Research on Medical Dialogue Generation of External Knowledge

Na Liu
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 690879168@qq.com

Feng Huang
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 851072437@qq.com

Xiaohui Su
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
E-mail: 287099144@qq.com

***Abstract*—Nowadays, the technology of medical dialogue generation has gradually attracted the attention of more researchers, and the demand for landing has gradually increased. Therefore, building a medical dialogue system that can automatically reply is conducive to improving the efficiency of clinical consultation and reducing the burden on doctors. This paper uses the method of fusing external knowledge to build a dialogue generation model, which greatly enhance the accuracy of the model and ameliorates the disadvantages of classical construction methods. Based on the large-scale pre-training model method, the doctor's response is generated by two-stage training, and the knowledge related to the medical background is added to generate the response that best fits the current context. In this paper, experiments were performed on the medical dialogue dataset KaMed and COVID-19, and the experimental data showed that compared with the traditional human-computer dialogue generation Seq2Seq model, the Perplexity value of this method decreased 1.91, compared with the VHRED model, B@1 value increased 0.3, and the B@2 value increased0.34, D@2 increased 2.14, it can be proved that the medical dialogue model proposed in this paper can provide doctors with response responses more effectively and enhance the accuracy of responses.**

Keywords-Medical Dialogue System; Knowledge Fusion Dialogue; Task-Oriented Dialogue Generation

I. INTRODUCTION

With the COVID-19 epidemic in 2019, an increasing number of businesses have launched a "home office" model, and online consultation has also become a new trend in the medical industry of various countries. In order to solve the problem of surge in user consulting demand and the shortage of medical personnel who provide online diagnosis and consultation services, it is imminent to develop a smart medical dialogue system that can provide users with online diagnosis and treatment solutions. These "online doctors" can greatly reduce the burden on human doctors, reduce the chance of cross -infection among patients, and improve the efficiency of medical resource operations.

The current task -type dialogue generation method can be divided into two types: traditional methods and deep learning methods. The advantage of using the traditional method to build a dialogue system is that it is simple, but only the answers to similar responses in the template can only be found, and high -skilled talent teams provide different templates in different fields. To other areas[2]. In deep learning methods, the method of decoder -based method creates a precedent for using neural network methods in task -type dialogue. At the same time, language models and neural network methods are used. The generated sentences have diversity and reduced artificial characteristics.

With the proposal and development of several large scale pre-training structures such as

DOI: 10.2478/ijanmc-2023-0063

Transformer and GPT, various variants in the field of task -type dialogue generation have also received widespread attention. This model can improve the ability of model modeling language through the pre -training mechanism to improve the model modeling language. , Cross -current -order attention mechanism can better process the structured MR (semantic information) input to achieve the best results at present. The filter mechanism is eliminated with medical knowledge with low historical compatibility with doctors and patients, enhance the selection of accurate knowledge in the reply, thereby improving the accuracy of the model, making it a more accurate and higher medical reference value.

II. RELATED WORK

In the task of establishing a task -oriented dialogue system, dialogue generation is an indispensable part of it. Many scholars are committed to studying various methods for dialogue generation. In 2015, we will be based on the decoder. The method is applied to the dialogue generation task, and a statistical language generator based on a combined cycle and convolutional neural network structure can be proposed. This structure can be trained on dialogue behavior and discourse. Essence Then in 2015, Wen et al. [4] referred to the research in the selection field, and proposed an Encoder-Decoder architecture that adapt to NLG based on a attention mechanism, which achieved a better effect than decoder-based methods.

In 2016, DUšEk et al. [5] explored the advantages and disadvantages of the two methods to generate sentence planning trees and directly generate natural languages with the SEQ2SEQ method. In the same year, Dukek et al. [6] introduced information about language models from the previous model. In the case of maintaining the overall framework of the model, they improved the input and spliced the user's words in front of the input MR (semantic information) three yuan group, as the front of the front; and the newly added one above Coder, coding user discourse alone.

In 2017, Van-Khanh Tran et al. [7] It proposed the encoder-polymer-decoder model based on the extension of the encoder-decoder architecture of recursive neural network.

In 2018, Wei Z, LiU Q et al. [8] proposed a dialogue system for automatic diagnosis, constructed the Medical DS dataset, and proposed a framework of the dialogue system based on strengthening learning (RL). So as to improve the accuracy of automatic diagnosis.

In 2019, SHANG-Yu Su et al.[9] proposed a new type of language understanding and generating learning framework based on dual supervision and learning, providing a method that uses puppets. Experiments show that this method improves significantly improvement can be significantly improved. Language understanding and generating learning performance.

In 2020, PENG et al. [10] proposed the SC-GPT model to enter the MR tank into the GPT pre-training model, and directly obtain the results using the sequence to sequence method.

In 2021, Yangming Li et al. [12] proposed a new heterogeneous rendering (HRM) framework that explains how the nerve generator render the input dialogue (DA) as discourse. For each generation step, the mode switch is concentrated from the renderer to select the appropriate decoder to generate items (words or phrases). This model can well explain the rendering process of the nerve generator.

On the basis of predecessors' research and deep learning development, although the natural language generation (NLG) mission in the dialogue system has achieved good performance, it is still found that if only dialogue data is used to train the model during the dialogue generation process, the model is found to train the model for training. It is easy to produce such security responses such as "good" and "can". This type of reply has no reference value to the therapy of physician. Thus, to solve these challenges, this article adds and has added and added to predecessors' research. The symptoms described by patients have high-fitting medical knowledge, followed by effectively incorporating knowledge and tracking the patient's speech, testing the doctor's behavior entity, and eventually generated a reply through professional medical knowledge,

patient's condition and doctor behavior to help Patients provide diagnosis and treatment opinions and suggestions in effective ways.

III. RESEARCH METHODS

A. Task definition

Medical dialogue generation tasks are to help doctors provide diagnosis and treatment suggestions in the case of consultation. This needs to be based on the history of the doctor and patient dialogue and related medical knowledge, to generate the same text reply to the user's and conform to the context and meet the medical principles. This article uses the dialogue history $H=\{\{h_1,r_1\},\{h_2,r_2\},...,\{h_i,r_i\},...,\{h_n,r_n\}\}$ and medical knowledge $K=\{k_1,k_2,...,k_i,...,k_n\}$ between patients and doctors. Among them, $\{h_i,r_i\}$ the history of the dialogue between doctors and patients, k_i is a collection of medical knowledge of the relevant dialogue history Each of these knowledge k_i contains head entities h , relationships r , and tail entities t that have some relationship with head entities. In the form of ternary groups, the model needs to generate a reply Y according to the history of dialogue H and related medical knowledge K Among them, $Y=\{y_1,y_2,...,y_i,...,y_n\}$, y_i it is the second in the reply. The goal of this article is to maximize the genetic reply Y , so that the response is accurate and referenceable.

B. Model overview

In this paper, a medical dialogue generation model that integrates external knowledge is proposed, as shown in Figure1. The model contains: doctor-patient dialogue coding, patient status tracking, doctor behavior detection, medical knowledge distillation, and reply generation. The model needs to obtain a vector with the context information of the doctor-patient dialogue by coding the doctor-patient dialogue to track the patient's condition status during the dialogue, and then predict the doctor's next round of actions, and finally fuse the medical knowledge extracted by the medical knowledge distillation module to produce relevant replies.

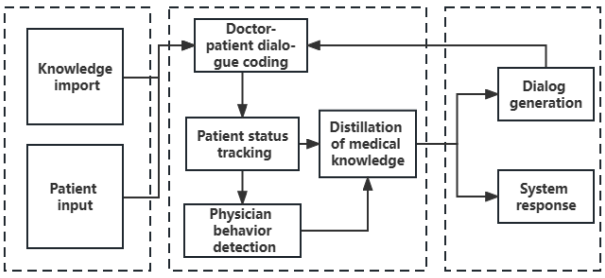

Figure 1. Structure diagram of knowledge-based medical dialogue generation

C. Doctor-patient dialogue coding module

For the doctor-patient dialogue coding module, this paper first uses the bidirectional gate circulation unit (BiGRU) to encode the doctor-patient dialogue history and convert it into a vector expression $\vec{H}=\{\{\vec{h}_1,\vec{r}_1\},\{\vec{h}_2,\vec{r}_2\},...,\{\vec{h}_i,\vec{r}_i\},...,\{\vec{h}_n,\vec{r}_n\}\}$ with doctor-patient context information.

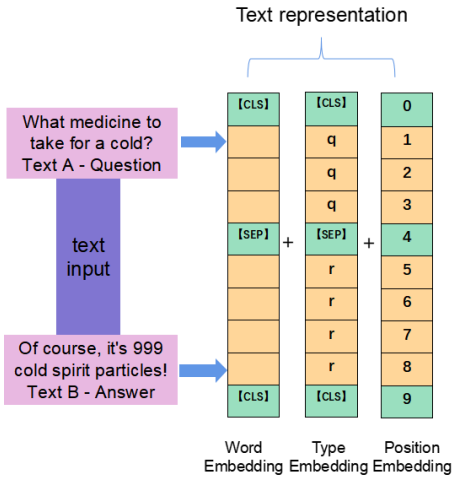

Figure 2. Input representation of doctor-patient dialogue coding module

The BiGRU neural network can greatly reduce the amount of calculation and complexity of the model. The input representation of its doctor-patient dialogue coding module is shown in Figure2.

After using BiGRU to encode the doctor-patient dialogue, a vector with the context information of the doctor-patient dialogue is obtained, where H is the doctor-patient dialogue history. Formally, the output of the encoder is shown in Equation (1):

$$E_h = BiGRU(H) \tag{1}$$

D. Patient status tracking module

The goal of the patient status tracking module is mainly to track the patient's condition status in the doctor-patient dialogue history, use tracking to the patient's state S to predict the doctor's next round of behavior B , and finally produce a response Y that is in line with the context of the doctor-patient dialogue. Therefore, a state tracker composed of variational autoencoder is added to the patient status tracking module to indirectly learn the posterior distribution $p(y \mid x)$ of the data to obtain the probability distribution of the patient state.

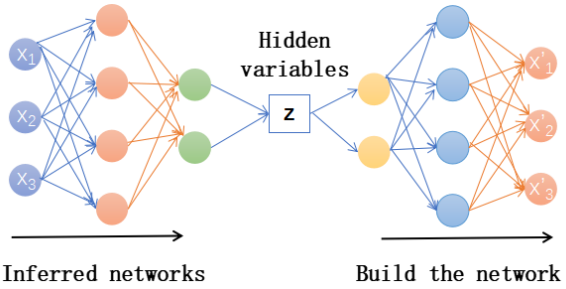

Figure 3. (Variational autoencoder) Generates a model structure diagram

The learning process first learns the joint probability distribution $p(x, y)$ from the dialogue data, and then reverses the posterior distribution by solving the prior probability and categorical conditional probability of the patient's state, and the posterior probability obtained at this time is the probability of the patient's state tracking distribution at time t as sought in this module. The structure of the model is shown in Figure3.

In each round of medical conversation that tracks the patient's status, the status tracker first samples a previous round of patient status s_{t-1}^{q} from the approximate posterior distribution $q_{\phi_s}(s_{t-1})$ of the state, and then enters the previous round of patient status s_{t-1}^{q} into BiGRU to obtain the current round of patient status s_t^q , that is, the prior patient status s_t^q of the T round is combined from the patient status s_{t-1}^{q} of the T-1 round, the doctor's response R_{t-1} , and the patient description U_t of the T round. The formula is shown in Equation (2).

$$p_{\theta_s}(S_t) = p_{\theta_s}(S_t \mid S_{t-1}, R_{t-1}, U_t) \tag{2}$$

Specifically, after obtaining the prior patient state $p_{\theta_s}(S_t)$ at time t, in the decoding stage, the model takes the doctor's behavior state $b_{t,0}^{S^p} = W_s^p[h_t^c; h_{t-1}^{S^q}]$ predicted at the very beginning moment (i.e., 0 moment) as the initial hidden state of the decoder (W_s^p in which a learnable parameter, $[\cdot;\cdot]$ is represented by a cascade operation of the vector), at the decoding moment of the ith word, the decoder decodes serially S_t , receives the encoding vector $e_{t,i-1}^{S^p}$ of the word of the previous moment as input, outputs the doctor's behavior state $b_{t,i}^{S^p}$ at the i moment, and maps it to the patient's state $b_{t,i}^{S^p}$ space. This article fixes the length of the patient state S_t to $|S|$, as shown in Equation (3).

$$p_{\theta_s}(S_t) = \prod_{i=1}^{|S|} soft\max(MLP(b_{t,i}^{S^p})) \tag{3}$$

Where MLP represents a multilayer perceptron (MLP), the state tracker is inferred to be similar to the above process, but additionally combines a vector representation R_t (i.e.,) as input. The BiGRU decoder is used $b_{t,0}^{S^q} = W_s^q[h_t^c; h_{t-1}^{S^q}; h_t^R]$ to initialize, where W_s^q is a learnable parameter, and at the ith decoding moment, output the predicted doctor's actions $b_{t,i}^{S^q}$ in the t round. Among them, the approximate posterior distribution is formulated as shown in Equation (4).

$$q_{\phi_s}(S_t) = \prod_{i=1}^{|S|} soft\max(MLP(b_{t,i}^{S^q})) \tag{4}$$

E. Doctor behavior detection module

The doctor behavior detection module is designed to predict the entity that the doctor will reply to in the next conversation, this paper builds a doctor behavior classifier in this module, which mainly divides the doctor behavior into four types

such as asking symptoms, diagnosis, prescribing drugs, small talk, etc., doctors should judge whether there is a shift between behavioral states according to the conversation history, that is, to determine whether the next sentence should continue to ask the patient's symptoms or give diagnostic suggestions, etc., the doctor behavior detection process is shown in Figure 4.

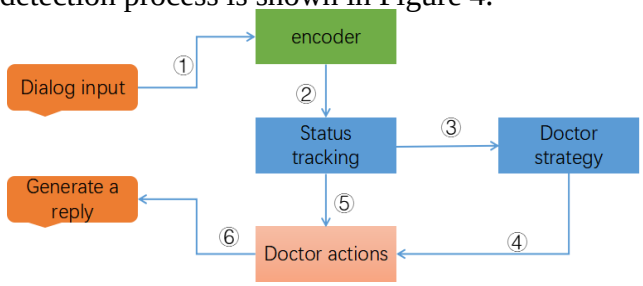

Figure 4. Flow chart of doctor behavior detection

Specifically, in this module, similar to the Patient Status Tracking Module, this paper models a priori distribution and an approximate posterior distribution of physician actions, and its prior strategy network and posterior strategy network are also based on encoder-decoder structures, and define the representation of physician behavior A_t at moment t as a physician action category A_t^c and a series of explicit keywords A_t^k, $A_t = \{A_t^c, A_t^k\}$, where A_t^k the length is fixed $|A|$.

For a priori strategy network, in the encoding stage, this paper first uses a BiGRU encoder to encode the patient state S_t^p at time t, and outputs a hidden vector $h_t^{S^p}$. In the decoding and reply stage, an action classifier A_t^c is designed to infer the doctor's action category, and the hidden layer vector h_t^c at time t is used to perform attention operations q_t on the obtained doctor's behavior, and an attention vector q_t about the doctor's behavior is calculated. Next, the action classifier combines the obtained attention vector to divide the physician's actions into four categories, namely, asking about symptoms, diagnosis, prescribing medication, and gossiping, and the doctor's prior behavior is formulaically described as shown in Equation (5).

$$P_{\theta_{a,c}}(A_t^c) = soft\max(W_c^p[h_t^{S^p}; h_t^c; q_t]) \quad (5)$$

Learners are described W_c^p, and the model samples the action categories $A_t^{c,p}$ from the prior behavior $P_{\theta_{a,c}}(A_t^c)$ of the doctor.

A_t^k are serialized by a BiGRU decoder, and at each decoding step, the context inference detector maps the output vector of the BiGRU decoder into the action space. The decoder is initialized by a doctor's behavior $b_{t,0}^{A^{k,p}} = W_k^p[h_t^{S^p}; h_t^c; e_t^{A^{c,p}}]$ at 0 moment, where $e_t^{A^{c,p}}$ is the embedded representation $A_t^{c,p}$. At the ith decoding step, the decoder outputs $b_{t,i}^{A^{k,p}}$. The contextual inference detector is based on the doctor's behavior output vector $b_{t,i}^{A^{k,p}}$ at i-moment, co-infers the action vector $A_{t,i}^k$ of the doctor at i-moment, learns from the original doctor-patient dialogue history and the detected patient state, and infers its prior distribution $A_{t,i}^k$ as shown in Equation (6) through a multilayer perceptron MLP.

$$p_{\theta_{a,d}}(A_{t,i}^k) = \frac{1}{z_A}\exp(MLP([h_t^{S^p}; h_t^c; e_t^{A^{c,p}}; b_{t,i}^{A^{k,p}}])) \quad (6)$$

Finally, the prior distribution of the doctor's behavior A_t is calculated as shown in Equation (7).

$$p_{\theta_a}(A_t) = p_{\theta_{a,c}} \cdot \prod_{i=1}^{|A|}[p_{\theta_{a,d}}(A_{t,i}^k) + p_{\theta_{a,g}}(A_{t,i}^k)] \quad (7)$$

The inference strategy network approximates the posterior distribution of action categories by extracting key information with directivity from previous doctors' responses R_t. Use a BiGRU encoder R_t to encode as a hidden vector h_t^R and a hidden vector S_t^q respectively $h_t^{S^q}$. Then obtain an approximate posterior distribution of the physician's action class at time t, as shown in Equation (8).

$$q_{\theta_{a,c}}(A_t^c) = soft\max(W_c^q[h_t^c; h_t^{S^q}; h_t^R]) \quad (8)$$

Finally, the model extracts the doctor's action vector to the T moment by approximating the posterior distribution $q_{\theta_{a,c}}(A_t^c)$ of the doctor's action category $A_t^{c,q}$ at time t. In order to enhance the influence of the information in the doctor's reply R_t at moment t, this paper inference policy network designs a posterior distribution A_t^k of contextual inference detector de-approximation. The decoder is represented by $b_{t,0}^{A^{k,q}} = W_k^q[h_t^c; h_t^{S^q}; e_t^{c,q}; h_t^R]$ initialization, where the word embed din e representation representing the doctor's behavior A. W_k^q is a matrix of learnable parameters. In the i-th decoding step, the decoder outputs a vector representation $b_{t,i}^{A^{k,p}}$, so the approximate posterior distribution of the i-th action keyword can be obtained in this paper, as shown in Equation (9).

$$q_{\phi_{a,d}}(A_{t,i}^k) = soft\max(MLP([h_t^c; h_t^{s^q}; e_t^{A^{c,q}}; b_{t,i}^{A^{k,q}}])) \quad (9)$$

Finally, the approximate posterior distribution of the doctor's behavior detection at the t moment is obtained by the doctor action category and the ith action keyword of the t moment, and the formula is shown in Equation (10).

$$q_{\phi_a}(A_t) = q_{\phi_{a,c}}(A_t^c) \cdot \prod_{i=1}^{|A|} q_{\phi_{a,d}}(A_{t,i}^k) \quad (10)$$

F. Medical Knowledge Distillation Module

The most important thing in the dialogue system that integrates knowledge is to screen the appropriate knowledge as the input of the dialogue system, but due to the large scale of external knowledge and many data types, all possible knowledge as input may lead to more noise and high computational intensity, so this paper adds a noise filtering mechanism in the medical knowledge distillation module to select better knowledge and generate a more accurate response.

Specifically, the medical knowledge distillation module of this paper first performs a global pooling operation on the encoded vector with dialogue history context information and the correct medical background knowledge vector information obtained to obtain the corresponding scalar representation sum $\{S_{k_1}, S_{k_2}, ..., S_{k_n}\}$, which uses the doctor-patient dialogue context vector and each dimension of information with medical background knowledge vector, so that the scalar obtained can represent the semantic information represented by the vector to a certain extent. Moreover, after the vector is converted to the corresponding scalar, the complexity of the feature representation is greatly reduced, and the complexity of the model calculation is greatly reduced.

Secondly, in order to allow simultaneous interaction between knowledge and knowledge and between knowledge and the historical context of doctor-patient dialogue, this paper stitches each piece of knowledge and the historical context of doctor-patient dialogue and performs a two-layer full connection operation to obtain the weight $r = \{r_1, r_2, ..., r_{N+1}\}$ assigned to each piece of knowledge and the historical context of doctor-patient dialogue, and the result of summing the weights of all knowledge is used k' as the representation of the selected knowledge. After global pooling and Layer 2 full connection operation, the model can learn more interactive information.

Where the weights can be expressed as shown in Equation (11):

$$r = \sigma(W_2\delta(W_1(S_{k_1}S_{k_2}, ..., S_{k_N}S_x)) \quad (11)$$

Where W_1 and W_2 is the learnable parameter, δ is the ReLU activation function, σ is the sigmoid activation function.

The calculation method of k' is shown in (12) (13).

$$k_i^{'} = r_i \cdot k_i (i = 1, 2, ..., N) \quad (12)$$

$$k' = \sum_{i=1}^{N} k_i^{'} \tag{13}$$

Where $k_i^{'}$ is the weight corresponding to each piece of knowledge, and the output k' of the knowledge distillation module is and will be used to participate in the generation of replies in the reply generation module.

G. Reply to the build module

The reply generation module is composed of the patient status obtained by the patient status tracking module, the next doctor behavior predicted by the doctor behavior detection module, and the knowledge that is highly consistent with the doctor-patient dialogue context screened out by the knowledge distillation module, and jointly generate a response through the reply generation module to give doctors a reference to the diagnosis and treatment response.

In the first stage of reply generation, the model uses a BiGRU encoder to encode the patient state S_t^q at t moment as an embeddable vector $\vec{S}_t^q$, followed by a word-level encoding matrix of the patient state S_t^q , in which $\vec{S}_t^q$ each row is represented as an embedding vector corresponding to a word. Similarly, physician behavior $A_t^{k,q}$ is encoded as an embeddable vector $\vec{A}_t^{k,q}$. Then the hidden vectors of the sum are calculated separately S_t^q and $A_t^{k,q}$ expressed as sum, and the reply decoder is based on a BiGRU unit in the decoding stage, $b_{t,0}^R = W_d[h_t^c; h_t^{S^q}; e_t^{A^{c,q}}; h_t^{A^{k,q}}]$ is initialized, where there W_d is a parameter matrix. In the i-th decoding step, the output $b_{t,i-1}^R$ of the decoder obtains a vector $b_{t,i}^h$ for the context representation H_t of the attention operation. At the same time, attention manipulation is performed for the patient's state S_t^q at time t and the doctor's behavior $A_t^{k,q}$ at moment t, and the hidden vector sum after attention operation is obtained, respectively. The relevant hidden vector and embedding vector $[b_{t,i}^h; b_{t,i}^s; b_{t,i}^a; e_{t,i-1}^R]$ are then fed into the BiGRU unit of the decoder and output the i word $b_{t,i}^R$ in the i-moment, which $e_{t,i-1}^R$ is demonstrated. The probability $R_{t,i}$ of the model generating a reply is the sum of the generation probability and the probability of copying knowledge, which is formulated as shown in equation (14)(15)and(16).

$$p_{\theta_g}(R_{t,i}) = p_{\theta g}^g(R_{t,i}) + p_{\theta g}^c(R_{t,i}) \tag{14}$$

$$p_{\theta_g}^g(R_{t,i}) = \frac{1}{z_R}\exp(MLP(b_{t,i}^R)) \tag{15}$$

$$p_{\theta_g}^c(R_{t,i}) = \frac{1}{z_R}\sum_{j:W_j=R_{t,i}} \exp(h_j^{W^T} \cdot b_{t,i}^R) \tag{16}$$

The generation probability is represented $p_{\theta_g}^g(R_{t,i})$ and the probability of copying knowledge is represented by $p_{\theta_g}^c(R_{t,i})$. z_R is a regular item shared with $p_{\theta_g}^c(R_{t,i})$. This article will write the sequence of R_{t-1} , U_t , $A_t^{k,q}$ and S_t^q to concatenate W , W_j to represent the j word in W .

IV. Experimental results and analysis

A. Datasets

To validate the essentiality of the essay method, this paper uses large-scale medical datasets KaMed and a large-scale COVID-19 Chinese dialogue dataset proposed by Zeng et al. [10] in 2020. Among them, COVID-19 contains more than 1,088 conversations, covering common COVID questions and answers, and providing granular entity-level annotations; KaMed includes more departments and a wider variety of diseases, including more than 17K medical conversations and 5682 entities. The dataset used in this article is shown in Table1.

TABLE I. Dataset statistics

Datasets	Dialog	Utterances	Tokens	Knowledge
KaMed	17864	153,000	6,663,272	Y
COVID-19	1088	9494	406,550	N

B. Experiment setup

The experimental hardware environment in this article uses Intel I9-10900KCPU, Nvidia Geforce 2080Ti*2, 64G RAM, and SSD512G. The software environment uses the Windows 10 system, and at the same time uses Pycharm, Visual Studio Code, Pytorch, Neo4j, etc. for development, and uses TensorBoard for dialogue display.

C. Evaluation indicators

In order to evaluate the linguistic quality of the responses generated in this paper, the metrics BLEU@N, Distinct@N, and Perplexity confusion level (PPL) are used to evaluate the model proposed in this paper.

D. Experimental analysis

To validate whether the model proposed in this paper is better than the previous model, this article uses the Seq2Seq end-to-end model with attention mechanism, the HRED model and the model in this paper are experimented with KaMed, COVID-19 two datasets, first of all, in the experimental process, and the confusion matrix is used to summarize the generated results. The visual attention score obtained by different models obtained by the experiment can be seen: different models have different maximum attention scores for the context, the basic Seq2Seq model, for "cold" and "medicine" such entities are not high, and HRED and VHRED compared with the Seq2Seq model, the attention score of such entities has increased a certain accuracy, and finally the model that integrates knowledge in this paper has the most accurate attention score for the conversation context, which is for "cold" and "taking medicine" The high attention scores of such entities indicate that the method presented in this paper can track and speculate on medical entities. The results of medical dialogue generation under different datasets of different models are shown in Table 2 below.

TABLE II. EXPERIMENTAL RESULTS

Datasets	Model	B@1	B@2	D@1	D@2	Perplexity
KaMed	Seq2Seq	2.71	1.58	1.24	6.85	24.82
	HRED	2.59	1.59	1.17	6.65	27.14
	VHRED	2.49	1.55	1.15	6.42	28.65
	Ours	2.79	1.89	1.58	8.56	22.91
COVID-19	Seq2Seq	3.13	5.70	5.5	29.0	53.3
	HRED	2.56	5.73	5.21	32.39	49.6
	VHRED	3.31	5.65	5.65	34.56	47.2
	Ours	3.57	5.90	5.89	31.21	40.8

Experiments show that the proposed method performs well on two different datasets. Compared with the baseline Seq2Seq method, the Perplexity value of the proposed method is reduced by 1.91 compared with the baseline Seq2Seq method, and the smaller the value, the higher the accuracy of the proposed method. Compared with the VHRED model, the B@1 value of this method increased by 0.3, the B@2 value increased by 0.34, and the D@2 value increased by 2.14, all of which indicate that this method is better and still the same on the dataset COVID-19. Figure5 shows the following example of the conversation under test of the model in this document. Experiments show that the proposed method.

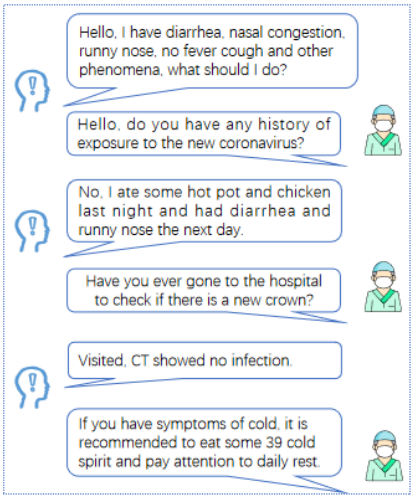

Figure 5. Example of model generation

Model generation sample For example, when the patient proposes diarrhea, nasal congestion, runny nose and other symptoms and asks what to do, the doctor needs to look for knowledge related to the patient's disease in the knowledge triad through the patient's description of the disease, see which conditions the patient's symptoms are related to, further inquire about the patient's condition, and finally give a reply that the doctor can refer to.

V. Conclusion and Outlook

In order to solve the problem of surging demand for user consultation and shortage of medical personnel providing online diagnosis and treatment consulting services in the field of intelligent healthcare, this paper constructs a knowledge-based medical dialogue generation model, although the traditional template-based method is simple, but it requires high-skilled talents to write different templates in different fields, the labor maintenance cost is high, and the dialogue system built by slot filling patients can only find the existing answer set, can not be expanded, and the user utilization efficiency is low.

Therefore, this paper proposes a medical conversation generation model that adds external knowledge, and under the premise of using pre-trained models to enhance the scalability and portability of the proposed model, this paper pioneeringly combines external knowledge to improve the medical response generation task by tracking the patient's status and detecting the doctor's behavior. A large number of experiments on KaMed and COVID-19 show that the medical dialogue generation model based on this paper is superior to most baseline models in BLEU value, Discrete value and Perplexity value, which verifies the effectiveness of this model.

In the next study, it is planned to integrate patients' emotional variables into the dialogue generation model to enable the model to better understand the sentences expressed by the patient's spoken expression, hoping to generate more personalized and targeted responses and further improve the performance of the model in generating responses.

References

[1] QIN Libo, LI Zhouyang, LOU Jieming, YU Qiying, CHI Wanxiang. Review of research progress on natural language generation in task-based dialogue systems [J]. Journal of Chinese Information Technology, 2022.

[2] ZHANG Xiaoyu, LI Dongdong, REN Pengjie, CHEN Zhumin, MA Jun, REN Zhaochun. Knowledge-aware medical dialogue generation based on memory network [J]. Computer Research and Development, 2022.

[3] Wen T H, Gasic M, Kim D, et al. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking [J]. 2015.

[4] Wen T H, Gasic M, Mrksic N, et al. Semantically conditioned lstm-based natural language generation for spoken dialogue systems [J]. 1508.01745, 2015.

[5] Dušek O, Jurčíček F. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings [J]. 2016.

[6] Dušek O, Jurčíček F. A context-aware natural language generator for dialogue systems [J]. 2016.

[7] Tran V K, Nguyen L M. Neural-based natural language generation in dialogue using rnn encoder-decoder with semantic aggregation [J]. 2017.

[8] Wei Z, Liu Q, Peng B, et al. Task-oriented dialogue system for automatic diagnosis[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2018: 201-207.

[9] Su S Y, Huang C W, Chen Y N. Dual supervised learning for natural language understanding and generation [J]. 2019.

[10] Peng B, Zhu C, Li C, et al. Few-shot natural language generation for task-oriented dialog [J]. 2020.

[11] Li Y, Yao K. Interpretable nlg for task-oriented dialogue systems with heterogeneous rendering machines[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15): 13306-13314.

[12] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks [J]. Advances in neural information processing systems, 2014, 27.

[13] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [J]. 2018.

[14] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [J]. 2019, 1(8).