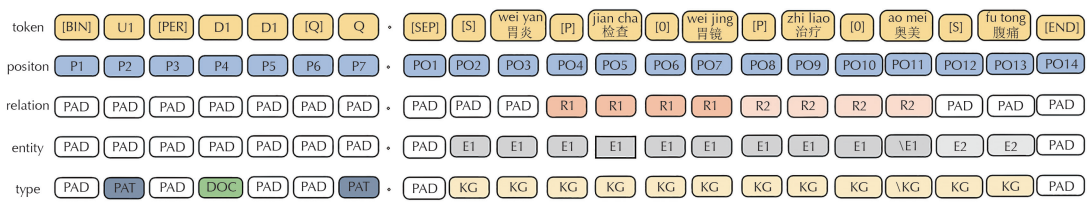


Figure 4. Construction of the input representation with multi-source embeddings.

Entity prediction module: This module predicts entity information in the current response based on the history of the doctor-patient conversation. In the coding layer (shown in Figure 5), different pre-training models can be used to encode the history of the doctor-patient dialogue. For the pre-training model in the general domain, unlabelled doctor-patient data can be used to continue the pre-training in order to improve the generalisation ability of the model for medical tasks.

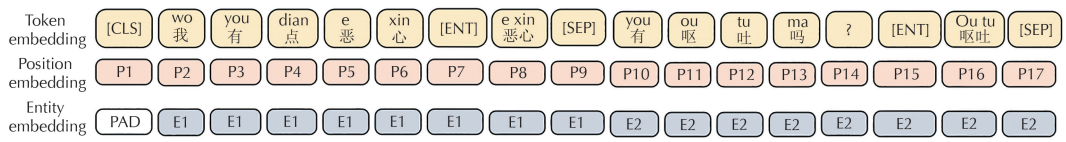


Figure 5. Input sequence in entity prediction module.

At the encoding layer, the conversation history is $U_{1:N} = \{U_1, U_2, \ldots, U_N\}$, and the historical entity is $E_{1:K} = \{E_1, E_2, \ldots, E_K\}$, where K is the number of historical entities. The context vector by which the encoder obtains the conversation history is expressed as follows:

$$H = Encoder(U_{1:N}, E_{1:K}) \tag{1}$$

The vector is transformed through the linear layer, and a sigmoid function is used to classify the model with multiple labels, which is expressed as follows:

$$\begin{cases} g = Linear(H) \\ p = Sigmoid(w \cdot g) \end{cases} \tag{2}$$

where $g \in \mathbb{R}^d$ is the logit vector obtained by applying a linear transformation to the context vector H, i.e., $g = Linear(H)$, and w denotes the weight vector of the linear layer. The sigmoid function is then applied to obtain the probability of each label. The binary cross-entropy loss is used to calculate the loss, as follows:

$$Loss = -ylog(p) - (1 - y)log(1 - p) \tag{3}$$

where $y \in \{0, 1\}$ indicates whether the prediction entity is a real entity.

Coding fusion module: This module incorporates entity prediction, conversational context, and partially generated results, ensuring that the response generation process considers more comprehensive information. Given query $Q \in R^{d_Q}$, key $K = R^{d_K}$, and value $V \in R^{d_V}$, each attention head H_i is calculated as follows:

$$H_i = F(W_i^Q Q,\ W_i^K K,\ W_i^V V) \in R^{p_V} \tag{4}$$

where F denotes the scaled dot-product attention function.

Learnable parameters include $W_i^Q \in R^{p_Q \times d_Q}$, $W_i^K \in R^{p_K \times d_K}$, and $W_i^V \in R^{p_V \times d_V}$.

Thus, the dialogue context E_{GH} with knowledge, the predicted entity E_{cur}, and the encoding of the right-shift output E_{pre} are fed into the Masked Multi-Head Cross-Attention (MMCA) module, which facilitates the fusion of multiple information sources, and its formulations are given in Equations (5) and (6):

$$output_{GH} = MMCA(E_{pre}, E_{GH}, E_{GH}) \tag{5}$$

$$output_{cur} = MMCA(E_{pre}, E_{cur}, E_{cur}) \tag{6}$$

In the formula, $output_{GH}$ and $output_{cur}$ fuse the encoding of the dialogue context of the prediction entity and the knowledge graph. The previously decoded text is obtained using Equation (7):

$$output_{pre} = MMCA(E_{pre}, E_{pre}, E_{pre}) \tag{7}$$

Finally, the obtained $output_{GH}$, $output_{cur}$, and $output_{pre}$ are averaged as shown in Equation (8):

$$output = w_{GH} * output_{GH} + w_{cur} output_{cur} + w_{pre} output_{pre} \tag{8}$$

where w_{GH}, w_{cur}, and w_{pre} are learnable scalar weights.

5. EXPERIMENTS AND ANALYSIS

The experiments were conducted using an Intel Xeon Gold 6138 CPU (32GB memory) and an RTX 3090 GPU (24GB) with CUDA 11.1. The tasks of Intent Recognition, Dialogue Strategy Learning, Medical Entity Recognition, and Medical Text Generation are compared using classic dialogue classification models, fine-grained dialogue classification models, text sequence annotation models, and dialogue response generation models, respectively.

5.1. Intention Recognition

This section compares between state-of-the-art intention recognition models based on pre-training. It considers the following:

(1) BERT [33] is a self-supervised language model that defines two types of tasks for pre-training. The basic model has 12 layers, 768 hidden layers, 12 heads, and 110 million parameters.

(2) Bert-wwm [34], based on BERT, is a Chinese full-word mask used to conduct the pre-training model.

(3) ALBERT-tiny [35] is a pre-training model based on two pre-training objectives: mask language modelling and sentence ordering prediction.

(4) Roberta-wwm [36] removes the task of predicting the next sentence and dynamically changes the mask mode applied to the training data, combining the advantages of RoBERTa and BERT-wwm.

(5) PCL-MedBERT [37] is a pre-trained language model used in the medical domain.

As shown in Table 5, RoBERTa-wwm achieves the best F1-score (82.09) among all intent recognition baselines.

Table 5. Intent recognition.

Model	Precision	Recall	F1-score
BERT	80.17	81.34	80.75
BERT-wwm	80.94	81.87	81.40
ALBERT-tiny	76.87	78.23	77.54
RoBERTa-wwm	81.84	82.34	82.09
PCL-MedBERT	80.78	81.69	81.23

5.2. Medical Entity Recognition

MER is a basic sequence labelling recognition method, mainly used to identify the categories and boundaries of medical entities in dialogue text. A bidirectional LSTM (Bi-LSTM) can capture bidirectional entity information accurately, combine it with a conditional random field (CRF) to effectively carry out NER. As shown in Table 6, BERT-BiLSTM-CRF obtains the best overall MER performance with an F1-score of 90.50.

Table 6. Medical entity recognition.

Model	Precision	Recall	F1-score
CRF	80.32	81.45	80.88
BiLSTM	87.12	87.65	87.38
BiLSTM-CRF	88.35	88.87	88.61
BERT-CRF	89.56	88.78	89.17
BERT-BiLSTM-CRF	90.34	90.67	90.50

5.3. Dialog Strategy Learning

Given the conversation history H_t and the current intent, the dialogue policy learning task aims to generate the next action A_t, which is represented as a set of slot-value pairs. The main models are as follows:

(1) Bert-WWM [34], based on BERT, adopts the Chinese full-word mask method to conduct a pre-training model.

(2) PCL-MedBERT [37] is a retrained language model in the medical domain.

(3) GPT-2 [38] is a one-way transformer model that uses autoregression.

(4) In contrast to GPT-2, DialoGPT [39] is pre-trained on large-scale dialogues and can capture more fine-grained dialogue flows.

As shown in Table 7, PCL-MedBERT yields the highest Action Precision (50.34), while BERT-wwm achieves the best Action F1 (49.98).

Table 7. Dialog policy learning.

Model	Action_Precision	Action_Recall	Action_F1
BERT-wwm	48.89	51.14	49.98
PCL-MedBERT	50.34	47.96	49.12
GPT-2	38.21	41.56	39.81
DialoGPT	39.65	41.89	40.74

5.4. Medical Text Generation

Main experiment: The main models are as follows:

(1) Seq2Seq [40] is a model with an attention mechanism. An encoder and decoder are used to complete the training task, transforming one input sequence to another.

(2) HRED [41] uses a hierarchical encoder to encode the history of the doctor–patient conversation so that the sentence-level representation of the preceding and subsequent messages can be obtained.

(3) GPT-2 [38] is a one-way transformer model that uses autoregression.

(4) MOSS [42] is a modular supervision network model in which an encoder-decoder training framework is used to identify the medical entities in a response using conversation-state tracking.

(5) MedDGBERT-GPT [6] uses MedDGBERT to encode a history of doctor-patient conversations. MedDGBERT is pre-trained using the MedDG dataset [6], and then GPT is used to generate final responses.

Table 8 presents the performance results of the comparison model (* indicates statistical significance, $p < 0.05$). It shows that the entity-aware medical dialogue model with knowledge integration proposed in this study outperforms other models on BLEU@N, distinct@N, and F1. Here, F1 is the entity-level F1 score computed on medical entities in the generated responses, using exact word matching against the reference entities. The proposed model shows better performance, which verifies its effectiveness.

Table 8. Performance comparison.

Model	BLEU1	BLEU2	BLEU3	BLEU4	Distinct1	Distinct2	F1
Seq2Seq	19.24	13.32	9.54	6.38	0.57	5.14	14.32
HRED	23.35	16.74	12.44	8.57	0.65	6.96	15.14
GPT-2	26.33	19.54	15.37	11.25	0.77	11.10	16.23
MOSS	28.47	21.32	17.23	13.64	0.74	10.92	17.54
MedDGBERT-GPT	30.56	23.78	19.43	15.45	0.53	10.94	18.23
Ours	**31.67***	**24.85***	**20.52***	**16.56***	**0.79***	**11.14***	**20.43***

Note: F1 denotes the entity-level F1 on medical entities in generated responses.

Ablation experiment: To further evaluate the effectiveness of different components in the proposed model, we conduct two ablation settings. w/o KG Embedding represents the model with the knowledge graph embedding module removed, while retaining the entity prediction and fusion modules. w/o Fusion Module represents the model with the coding fusion module removed. As shown in Table 9, both ablation variants perform worse than the full model, demonstrating the effectiveness of each component. Specifically, w/o KG Embedding yields the largest performance drop, indicating that the knowledge graph embedding module provides useful structured medical knowledge and entity relations for response generation. w/o Fusion Module also performs worse than the full model, suggesting that the coding fusion module plays an important role in integrating dialogue context, predicted entities, and knowledge information. Overall, these results verify that both the knowledge graph embedding module and the fusion module contribute substantially to medical dialogue generation.

Table 9. Ablation analysis.

Model	BLEU1	BLEU2	BLEU3	BLEU4	F1
w/o KG Embedding	30.78	23.69	19.47	15.67	18.12
w/o Fusion Module	31.13	24.52	20.32	16.28	19.67
Ours	**31.67***	**24.85***	**20.52***	**16.56***	**20.43***

5.5. Example Analysis

In this study, we developed a WeChat mini program to evaluate medical dialogue generation. The medical dialogue system comprises three modules: the NLU module, which obtains the intention expressed in the patient text and extracts the corresponding medical entity; the knowledge fusion module, a fusion of the current questions, predicted entities, dialogue history text, entity relationship, and role information; and the dialogue generation module, which converts the fused information into understandable text information.

The example of the proposed model presented above shows that the doctor role is well-modelled and that the generated results contain information that is richer and more coherent than the results of prior models.As shown in Figure 6, the deployed WeChat mini program demonstrates the capability of generating coherent and medically relevant responses in an online setting.

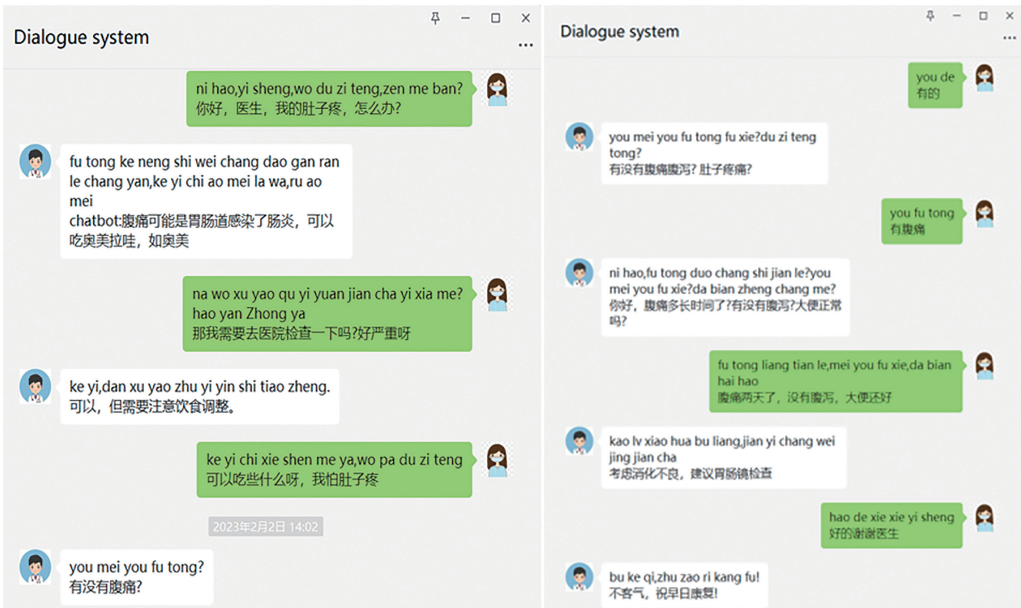


Figure 6. WeChat mini program effect figure.

6. CONCLUSION AND FUTURE WORK

This study provides a multi-task multi-label medical dialogue dataset as an important basis for the construction of a more effective medical consultation system. The current generation of medical dialogue is beset with problems, such as insufficient medical entities in responses and chaotic relationships between medical entities, which reduce the credibility of medical dialogue generation and affect patients' judgment. To solve these problems, this study proposes a medical dialogue text generation model based on knowledge fusion. The study's main findings are as follows: (1) The knowledge representation of medical dialogue datasets in previous studies is incomplete and has problems, such as a single label or a single training task. (2) Relationships between medical entities are disordered in previous studies. This study corrects the relationships between entities by embedding knowledge. (3) The study increases the number of medical entities in responses through the fusion module and entity prediction. (4) This study built a gastroenterology dialogue dataset, tested the performance of various medical conversion tasks, and can conduct a dialogue online.

The model in this paper was trained on a single NVIDIA RTX 3090 GPU (24GB). And actual online deployment was carried out using WeChat mini programs, which can respond to patient needs in real time. It can address the practical needs of gastrointestinal patients, and in the future, we hope to build a single model medical dialogue system for general medical departments, which can reduce the training cost of the model while ensuring generalization accuracy through transfer model. Utilize unsupervised data or manual feedback data for continuous learning to further improve performance.

AUTHOR CONTRIBUTIONS

Yupeng Liu, design, writing, revision; Gang Liu, writing, experiment; Yuhao Zhang, data construction, revision, writing; Jiang Zhang, data construction, revision, writing; Han Zhang, data construction, revision, writing; Yi Qin, submission, revision.

REFERENCES

[1] Q. Chen et al., "Enhancing dialogue symptom diagnosis with global attention and symptom graph," in *Proc. Conf. Empirical Methods Nat. Lang. Process. and 9th Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5032–5041, doi: 10.18653/v1/D19-1508.

[2] X. Shi, H. Hu, W. Che, et al., "Understanding medical conversations with scattered keyword attention and weak supervision from responses," *Proc. AAAI Conf. Artif.* Intell., vol. 34, no. 5, pp. 8838–8845, 2020, doi: 10.1609/aaai.v34i05.6412.

[3] Y. Zhang, Z. Jiang, T. Zhang, et al., "MIE: A medical information extractor towards medical dialogues," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguist.* (ACL), 2020, pp. 6460–6469, doi: 10.18653/v1/2020.acl-main.576.

[4] Z. Wei, Q. Liu, B. Peng, et al., "Task-oriented dialogue system for automatic diagnosis," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguist.* (Short Papers), 2018, pp. 201–207, doi: 10.18653/v1/P18-2033.

[5] W. Yang, G. Zeng, B. Tan, et al., "On the generation of medical dialogues for COVID-19," *arXiv:2005.05442*, 2020. [Online]. Available: https://arxiv.org/abs/2005.05442.

[6] W. Liu, J. Tang, Y. Cheng, W. Li, Y. Zheng, and X. Liang, "MedDG: An entity-centric medical consultation dataset for entity-aware medical dialogue generation," in *Natural Language Processing and Chinese Computing*, LNCS, vol. 13551, Cham, Switzerland: Springer, 2022, pp. 447–459, doi: 10.1007/978-3-031-17120-8_35.

[7] O. Vinyals and Q. Le, "A neural conversational model," arXiv:1506.05869, 2015. [Online]. Available: https://arxiv.org/abs/1506.05869.

[8] Z. Tian, R. Yan, L. Mou, et al., "How to make context more useful? An empirical study on context-aware neural conversational models," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguist.* (Short Papers), 2017, pp. 231–236, doi: 10.18653/v1/P17-2036.

[9] C. Xing, Y. Wu, W. Wu, et al., "Hierarchical recurrent attention network for response generation," *Proc. AAAI Conf.* Artif. Intell., vol. 32, no. 1, 2018, doi: 10.1609/aaai.v32i1.11965.

[10] H. Chen, Z. Ren, J. Tang, et al., "Hierarchical variational memory network for dialogue generation," in *Proc. Int. World Wide Web Conf.* (WWW), 2018, pp. 1653–1662, doi: 10.1145/3178876.3186077.

[11] H. Zhang, Y. Lan, L. Pang, et al., "ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation," *arXiv:1907.05339*, 2019. [Online]. Available: https://arxiv.org/abs/1907.05339.

[12] A. Kumar, O. Irsoy, P. Ondruska, et al., "Ask me anything: Dynamic memory networks for natural language processing," in *Proc. Int. Conf. Mach. Learn.* (ICML), 2016, pp. 1378–1387.

[13] L. Mou, Y. Song, R. Yan, et al., "Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation," *arXiv:1607.00970*, 2016. [Online]. Available: https://arxiv.org/abs/1607.00970.

[14] L. Yao, Y. Zhang, Y. Feng, et al., "Towards implicit content-introducing for generative short-text conversation systems," in *Proc. Conf. Empirical Methods Nat. Lang. Process.* (EMNLP), 2017, pp. 2190–2199, doi: 10.18653/v1/D17-1233.

[15] C. Xing, W. Wu, Y. Wu, et al., "Topic aware neural response generation," *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017, doi: 10.1609/aaai.v31i1.10981.
[16] N. Dziri, E. Kamalloo, K. W. Mathewson, et al., "Augmenting neural response generation with context-aware topical attention," *arXiv:1811.01063*, 2018. [Online]. Available: https://arxiv.org/abs/1811.01063.
[17] Y. Jo, M. Yoder, H. Jang, et al., "Modeling dialogue acts with content word filtering and speaker preferences," in *Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP)*, 2017, pp. 2179–2189, doi: 10.18653/v1/D17-1232.
[18] B. Zhang and X. Zhang, "Hierarchy response learning for neural conversation generation," in *Proc. EMNLP-IJCNLP*, 2019, pp. 1772–1781, doi: 10.18653/v1/D19-1186.
[19] T. H. Wen, D. Vandyke, N. Mrksic, et al., "A network-based end-to-end trainable task-oriented dialogue system," arXiv:1604.04562, 2016. [Online]. Available: https://arxiv.org/abs/1604.04562.
[20] A. Bordes, Y. L. Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog," *arXiv:1605.07683*, 2016. [Online]. Available: https://arxiv.org/abs/1605.07683.
[21] A. Madotto, C. S. Wu, and P. Fung, "Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems," *arXiv:1804.08217*, 2018. [Online]. Available: https://arxiv.org/abs/1804.08217.
[22] M. Eric and C. D. Manning, "Key-value retrieval networks for task-oriented dialogue," *arXiv:1705.05414*, 2017. [Online]. Available: https://arxiv.org/abs/1705.05414.
[23] W. Lei, X. Jin, M. Y. Kan, et al., "Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2018, pp. 1437–1447, doi: 10.18653/v1/P18-1133.
[24] W. Wu, Z. Guo, X. Zhou, et al., "Proactive human-machine conversation with explicit conversation goals," *arXiv:1906.05572*, 2019. [Online]. Available: https://arxiv.org/abs/1906.05572.
[25] E. Dinan, S. Roller, K. Shuster, et al., "Wizard of Wikipedia: Knowledge-powered conversational agents," *arXiv:1811.01241*, 2018. [Online]. Available: https://arxiv.org/abs/1811.01241.
[26] H. Zhou, C. Zheng, K. Huang, et al., "KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation," *arXiv:2004.04100*, 2020. [Online]. Available: https://arxiv.org/abs/2004.04100.
[27] K. Zhou, Y. Zhou, W. X. Zhao, et al., "Towards topic-guided conversational recommender system," *arXiv:2010.04125*, 2020. [Online]. Available: https://arxiv.org/abs/2010.04125.
[28] Q. Qian, M. Huang, H. Zhao, et al., "Assigning personality/profile to a chatting machine for coherent conversation generation," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 4279–4285.
[29] S. Zhang, E. Dinan, J. Urbanek, et al., "Personalizing dialogue agents: I have a dog, do you have pets too?," *arXiv:1801.07243*, 2018. [Online]. Available: https://arxiv.org/abs/1801.07243.
[30] P. E. Mazaré, S. Humeau, M. Raison, et al., "Training millions of personalized dialogue agents," *arXiv:1809.01984*, 2018. [Online]. Available: https://arxiv.org/abs/1809.01984.
[31] B. Kim, J. Ahn, and G. Kim, "Sequential latent knowledge selection for knowledge-grounded dialogue," *arXiv:2002.07510*, 2020. [Online]. Available: https://arxiv.org/abs/2002.07510.
[32] R. Lian, M. Xie, F. Wang, et al., "Learning to select knowledge for response generation in dialog systems," *arXiv:1902.04911*, 2019. [Online]. Available: https://arxiv.org/abs/1902.04911.
[33] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805.

[34] Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang, "Pre-training with whole word masking for Chinese BERT," *IEEE/ACM Trans*. Audio, Speech, Lang. Process., vol. 29, pp. 3504–3514, 2021, doi: 10.1109/TASLP.2021.3124365.
[35] Z. Lan, M. Chen, S. Goodman, et al., "ALBERT: A lite BERT for self-supervised learning of language representations," *arXiv:1909.11942*, 2019. [Online]. Available: https://arxiv.org/abs/1909.11942.
[36] Y. Liu, M. Ott, N. Goyal, et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv:1907.11692*, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692.
[37] T. Liu, B. Qin, M. Liu, R. Xu, B. Tang, and Q. Chen, "PCL-MedBERT: Pre-training model for Chinese medical text processing," 2020. [Online]. Available: https://code.ihub.org.cn/projects/1775.
[38] A. Radford, J. Wu, R. Child, et al., Language models are unsupervised multitask learners. OpenAI, 2019.
[39] Y. Zhang, S. Sun, M. Galley, et al., "DialoGPT: Large-scale generative pre-training for conversational response generation," *arXiv:1911.00536*, 2019. [Online]. Available: https://arxiv.org/abs/1911.00536.
[40] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *arXiv:1409.3215*, 2014. [Online]. Available: https://arxiv.org/abs/1409.3215.
[41] I. Serban, A. Sordoni, Y. Bengio, et al., "Building end-to-end dialogue systems using generative hierarchical neural network models," *Proc. AAAI Conf. Artif. Intell.*, vol. 30, no. 1, 2016.
[42] W. Liang, Y. Tian, C. Chen, et al., "MOSS: End-to-end dialog system framework with modular supervision," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, pp. 8327–8335, 2020, doi: 10.1609/aaai.v34i05.6349.

AUTHOR BIOGRAPHY

Yupeng Liu received the Ph.D. degree from the Harbin Institute of Technology. He is currently a Professor in Harbin University of Science and Technology. His research interests include natural language processing and medical intelligence.

Gang Liu is the graduate degree in Harbin University of Science and Technology. His research interests include natural language processing and medical intelligence.

Yuhao Zhang is the graduate degree in Harbin University of Science and Technology. His research interests include natural language processing and medical intelligence.

Jiang Zhang is the graduate degree in Harbin University of Science and Technology. His research interests include natural language processing and medical intelligence.

Han Zhang is the graduate degree in Harbin University of Science and Technology. His research interests include natural language processing and medical intelligence.

Yi Qin is the graduate degree in Harbin University of Science and Technology. His research interests include natural language processing and medical intelligence.