ISSN: 2414-1895 DOI: 10.6919/ICJE.202309_9(9).0020

Research on Medical Dialogue Generation based on Pre-trained Models

Na Liu[a], Xiaohui Su[b], and Feng Huang[c]

School of Computer Science and Engineering, Xi'an Technological University, Xi'an 710021, China

[a]18391334836@163.com, [b]287099144@qq.com, [c]851072437@qq.com

Abstract

China is gradually entering an aging society, the online medical industry is becoming more and more important, and the requirements for intelligent online medical dialogue are getting higher and higher. The traditional dialogue system built using template-based or sentence-based planning methods is simple, but it requires high-skilled talents to provide different templates in different fields, which requires a lot of material, human and financial resources, poor portability and scalability, and is difficult to transfer to other fields. Based on the previous research on large-scale pre-training models to build a medical dialogue generation system, this thesis builds an intelligent dialogue system based on the Seq2Seq method, adds a fusion knowledge module, and adds a noise filter mechanism in the decoding process to screen out medical knowledge that does not match well with the history of doctor-patient dialogue. Finally, experiments are carried out on the large-scale medical dialogue dataset KaMed and COVID-19, and the experimental data show that compared with the traditional human-computer dialogue generation Seq2Seq model, the Perplexity value of the proposed method is reduced by 12.7%, and the B@2 value is increased by 5.72%, which greatly improves the accuracy of the model, makes its response more accurate and has a higher medical reference value.

Keywords

Medical Conversation Generation; Pre-trained Models; Task-oriented Conversation Generation; Medical Knowledge.

1. Introduction

With the improvement of people's living standards, health has gained more and more attention from the public, and timely and accurate medical diagnosis has become the goal pursued by everyone. However, the Chinese population base is huge, the distribution of medical resources is extremely uneven, and the most serious problem at this stage is that there is a huge gap in the number of doctors. Therefore, in recent years, with the development of the Internet, a group of online medical dialogue systems similar to Chunyu Doctor, Lilac Doctor, and Good Doctor have emerged, which break the limitations of time, space and geography, so that patients can enjoy high-level and high-quality services from superior medical experts without leaving home, which not only helps patients avoid the rush to and from the hospital, reduces the cost of medical treatment for patients, but also wins valuable time for patients and greatly reduces the burden of human doctors.

Current task-based dialogue generation methods can be classified into two categories [2]: traditional methods and deep learning methods. Among them, traditional methods include (1) template-based methods, (2) sentence planning-based methods, (3) class-based methods, and (4) phrase-based methods. Deep learning methods include (1)decoder-based methods, (2) sequence-to-sequence-based methods, and (3) pre-trained model-based methods.

The advantages of using traditional methods to build dialogue systems are their simplicity, efficiency, and the accuracy, ease of control, and high quality of the generated responses, but their poor portability, the need to write different templates for different domains, and the high workload of manual writing and maintenance. Among the deep learning approaches, the decoder-based approach pioneered the use of neural network methods in task-based dialogue generation, using both language models and neural network methods to generate utterances with diversity and reduced manual feature intervention. However, its drawback is that one-hot encoding using semantic control contains less information, the design is simpler, and the criteria for generating utterances are too simple to get better results sometimes. Since then, as models of the Seq2Seq class have achieved better results in fields such as machine translation, there has been some work in recent years to explore the application of Seq2Seq in task-based conversational natural language generation. Its model is able to encode MR (semantic information) input better through the encoder-decoder architecture of attention mechanism, and it obtains better results by adding the semantic information contained in the input to the decoding stage. However, it still has drawbacks, mainly that for MR input, the method only treats it as a sequence for encoding and ignores the structured information in it. the autoregressive modeling paradigm used by Encoder-decoder cannot be processed in parallel, resulting in a longer time for model training.

With the proposal and development of several large-scale pre-training structures such as Transformer and GPT, its various variants in the field of task-based dialogue generation have also received widespread attention. The model can improve the model's ability to model language through pre-training mechanism, and the attention mechanism across temporal sequences can better process structured MR (semantic information) input, achieving the best results so far. Therefore, in this paper, based on the previous research on large-scale pre-training model to build a medical dialogue generation system, we add a fused knowledge module, add a noise filter mechanism in the process of decoding to eliminate medical knowledge that does not match well with the history of doctor-patient dialogue, enhance the selection of accurate knowledge in the responses, and then improve the accuracy of the model so that its responses are more accurate and have higher medical reference values.

2. Related Work

Online consultation and consultation makes the two parties no longer restricted by geography, but the traditional pipeline approach [2] can only build dialogue systems for specific domains, which limits the flexibility of the model, especially when large-scale labeled data are often not available due to privacy-protected medical industry data not being publicly available, and it is more difficult to study medical dialogue generation using traditional methods. Therefore, many scholars at home and abroad have devoted themselves to studying various approaches to dialogue generation, and in 2015 Wen et al [3] first applied decoder-based methods to dialogue generation tasks, proposing a statistical language generator based on a joint recurrent and convolutional neural network structure that can be trained on dialogue behaviors, discourse pairs, without any semantic alignment or predefined syntax trees. Then in 2015 Wen et al [4] proposed an Encoder-Decoder architecture based on an attention mechanism adapted to NLG with reference to research in the field of selection, achieving better results than the decoder-based approach.

In 2016 Dušek et al [5] explored the advantages and disadvantages of both approaches, generating sentence planning trees with the Seq2Seq method and generating natural language responses directly. In the same year Dušek et al [6] introduced information about language models from their previous model. While keeping the general framework of the model unchanged, they improved the input by splicing the user's words in front of the MR (semantic information) triad of the input as the antecedent above; and a new above encoder was added to encode the user's words separately.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202309_9(9).0020

In 2017 Van-Khanh Tran et al [7] it proposed an encoder-aggregator-decoder model based on an extension of the encoder-decoder architecture of recurrent neural networks. The proposed model can jointly train sentence planning and surface implementation to produce natural language discourse.

In 2018 Wei Z, Liu Q et al [8] proposed a dialogue system oriented to automatic diagnosis, constructed the Medical DS dataset, and proposed a framework for reinforcement learning (RL) based dialogue system that can improve the accuracy of automatic diagnosis by collecting medical conditions from patients through dialogue.

In 2019 Shang-Yu Su et al [9] proposed a new framework for language comprehension and generative learning based on double-supervised learning, providing an approach that exploits pairwise duality, and experiments show that such an approach improves can significantly improve the performance of language comprehension and generative learning.

In 2020 Peng et al [10] proposed the SC-GPT model, where MR is tiled into the GPT pre-trained model and the results are obtained directly using a sequence-to-sequence approach.

In 2020 Yang et al [11] trained several large-scale generative models such as Transformer, GPT, and BERT-GPT, and then fine-tuned them in the COVID-Dialog task, which produces responses that are likely to be physician-like, relevant to conversation history, and have clinical applications.

In 2021 Yangming Li et al [12] proposed a new heterogeneous rendering machine (HRM) framework that explains how the neural generator renders input dialog acts (DA) into discourse. For each generation step, the mode switcher selects the appropriate decoder from the set of renderers to generate items (words or phrases), and this model can well explain the rendering process of the neural generator.

Based on previous research and the development of deep learning, natural language generation (NLG) tasks in conversational systems have achieved good performance. However, traditional NLG models rely on large amounts of labeled data for training, limiting their flexibility to scale domains, and it is not desirable to collect rich labeled datasets for each new domain in real application scenarios. Medical dialogue systems today face a series of challenges, requiring a large number of manually labeled doctor-patient dialogue states, while hospital data is extremely private and the data is basically not publicly available, so dialogue systems built using existing methods cannot give professional responses to the symptoms described by patients. In this paper, based on a pre-trained model, medical knowledge and noise filtering mechanisms are incorporated to make it possible to complete this medical dialogue generation process by encoding and decoding the doctor-patient context and relevant medical knowledge to help patients provide medical advice and suggestions using an effective way.

3. Research Methods

In this paper, we use a pre-trained model based on a two-stage approach, where the pre-trained model is first trained using a generic dataset and then fine-tuned on the COVID-19 dataset, whose inputs to the fine-tuned part of the model are doctor-patient dialogue contexts and relevant medical knowledge, and the outputs are medical diagnoses with the correct medical context that can be drawn on by doctors. In order to filter the appropriate knowledge as the input of the dialogue system, a noise filtering mechanism and a multi-headed attention mechanism are incorporated in this paper to extract the features of the doctor-patient context encoding and knowledge encoding, and subsequently, the knowledge gating unit calculates a reduction weight based on the matching degree of the knowledge and the context so that the medical dialogue generation is completed by encoding the doctor-patient context and the relevant medical knowledge process to help patients provide medical advice and recommendations in an effective manner.

3.1 Problem Definition

Medical conversation generation requires a round of doctor-patient conversation by a given history of doctor-patient conversations, with a representation of the patient's questions and a representation

ISSN: 2414-1895 DOI: 10.6919/ICJE.202309_9(9).0020

of the responses generated by the doctor based on the patient's questions, then a round of doctor-patient conversations, and the model takes the history of doctor-patient conversations as input. The formula is shown in equation (1).

$$\mathrm{H} = \{(q_1, r_1), (q_2, r_2), \ldots, (q_t, r_t)\} \quad (1)$$

In order to unify the data for easy coding, in the data pre-processing stage, we eliminate the conversation combinations with only patient questions or more than 6 rounds of doctor-patient conversation history, and then the model represents them as vectorized semantic representations by coding the processed doctor-patient conversations, as shown in equation (2).

$$\vec{\mathrm{H}} = \{(\vec{q}_1, \vec{r}_1), (\vec{q}_2, \vec{r}_2), \ldots, (\vec{q}_t, \vec{r}_t)\} \quad (2)$$

After encoding the history of the doctor-patient conversation, the pre-trained model fine-tunes the task, which in turn generates a response vector for the doctor's reference through the Decoder stage, which is then decoded into a semantically correct response with the correct medical context y. The formal definition is shown in equation (3).

$$\mathrm{y} = NLG(H(q_i, r_i), K) \quad (3)$$

3.2 Problem Study

3.2.1 A Medical Conversation Generation Model with Fused Knowledge

In this paper, we propose a pre-trained model-based dialogue generation model incorporating medical knowledge, which completes this medical dialogue generation process by encoding the doctor-patient context and adding knowledge encoding in the second stage for fine-tuning, making it possible for users to get a referenceable medical diagnosis without having to consult online.

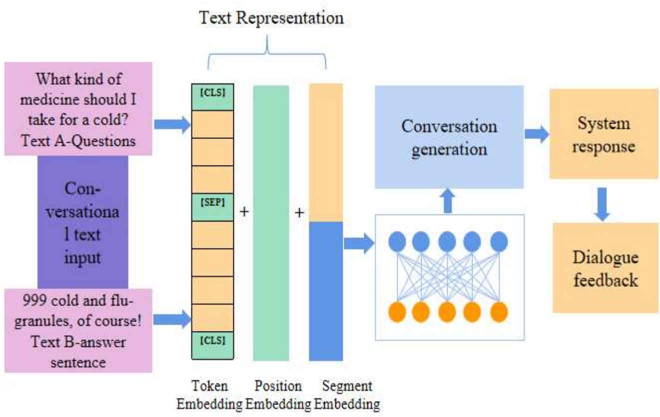

Fig. 1 MK-GPT-2 Dialogue Generation Model

The method model used is shown in Fig 1 above, which employs an end-to-end model that firstly characterizes the text of the doctor-patient context as well as medical knowledge, and secondly fuses it with the conversation context information and entity information through the mechanism of Masked Muti-head Cross Attention by means of Encoder-Decoder, so the final response has the predicted entity information, and the information to predict the next word is formulated as equation:

$$\mathrm{P(n)} = softmax(h_i(q_i, r_i) \bullet W_e^T) \tag{4}$$

where is the doctor-patient conversation history and represents a T-dimensional utterance vector. The most likely next word and the best set of entities are predicted by inference, and the optimization of F1 index is performed by candidate region sorting search, and if no corresponding content is matched then the bm25 retrieval algorithm is used to find the most similar knowledge data in the constructed retrieval model. It is calculated as shown in equations (5) and (6).

$$\mathrm{L} = \sum_{\mathrm{i \in Q}} \log \frac{r_i + 0.5 / R - r_i + 0.5}{n_i - r_i + 0.5 / N - R - n_i + r_i + 0.5} \times \frac{(k_1+1) f_i}{K + f_i} \times \frac{(k_2+1) tf_{tq}}{k_2 + tf_{tq}} \tag{5}$$

$$\mathrm{K} = k_1 [((1-b) + b \times \frac{L_d}{L_{ave}})] \tag{6}$$

Where: is the number of related documents containing the word; is the total number of related documents; is the total number of documents; is the frequency of the word in the documents; , and are empirical parameters. In the experiments, the value of b is set to 0.75. In this paper, each set of entity names and their corresponding synonyms are considered as one document for the construction of the bm25 search model.

After the model matches the most similar knowledge to the question, the predicted entity labels and dialogue context are fed into a medical dialogue generation model based on adaptive fusion of coded information, and finally the results are corrected by a consistency detection model based on contrast learning to generate a system response and thus give feedback to the user.

Bi-GRU is also used as the neural network unit because the bi-directional GRU captures not only the above conversation information but also the below conversation information. By incorporating the knowledge into the model, the accuracy of the model is greatly improved and the generated responses are more in line with the medical doctor's response to the patient's condition. The model uses the context of the doctor-patient conversation and external medical knowledge to generate a medical response, and then generates a system response to give the patient feedback on the conversation to complete the online medical consultation.

3.2.2 Encoder Module

The encoder module uses a bi-directional gate loop unit (Bi-GRU) to encode the input vectors including Word Embedding, Type Embedding and Position Embedding, and the doctor-patient conversation history is encoded as to use the Bi-GRU's gate control unit to selectively information transfer and remember key information in the doctor-patient conversation. Compared with the Long Short Term Memory Network (LSTM), the Bi-GRU used can greatly reduce the computational effort and the complexity of this model, which can solve the problem of excessive computational complexity of traditional RNN and LSTM. Formally, the output of the encoder is computed as follows:

$$\mathrm{E}_c = BiGRU(\mathrm{X}) \tag{7}$$

$$\mathrm{E}_k = BiGRU(K_c) \tag{8}$$

$$\mathrm{E}_G = \mathrm{Bi}GRU(G'_{next}) \tag{9}$$

Volume 9 Issue 9, 2023
DOI: 10.6919/ICJE.202309_9(9).0020

3.2.3 Integration Knowledge Module

The most important thing for the dialogue generation model of fused knowledge is to filter the appropriate knowledge as the input of the dialogue system, but due to the large size of external knowledge and the large number of data types, using all possible knowledge as input may lead to more noise and high computational effort. To solve these problems, a noise filter is introduced in the fused knowledge module to select better knowledge. The fusion knowledge module filters knowledge features by Knowledge Gate. Specifically, the filter first takes the output of the previous layer as the question Q and extracts the features encoded by the doctor-patient context and the knowledge encoded by the multi-headed attention mechanism: then, the Knowledge Gate control calculates a reduction weight based on the matching of knowledge and context. Finally, the filter uses averaging over the context features and knowledge features and uses the result as the output:

$$\alpha_{\mathrm{g}} = Sigmoid(W_g O_{dec}) \tag{10}$$

$$\mathrm{H} = \mathrm{W}_v O_{dec} \tag{11}$$

$$\mathrm{P}_o(y_j) = soft\max([\begin{matrix} \alpha_g H(y_j \notin \kappa') \\ H(y_j \in \kappa') \end{matrix}]) \tag{12}$$

which are trainable parameters. The noise filter controls the flow of knowledge. The reduction weights are reduced when no knowledge is required for generating responses or when the input knowledge is not contextually relevant, and vice versa. Finally, the fusion knowledge module gives the response generation to filter the medical knowledge that fits better to get a more accurate response. The structure of the Fusion Knowledge module is shown in Fig 2 below.

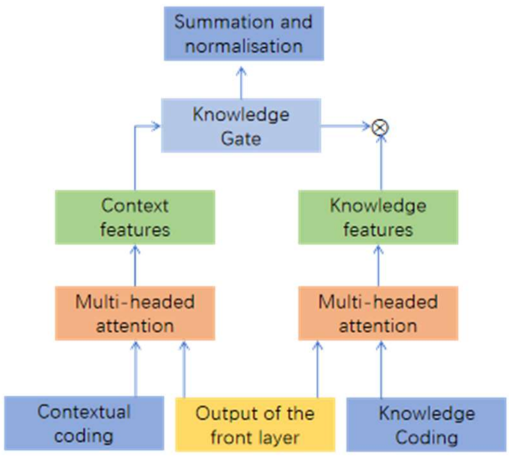

Fig. 2 Diagram of the Fusion Knowledge Module

3.2.4 Decoder Module

The decoder module merges the attention mechanism and the noise filtering mechanism to generate responses that are contextually relevant to the doctor-patient dialogue. The sequential attention mechanism aims to enhance the guidance of dialogue generation goals by simulating human cognitive processes, and the noise filtering mechanism aims to select more appropriate knowledge. In the dialogue generation process, humans will first form an overall concept of the dialogue and then give a reply response given the current doctor-patient dialogue context. So the model would have the

Volume 9 Issue 9, 2023
DOI: 10.6919/ICJE.202309_9(9).0020

decoder first process different parts of the encoder output at different layers and then combine these layers in a specific order similar to human cognition to finally generate a referenceable medical response for the patient. The order in which features are extracted in the decoder module is as follows:

$$O_p = MuitHead(I(Y_p), I(Y_p), I(Y_p)) \tag{13}$$

$$O_G = MuitHead(O_p, E_G, E_G) \tag{14}$$

$$O_{KG} = NF(O_G, E_C, E_K) \tag{15}$$

$$O_{dec} = FFN(O_{KG}) \tag{16}$$

4. Experimental Validation

4.1 Dataset

To evaluate the effectiveness of the proposed method, this paper conducts comparative experiments on a large-scale medical dataset KaMed, and a large-scale COVID-19 Chinese conversation dataset proposed by Zeng et al [10] in 2020. Among them, COVID-19 contains more than 1088 conversations about common new coronary pneumonia questions and answers, and provides fine-grained entity-level annotation; KaMed contains more departments and richer disease categories, and contains more than 17K medical conversations and 5682 entities. The datasets used in this paper, as well as the example KaMed dataset , are shown in Table 1, Table 2 below.

Table 1. Statistical table of the dataset

datasets	dialog	utterances	tokens	knowledge
KaMed	17864	153,000	6,663,272	Y
COVID-19	1088	9494	406,550	N

Table 2. Example of KaMed dataset

HEAD	RELATION	TAIL
Fatigue	related_to	dizzy
Fever	related_to	fatigue
Lactulose	to_treat	constipation
Lactulose	adverse_reaction	stomachache

4.2 Experimental Setup

(1) Hardware

The computer CPU for the experiment of this paper: Intel I9-10900K, graphics card Nvidia Geforce 2080Ti*2, 64G RAM, SSD 512G.

(2) Software

Systems: Windows 10, Ubuntu Server 2.0;

Development platform and tools: Pycharm, Visual Studio Code, Pytorch, Neo4j, etc.;

Development environment: Python-3.7, Pytorch, TensorBoard, etc.

(3) Parameter setting

Table 3. Model parameters

Parameter	Meaning	Values
epochs	Training rounds	50
batch_size	Training size	24
lr	Learning Rate	1.5e-4
warmup_steps	warm up	2000

4.3 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. Among them:

1) BLEU [12]: this evaluation metric was first used to evaluate machine translation, which is essentially to determine the similarity between the expected and response sentences by calculating the co-occurrence word frequency of the two sentences, i.e., the BLEU score is calculated by counting the number of occurrences of n-gram phrases in the generated and real responses in the whole training corpus. In general, the larger the BLEU value is on a given dataset, i.e., the better it is considered. The calculation formula is shown in equation (17) (18):

$$\mathrm{BLEU} = \mathrm{BP} \bullet \exp\left(\sum_{n=1}^{N} w_n \log P_n\right) \tag{17}$$

$$\mathrm{BP} = \begin{cases} 1 & ,c > r \\ e^{1-r/c} & ,c \le r \end{cases} \tag{18}$$

2) Distinct [13]: used to measure the diversity of the text in the generated scenes, and the larger its value indicates the higher diversity of the generated text. The calculation formula is shown in equation (19):

$$\mathrm{Distinct}_n = \frac{Count(unique \quad \mathrm{ngram})}{Count(word)} \tag{19}$$

where: represents the number of unduplicated words in the response and represents the total number of words in the response.

3) Perplexity perplexity (PPL): Used to estimate the accuracy of a language model, this evaluation metric reflects how accurate the model is in generating its own target sequence. Generally speaking,

the smaller the perplexity value, the better the model is for a given data set. The calculation formula is shown in equation (20):

$$\mathrm{PP(W)} = \mathrm{P}(\omega_1\omega_2\cdots\omega_N)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(\omega_1\omega_2\cdots\omega_N)}} \tag{20}$$

Where: W represents the current sentence evaluated, N is the length of the current sentence being evaluated, and inside the root sign is the inverse of the sentence probability, so obviously if the value is smaller, i.e., the less confusion, the better the sentence is (larger probability, i.e., larger value).

4.4 Experimental Results

In order to evaluate whether the model proposed in this paper is better than the previously proposed model, this paper uses Seq2Seq end-to-end model with attention mechanism, HRED model and our model are experimented with KaMed, COVID-19 two datasets, first of all, in the experimental process, in order to determine the maximum attention score of each word in the response to the context and related knowledge, the confusion matrix is used to summarize the generated results, and the visualization matrix of different models is shown in Fig. 3~6 below.

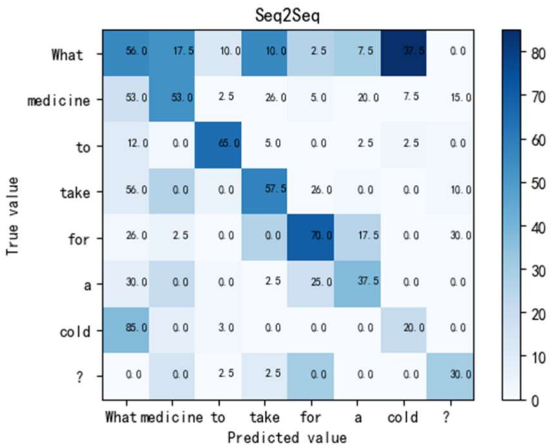

Fig. 3 Seq2Seq's attention score

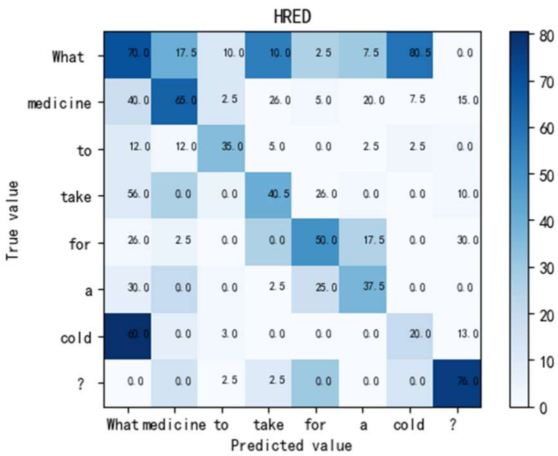

Fig. 4 HRED's attention score

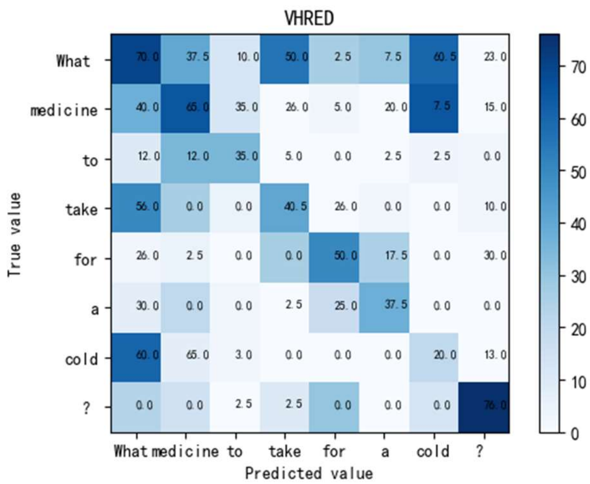

Fig. 5 VHRED's attention score

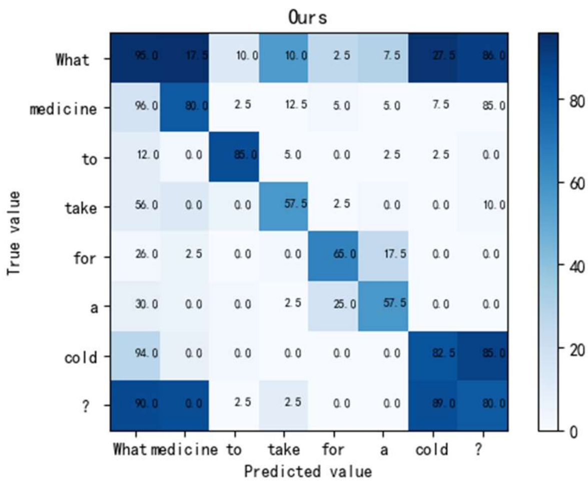

Fig. 6 Ours attention score

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 to the Seq2Seq model, the attention score of such entities has increased to a certain accuracy, and finally our model that integrates knowledge has the most accurate attention score for the conversation context, which is for "cold" and "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 4 below.

Table 4. Experimental results of different models on different data sets

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	**2.86**	**1.86**	**1.58**	**8.56**	**23.91**
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	**3.56**	**5.90**	**5.89**	31.21	**40.8**

Experiments show that our method performs better on two different datasets. Using our method on the KaMed dataset has a Perplexity value that is 0.91 less than the baseline Seq2Seq method, and the smaller the value, the higher the accuracy of our method. Compared with the VHRED model, our method has increased B@1 values by 0.37, B@2 values by 0.31, and D@2 values by 2.14, all indicating that our method is better and still our method is better on the dataset COVID-19. An example of the dialogue under test of our model is shown in Fig 7 below:

Volume 9 Issue 9, 2023
DOI: 10.6919/ICJE.202309_9(9).0020

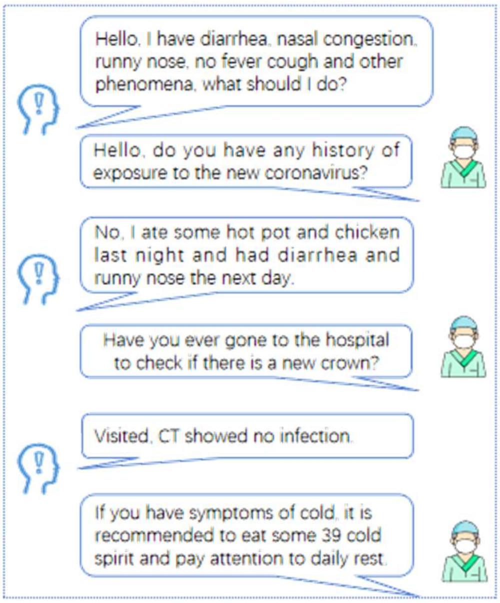

Fig. 7 Example of model generation

5. Conclusion and Perspectives

In this paper, a medical dialogue generation mode is provided by adding external knowledge, which pioneers the integration of a large amount of external knowledge under the condition that the scalability and portability of the model are improved by using pre-trained models, and realizes the task of medical response by using knowledge ontology tracking and two-stage cognitive entity prediction. Based on a large number of experimental data on COVID-19 on KaMed and Online, the validity of the proposed model is verified based on the fact that the medical dialogue generation mode proposed in this paper is higher than most of the baseline patterns in BLEU value, Discrete value and Perplexity value.

In the next step of development, it is planned to add the patient's thinking and emotional factors to the dialogue generation mode, so that the model can better understand the sentences understood by the patient, and hope to get more personalized and targeted feedback, so as to improve the performance of pattern generation and response.

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, 2022,36(01):1-11+20.

[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 (12):2889-2900.

[3] Wen T H, Gasic M, Kim D, et al. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking[J]. arXiv preprint arXiv:1508.01755, 2015.

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

[5] Dušek O, Jurčíček F. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings[J]. arXiv preprint arXiv:1606.05491, 2016.

[6] Dušek O, Jurčíček F. A context-aware natural language generator for dialogue systems[J]. arXiv preprint arXiv:1608.07076, 2016.

[7] Tran V K, Nguyen L M. Neural-based natural language generation in dialogue using rnn encoder-decoder with semantic aggregation[J]. arXiv preprint arXiv:1706.06714, 2017.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202309_9(9).0020

[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]. arXiv preprint arXiv:1905.06196, 2019.

[10] Peng B, Zhu C, Li C, et al. Few-shot natural language generation for task-oriented dialog[J]. arXiv preprint arXiv:2002.12328, 2020.

[11] Yang W, Zeng G, Tan B, et al. On the generation of medical dialogues for COVID-19[J]. arXiv preprint arXiv:2005.05442, 2020.

[12] 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.

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

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

[15] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[16] Compton R, Valmianski I, Deng L, et al. MEDCOD: A Medically-Accurate, Emotive, Diverse, and Controllable Dialog System[C]//Machine Learning for Health. PMLR, 2021: 110-129.

[17] Rojowiec R, Roth B, Fink M. Intent recognition in doctor-patient interviews[C]//Proceedings of the Twelfth Language Resources and Evaluation Conference. 2020: 702-709.

[18] Lin S, Zhou P, Liang X, et al. Graph-evolving meta-learning for low-resource medical dialogue generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(15): 13362-13370.