Intelligent Computing Technology and Automation
Z. Hou (Ed.)

doi:10.3233/ATDE231211

Question-and-Answer Juxtaposition Based Chat Model for Multi-Turn Medical Inquiry

Yujiang LIU[a,b,1], Lijun FU[b], Xiaojun XIA[a,b] and Zhijun CHANG[a,b]
[a] *Department of Computer Application Technology, Shenyang Institute of Computing Technology Co.Ltd., Shenyang, Liaoning, 110168, China*
[b] *University of Chinese Academy of Sciences, Beijing, 100049, China*

Abstract. AI chatbots talk to patients in multi-turn conversations in order to imitate real doctors and meet the patients' needs during prior medical inquiries. However, existing works ignore the fact that doctors can convey two different tones answers to patients through feedback information, that is, question and statement. This inflexible method urges us to seek a new response strategy that outputs a tone indicator and uses it to limit the range of words output. In this paper, we propose a novel chat model based on question-and-answer juxtaposition (QAJCM), which simultaneously optimize three parts: the basic response, the generated counterpart, and the type of the basic response. These are all produced from the basic response, the real result. In specific, the model outputs a question, a statement, and a type to ensure that the type gives correct choice in the first two parts. Experiments on MedDG and HaoDF datasets show that the BLEU-average scores are up to 20.28 and 3.03, which are 45% and 49% higher than the baselines, respectively. According to the experimental results, we have made obvious improvements and verify that our work can respond to patients in a correct tone in the medical inquiry scenario.

Keywords. AI chatbots; medical inquiries; tone; question and statement

1. Introduction

Multi-turn medical inquiry refers to the process of engaging in a back-and-forth conversation with a user to provide relevant and accurate medical information. Machine medical inquiry applications such as AI chatbots and virtual assistants have been developed to help users obtain medical information, arrange appointments, and answer general health-related questions [1-2]. These systems leverage techniques such as natural language processing, representation learning, and knowledge graphs [3] to understand user queries and provide appropriate responses. They are developed with medical parameterized priori knowledges stored in the parameters of the neural networks [4-5]. While there have been significant advancements in machine learning and natural language processing, existing chat models still face challenges when it comes to multi-turn medical inquiries [6]. These models do not pay attention to the tones and the corresponding content.

[1] Corresponding Author: Yujiang Liu; Department of Computer Application Technology, Shenyang Institute of Computing Technology Co.Ltd., Shenyang, Liaoning, 110168, China; liuyujiang16@mails.ucas.ac.cn

The doctors' tones can be divided into two parts: question and statement. This question means that the doctors intend to ask more details about the patients' symptoms. Then, the statement indicates that the doctors will arrive at a conclusion based on previous communication with the patient. Although some previous works have also noticed this phenomenon, they do not solve it by considering the tones [7-10]. The indistinguishable tones of the previous works may cause the following three problems:

Conflating questions and answers: One major issue with existing chat models is their inability to distinguish between questions and answers clearly. For example, a response should be a statement but the patient receives a question. This can lead to confusion, because the chatbot may provide an irrelevant or incorrect response to the user's query. After several rounds of chatting, the reply may be lengthy and the patient's purpose may be missed.

Lack of a clear point of views: This situation corresponds to endless problems of the responses. This question gives patients a sense of indecision. This is because the responses cannot be switched to a statement so the patients always receive indefinite information. Users expect accurate and reliable information when discussing health-related topics, but the lack of a clear stance can undermine the credibility of the chatbot.

Jump to a conclusion: It seems that the chatbots has solved the problem of the patients but it is very likely that the words are useless. This causes great hardship to maintain context in multi-turn conversations, leading to disjointed and unhelpful exchanges. It is especially problematic in medical consultation, because understanding the background and the patient's concerns is crucial for providing accurate and helpful information.

To address these challenges, further research and development is required to improve the performance of AI-based medical inquiry models. Although the generated text embodies the tones [11], the chat model does not actively control the text with a certain tone. According to the task of dialogue scenario, the data contains history, prompt and response. As shown in Figure 1, if the response is converted to two parts [12], the content output and the tone output, the problem can be properly solved. This thought presents a clear outline for us to overcome the difficulty of mixed tones in multi-turn medical inquiry.

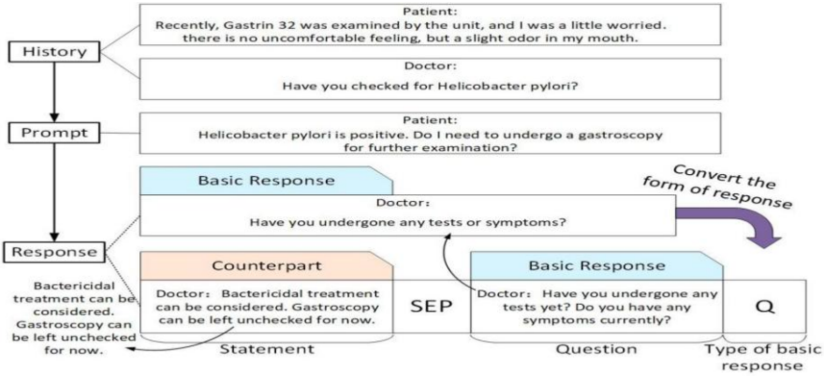

History
Prompt
Response
Patient:
Recently, Gastrin 32 was examined by the unit, and I was a little worried. there is no uncomfortable feeling, but a slight odor in my mouth.
Doctor:
Have you checked for Helicobacter pylori?
Patient:
Helicobacter pylori is positive. Do I need to undergo a gastroscopy for further examination?
Basic Response
Doctor:
Have you undergone any tests or symptoms?
Convert the form of response
Bactericidal treatment can be considered. Gastroscopy can be left unchecked for now.
Counterpart
Doctor: Bactericidal treatment can be considered. Gastroscopy can be left unchecked for now.
SEP
Basic Response
Doctor: Have you undergone any tests yet? Do you have any symptoms currently?
Q
Statement
Question
Type of basic response

Figure 1. Wrong case A demo of the response transition process. The response is converted into a basic response (the source text), the counterpart, and the type. We define the statement at the beginning of the combination sequence. It is separated from the question by the delimiter "SEP". The type contains only two indicators: "Q" means to choose the question while "S" means to select the statement.

First, we train a sentence classification model for the response content, which is considered as the basic response. We use it to label all the basic responses and the sentence type of the training data. Second, considering the basic response of each data has only one tone, we train two mono-tone chat models to generate supplement response for the responses, which is regarded as the counterpart. After these two steps, we get the training data with the format that we need, an original response modified to a basic response type and a union response content, which contains a basic response and a counterpart. Finally, we purpose a mixed chat model with a segmentation result layer for the three outputs. Inspired by the joint model [13-14], we design an internal joint loss and realize joint adjustment of multiple results. These outputs share the same input and hidden layers, but are monitored by the different functions. In addition, we design three masks to cover the different parts of the hidden states to increase the discrimination of the results. The result of the type can supervise the content of the union response at the same time. In short, our mixed chat model can output the tone type and the union response, so that we can find the correct response with the type. Experiments show that our model outperforms state-of-the-art methods with medical inquiry datasets. And our contributions are as follows:

We prepare the response data in two sources. We organize the data into two parts — questions and statements. This will ensure the model outputs two tone types of responses as two candidates.

We use a binary mixture loss and a set of three masks to optimize the model to generate an auxiliary decision. The combination of three loss functions is designed to optimize the model performance. This will help the model to consider different aspects of the generated information, such as their relevance and accuracy.

2. Related work

2.1 The development of the multi-turn medical inquiry in machine learning

The development of multi-turn medical inquiry in machine learning has advanced significantly in recent years, offering enhanced AI-driven conversational agents that can assist users in obtaining medical information and providing personalized healthcare advice [15-17]. Initially, medical chatbots and conversational agents were limited to single-turn interactions, where users would ask a question and receive a predefined response [18]. These early approaches lacked the ability to engage in dynamic and context-aware dialogues. The introduction of context-aware conversational agents provides multiple rounds of functions, enabling them to maintain context and respond to user inputs intelligently [19]. This is made possible through advances in natural language processing (NLP) and natural language understanding (NLU) techniques, allowing the inquiry chat models to better comprehend user queries and provide relevant answers. The integration of medical knowledge base provides data foundation for the AI methods [20-21]. With the addition of medical knowledge base, such as electronic health records, medical literature, and expert-generated content, AI models can provide accurate personalized information [22-24]. This ensures that users can receive reliable and medically reasonable suggestions. More important, the medical AI community has realized the importance of collaboration and standardization in promoting this field [25]. In summary, multi-turn medical inquiry has come a long way, evolving from simple question-answer applications to sophisticated AI-driven conversational agents. These advancements have the potential to revolutionize healthcare by providing users with accurate, personalized, and context-aware medical information, and ultimately improving patients' outcomes and overall medical care experiences [26].

2.2 Recently chat models

Recently, generative pre-trained model has risen in dialogue field, which has significantly impacted the field of medical inquiry. The original GPT model, released in 2018, was built on the Transformer architecture introduced by Vaswani et al. in 2017 [27]. It makes use of unsupervised pre-training on large amounts of text data followed by fine-tuning on specific tasks. An improved and larger version named GPT-2, was proposed in 2019 [28]. It consists of 1.5 billion parameters and showcases a wide range of capabilities, including language translation, question-answering, and text summarization. GPT-3 [29] demonstrate remarkable performance in various tasks with minimal fine-tuning, including translation, summarization, and code generation, with 175 billion parameters, making it the largest language model after 2020. ChatGPT comes from InstructGPT [30], which is a variant of GPT-3 and specifically designed for dialogue-based applications [31]. It has been fine-tuned to understand and respond to user queries more effectively, providing a more interactive and engaging experience [32]. Following the same lineage as the GPT models, ChatGLM [33-34] is released in this year. It is designed to improve upon the limitations of earlier GPT models, such as handling multi-turn conversations and providing more coherent responses. Several works have employed it in their models [35-36]. In this paper, we choose ChatGLM as our basic model, because it is open source and contains priori knowledge. We use the embedding part and the encoding layers.

3. Method

3.1 Problem overview

We regard this task as a response to the dialogue in the medical care scene, with the length of n prompt sentence $P = \{x_1, x_2, \dots, x_n\}$, a history information groups $H = \{[S_1, D_1], [S_2, D_2], \dots, [S_m, D_m]\}$ with m pairs of the patient's and doctor's word. We allow these to be encoded with our chat model and output two results, the basic response type t and the union response content R of the simulated doctor. The overview of QAJCM is shown in Figure 2.

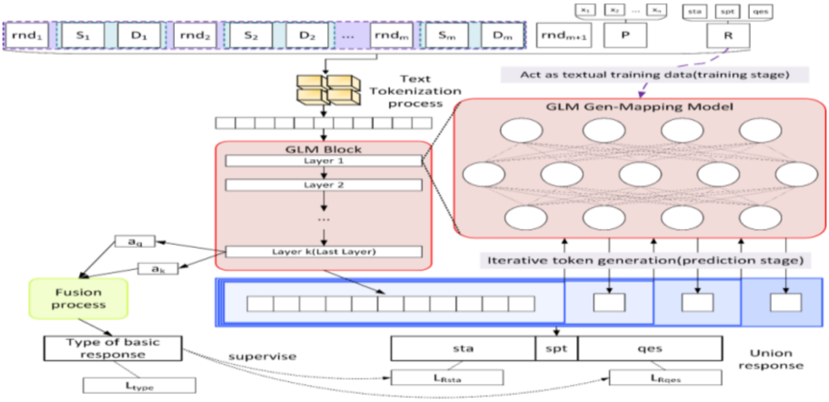


Figure 2. Overview of the QAJCM

3.2 Data generation method

We generate counterpart for each sample with question or statement, because the original data doesn't contain these data. In this method, BERT model [37] is used to generate a sentence classification model and help to divide the question response and the statement response in each stage of the conversation. After that, we train two basic chatGLM for generating a question-based chat model and a statement-based chat model. If a basic response is a question, it can be regarded as lack of the statement. The counterpart allows the mixed chat model to fine-tune the training data with the minimum cost of the loss. In this way, each sample will contain a basic response and a text with the corresponding supplementary tone for the counterpart.

The preparations provide a new output form for the union response content R. No matter what the type of the basic response is, its order is the statement part, then the question. These two parts are linked by special tokens "/a/". This can be described as follows:

$$R = concat[sta, spt, qes] \tag{1}$$

Where sta and qes are the statement part and question part; spt represents the special tokens. The function "$concat$" means to connect them into a sequence. Therefore, R can be considered as a union response.

3.3 Encoder part

The encoder gives the token representation of the combined sequence in the sample. For the consideration of the prior knowledge, we choose the chatGLM-6B model [34]. It absorbs the deep representation and outputs a sequence of token vector. We follow the encoding principles of chatGLM-6B, with $V = \{rnd_1, [S_1, D_1], rnd_2, [S_2, D_2], \dots, rnd_m, [S_m, D_m], rnd_{m+1}, P, R\}$ as the input. The $rnd_1, rnd_2, \dots, rnd_{m+1}$ denotes the special tokens "$Round_i(1 <= i <= m + 1)$" for separating the dialogue rounds in history. According to it, the prompt and the union response are merged into history so that the encoder will capture all the context information. It will be used for the next calculation.

3.4 Separated method of result set prediction

The GLMBlock in every layer of the Encoder outputs three vectors which can be used to predict the type and the union response. They are GLM hidden state, attention query, and attention key of the corresponding layer outputs. We design three masks to divide the intermediate results. This novel method can effectively reduce mutual interference. We collect all these vectors from the last layer, and do the following.

For the type t, we employ the two attention vectors as the intermediate vector. The history-prompt mask is used to cover the non-history and non-prompt part with value of 0, remaining others with value of 1. We design a weighted average pooling method for getting the compressed representation for the two attention vectors. After that, the representation is sent to a linear mapping with a sigmoid function and converted into a binary result. This result can represent the type t of the sample. This fusion process can be described as follows:

$$avg_{att} = Avgpool(a_q, a_k) \tag{2}$$

$$h_t = avg_{att} * Softmax(m_t) \tag{3}$$

$$p_t = Sigmoid(W_t * h_t + b_t) \tag{4}$$

Where the avg_{att} is the average pooling result of merging two attention vectors; the h_t is the compressed representation of the type result after the above multiplied by history-prompt mask m_t; the p_t is the prediction result of the type t; the W_t is the trainable weights and the b_t is the bias. We adopt sigmoid function to get a value between 0 and 1.

We multiply the question mask and the statement mask by the two duplicated GLM hidden states to generate two different candidate vectors for the union response. The question mask covers the non-question response part with value of 0, leaving the value of 1 for other parts. The statement mask assigns weight to the non-statement part and other parts with the same pattern. By the way, the hidden state can be transformed into two different distributions of values and the distinction of representation can be enhanced. These two candidate vectors are mapped to the result space by a linear function, and their output dimension are the same as the number of the words in the space. The method of getting candidate vectors is described as follows:

$$cv_1 = m_q * h \tag{5}$$

$$cv_2 = m_s * h \tag{6}$$

$$p_q = W_o cv_1 + b_o \tag{7}$$

$$p_s = W_o cv_2 + b_o \tag{8}$$

Where the cv_1 and cv_2 are the coverage values in the question mask m_q and the statement mask m_s; the p_q and p_s are the forecasting question and statement parts in our model. We allow the coverage values to share the same word mapping function with the trainable weights W_o and bias b_o.

3.5 Loss Calculation Strategy

The loss function is designed as the joint loss. The joint loss establishes an association between the basic response type t and the union response content R, thus allowing the parameters to be collaborative adjustment in the backpropagation stage. We employ binary cross entropy for the type loss, and cross entropy for both the question part and the statement part. Because the counterpart in the union response is necessary to obtain the output, but it needs to be weakened in the backpropagation, we propose a new combination scheme by using the real type of the basic response. This type will be changed into two binary one-hot values. If the true type value is 0, it means that the basic response is a statement and we should discard the question part. Therefore, the front value from it is 1 and the back value is 0. After that, we multiply it with the corresponding loss to ensure that the total loss is not affected. However, this process can only ignore the independent part, but cannot enhance the semantic representation of the basic response part. Thus, we add a type coefficient from the forecast type value and make it a new multiplier. Considering that the predicted value is between 0 and 1, we add number 1 to the coefficient, so that the corresponding loss can contribute twice as much to the total loss at most. This process is a kind of inner joint model which uses the same input but gets two results from different loss functions, and merges them with self-adaptive coefficients together. The method of calculating losses is described as follows:

$$L_{type} = -w_n * [r_t \log(p_t) + (1 - r_t) \log(1 - p_t)] \tag{9}$$

$$L_{Rqes} = -\frac{1}{u} \sum {}_{z=1}^{u} y_z^{(q)} log\left(p_z^{(q)}\right) \tag{10}$$

$$L_{Rsta} = -\frac{1}{j} \sum {}_{z=1}^{j} y_z^{(s)} log\left(p_z^{(s)}\right) \tag{11}$$

$$r_{front}, r_{back} = \neg xor(r_t, 0), \neg xor(r_t, 1) \tag{12}$$

$$L_{total} = L_{type} + r_{front} * (2 - p_t) * L_{Rsta} + r_{back} * (1 + p_t) * L_{Rqes} \tag{13}$$

Where the w_n is the trainable weights of type t; the r_t is the true type of the basic response (same to t); the L_{Rqes} and the L_{Rsta} are the losses of the question part and the statement part; the y_z and the p_z are the real token and the predicted token in the union response, with q and s to indicate the question and the statement sources; the r_{front} and the r_{back} represent the front value and the back value from r_t, which are generated by exclusive or operation(XOR) and reverse option; the L_{total} is the total loss. The $(2 - p_t)$ and $(1 + p_t)$ are strengthen factor, which come from 1 plus the unenhanced state $(1 - p_t)$ and p_t for the statement part and the question part.

4. Experiments

4.1 DataSet

In order to make an impartial and significant comparison, we follow the previous works to evaluate two Chinese medical inquiry datasets, MedDG [38] and HaoDF. Both of them are processed into JSON files, and the content structure of the file is a doctor's words followed by a patient'. The essential difference between them lies in the source of data. MedDG is an open-source dataset which can be downloaded from the website. It has separate training, testing, and validation dataset, and we use the testing part to compare with other baselines. HaoDF is private data that we collect from the internet and sponsors. We divide the training, testing, and validation part at a ratio of 8:1:1, and record the performance of the testing data. Table 1 shows the detailed information of the datasets, such as numbers of the samples and their average turns. Except for the performance comparison of the testing data, we also study several revised models and changed data.

Table 1. Basalt fiber physical index

Dataset	Train	Valid	Test	Average talking turn
MedDG	14864	2000	1000	11.62
HaoDF	12597	1000	1000	8.25

4.2 Evaluation

We compare our model with basic pre-trained chat models and other NLP models. Some MedDG experimental results of these models are from the record of original papers. HaoDF results are produced by us with our training resources. We adopt the word-overlap based metrics, BLEU [39] scores, which scrutinizes each predicted word with the basic response. According to the type of the basic response, we select the corresponding part as the real answer. Specifically, we use BLEU-2, BLEU-3, BLEU-4 and the average value of them. We will discuss the joint strategy in section 4.4 and give statistical results to show the performance of our model.

Because the medical response generation is an NLG task, we do not compare our methods with the works that show the results of intention slot task but use the same dataset. They [40-41] only predict the keywords in the response rather than generating the whole sentence. We compare our results with the baselines of external knowledge strategies [42-44] and internal enhancement strategies [45-48].

4.3 Implementation details

We use chatGLM-6B as the encoder for the inputs, which contains prior knowledge of the parameters of the network. We set the batch size to 1 as the number of the parameters is too large for the GPU memory. The epoch of all training dataset is set to 20. We choose the AdamW optimizer and set decaying strategy for the learning rate. The initial learning rate is 2e-2. The training equipment we use contains 2 * Nvidia Tesla M40. We get the best performance model on testing dataset and compare it with other baselines.

4.4 Main result

Table 2 and Table 3 show the results of our model processing the two datasets. Compared with other baselines, our model outperforms all other models in BLEU-average. To be specific, our model has been improved by 5.8 BLEU-average on MedDG and 1.0 BLEU-average on HaoDF. This can prove that our model is valid.

The main reasons for the improvement are multi results output strategy and internal joint loss method. First, the two outputs take the type as an indicator and the union response as a candidate set of two different contents. The indicator determines which part of the contents should be chosen. Because the generated words come out one by one, if the tone is wrong, they will suffer from error propagation. However, the indicator provides a definite tone, and no error is allowed. This makes a tremendous contribution to performance improvement. Second, the internal joint loss ensures collaborative calculation and finetuning. Although both outputs are controlled by different loss functions, they share the same encoding layers. This illustrate that the difference between them comes from the last layer. Therefore, the internal joint loss can coordinate the parameters of each layer and adjust them comprehensively. This optimization idea guarantees high performance and low computational overhead. By contrast, other baselines do not separate the question part and statement part for the original response, so their performances are lower than ours.

Table 2. Main results of the MedDG dataset. Note that * means the reproductions of the baselines.

Methods	BLEU-2	BLEU-3	BLEU-4	BLEU-average
MKA[42]	8.09	5.65*	2.87	5.54
TAMDE[43]	10.11*	6.83*	5.18*	7.37
TAMDG[44]	18.01*	13.72*	5.84*	12.53
GDEMR[45]	19.12*	14.13*	6.12*	13.12
SVRMDG[46]	20.46*	15.02*	6.31*	13.93
EDG[47]	7.80	5.42*	2.81*	5.34
PlugMed[48]	21.13*	14.95*	6.00	14.03
QAJCM(Ours)	33.70	17.67	9.47	20.28

Table 3. Main results of the HaoDF dataset. Note that * means the reproductions of the baselines.

Methods	BLEU-2	BLEU-3	BLEU-4	BLEU-average
MKA[42]	1.97*	0.40*	0.06*	0.81
TAMDE[43]	2.54*	0.69*	0.10*	1.11
TAMDG[44]	4.41*	1.02*	0.12*	1.85
GDEMR[45]	4.50*	1.05*	0.12*	1.89
SVRMDG[46]	4.65*	1.23*	0.18*	2.02
EDG[47]	1.94*	0.38*	0.06*	0.79
PlugMed[48]	4.85*	1.02*	0.19*	2.03
QAJCM(Ours)	7.66	1.22	0.20	3.03

4.5 Ablation study for model

We delete the strengthen factor to compare the unenhanced loss with the original model. As shown in table 4, the BLEU scores witness a slight drop. When the loss is fed back into the correct words of the basic response, the unenhanced one gives fewer adjustments. In addition, with the increase of training epochs, this gap will increase with the passage of time. Therefore, the strengthen factor is necessary for the internal joint loss.

We also compare the loss of the original model with that of the simple summation. In this experiment, we neglect all the loss ratios and just add them up. As shown in table 4, the BLEU scores decrease. Actually, the simple sum disables the mask of the counterpart. Therefore, the accumulated loss adds an unnecessary value and may lead to the inaccuracy of the backpropagation stage.

Table 4. Ablation Study for model. We use the MedDG dataset for this group of experiments.

Methods	BLEU-2	BLEU-3	BLEU-4	BLEU-average
QAJCM(Src)	33.70	17.67	9.47	20.28
- Strengthen Factor	33.65	17.62	9.45	20.24
-Ratio of Losses	31.92	16.45	8.3	18.89

4.6 Ablation study for data

We convert the task into a non-joint task, in which the three parts of the result are concatenated into a sequence. The model only needs to generate the words one after another. As shown in Table 5, the BLEU scores decrease by nearly 1.5 because the inherent inconsistency of this type of output reduces the flexibility of the model. The type of the basic response contains only two results, but it is bound to the content of the union response. This leads to a limited optimization and impairs performance.

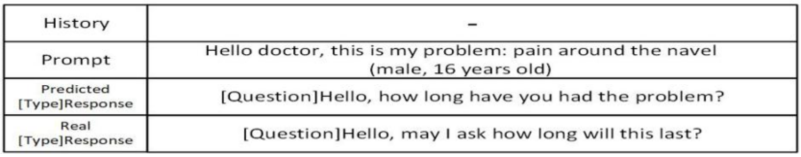
Another study is to split model according to the type of the output. We use two models, one for predicting the type of the basic response and the other for predicting the content of the union response. As shown in Table 5, the BLEU scores decrease by 2.36. When splitting, the loss of the type will not affect the content. In this way, the content model globally optimizes both the basic response and the counterpart. However, the counterpart should be overlooked in the training stage. As a result, the content model pays more attention on the counterpart, and the basic response is not optimized enough. Furthermore, since this study contains two independent models, more time is needed to train data.

Table 5. Ablation Study for data. We use the MedDG dataset for this group of experiments.

Methods	BLEU-2	BLEU-3	BLEU-4	BLEU-average
QAJCM(Src)	33.70	17.67	9.47	20.28
Non-joint QAJCM	30.83	16.58	8.67	18.69
Splited Two Models of QAJCM	30.11	15.71	7.94	17.92

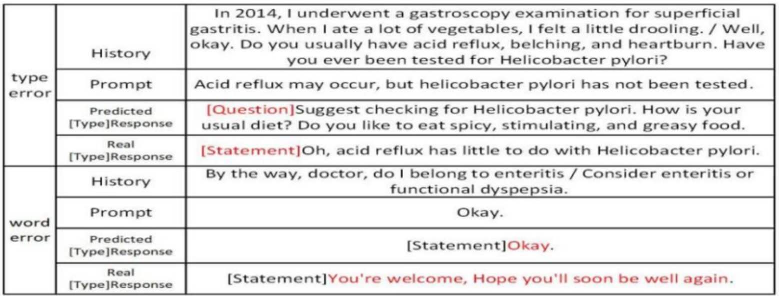
4.7 Case Study

We show a right case and a wrong case to illustrate the effectiveness and the shortcomings of our model. The cases are from the MedDG dataset. We study both the type part and the content part. The red background color indicates the type error and the word error. The right case is shown in Figure 3 and the wrong case is shown in Figure 4.



History	–
Prompt	Hello doctor, this is my problem: pain around the navel (male, 16 years old)
Predicted [Type]Response	[Question]Hello, how long have you had the problem?
Real [Type]Response	[Question]Hello, may I ask how long will this last?

Figure 3. Right case



type error	History	In 2014, I underwent a gastroscopy examination for superficial gastritis. When I ate a lot of vegetables, I felt a little drooling. / Well, okay. Do you usually have acid reflux, belching, and heartburn. Have you ever been tested for Helicobacter pylori?
	Prompt	Acid reflux may occur, but helicobacter pylori has not been tested.
	Predicted [Type]Response	[Question]Suggest checking for Helicobacter pylori. How is your usual diet? Do you like to eat spicy, stimulating, and greasy food.
	Real [Type]Response	[Statement]Oh, acid reflux has little to do with Helicobacter pylori.
word error	History	By the way, doctor, do I belong to enteritis / Consider enteritis or functional dyspepsia.
	Prompt	Okay.
	Predicted [Type]Response	[Statement]Okay.
	Real [Type]Response	[Statement]You're welcome, Hope you'll soon be well again.

Figure 4. Overview of the QAJCM

For the right case, we come to the conclusion that our model can generate a meaningful covering sentence compared with the basic response. In addition, it can also generate an inference sentence for the counterpart. The prediction type gives their choice and output the correct response.

For the wrong case, we can divide it into two parts, type error and word error. The type error may occur when the response of the testing data is highly subjectivity. Suppose a doctor ought to answer a question to the patient and ask for more details, but only provide a diagnosis. If this conversation is recorded into the dataset, it will misguide the model training or increase the error rate. The word error may occur when the predicted words do not cover the words of the basic response words. If the doctor's reply contains some dialect words or supplements, it may not be in line with the forecast.

5. Summary

In this paper, we propose a new internal joint loss method to predict the tone of the upcoming response and output the corresponding reply. This method limits the probability space of the output by modifying its structure. Such approach provides a clear tone and accurate information. In this way, we offer a tone-based solution in the AI medical inquiry scene. This tone can guide the chat model to output an appropriate response for patients, and reduce the workload of doctors by assisting doctors to cognize patients' situation in advance. Experimental results show that our method is superior to all baselines and has obvious improvement on two datasets. Further research on model structure and data form shows that this method has popularization and practical value. In future, we will try to find a detailed scheme to divide the questions and the statements into more complex tones to study how the combination of tones affects the response.

Acknowledgements

Upon the completion of the thesis, I would like to take this opportunity to express my sincere gratitude to my two supervisors, Professor Fu Lijun and Professor Xia Xiaojun, who have given me important guidance on the thesis. This work was supported in part by the National Social Science Foundation of China under Grants 21BTQ06. The work of Yujiang Liu, Lijun Fu, Zhijun Chang was supported in part by the Beijing Zhongke Ruitong Information Technology Co., Ltd, Beijing, China.

References

[1] Tripathy, Sushreeta, Rishabh Singh, and Mousim Ray. "Natural Language Processing for Covid-19 Consulting System." Procedia Computer Science 218 (2023): 1335-1341.

[2] Chang, I. C., Shih, Y. S., & Kuo, K. M. (2022). Why would you use medical chatbots? interview and survey. International Journal of Medical Informatics, 165, 104827.

[3] Xia F, Li B, Weng Y, He S, Liu K, Sun B, Li S, Zhao J. MedConQA: Medical Conversational Question Answering System based on Knowledge Graphs. InProceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations 2022 Dec (pp. 148-158).

[4] Rajula HS, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina. 2020 Sep 8;56(9):455.

[5] Qiu XZ, Yuan CW, Bi N, Huang MC, You CW. Exploring the Challenges and Opportunities in Developing Systems to Improve Alcohol Use Disorder through Chatbot Technology. InExtended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems 2023 Apr 19 (pp. 1-5).

[6] Luo B, Lau RY, Li C, Si YW. A critical review of state‐of‐the‐art chatbot designs and applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2022 Jan;12(1):e1434.

[7] Yang, Tianqing, Tao Wu, Song Gao, and Jingzong Yang. "Dialogue Logic Aware and Key Utterance Decoupling Model for Multi-Party Dialogue Reading Comprehension." IEEE Access 11 (2023): 10985-10994.

[8] Y. Li and H. Zhao, ‘‘Self- and pseudo-self-supervised prediction of speaker and key-utterance for multi-party dialogue reading comprehension,’’ in Proc. Findings Assoc. Comput. Linguistics: EMNLP, Punta Cana, Dominican Republic, Nov. 2021, pp. 2053–2063. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.176

[9] C. Li and J. D. Choi, ‘‘Transformers to learn hierarchical contexts in multiparty dialogue for span-based question answering,’’ in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, Jul. 2020, pp. 5709–5714. [Online]. Available: https://aclanthology.org/2020.acl-main.505

[10] J. Liu, D. Sui, K. Liu, and J. Zhao, ‘‘Graph-based knowledge integration for question answering over dialogue,’’ in Proc. 28th Int. Conf. Comput. Linguistics, Barcelona, Spain, Dec. 2020, pp. 2425–2435. [Online]. Available: https://aclanthology.org/2020.coling-main.219

[11] Kusal S, Patil S, Choudrie J, Kotecha K, Mishra S, Abraham A. AI-based Conversational Agents: A Scoping Review from Technologies to Future Directions. IEEE Access. 2022 Aug 23.

[12] TAŞAR DE, Şükrü OZ, KUTAL S, ÖLMEZ O, GÜLÜM S, Fatih AK, BELHAN C. Performance Trade-Off for Bert Based Multi-Domain Multilingual Chatbot Architectures. Journal of Artificial Intelligence and Data Science.2021;1(2):144-9.

[13] Zheng, Hengyi, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. "PRGC: Potential Relation and Global Correspondence Based Joint Relational Triple Extraction." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 6225-6235. 2021.

[14] Li Z, Fu L, Wang X, Zhang H, Zhou C. RFBFN: A Relation-First Blank Filling Network for Joint Relational Triple Extraction. InProceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop 2022 May (pp. 10-20).

[15] Zhu Y, Feng S, Wang D, Zhang Y, Han D. Knowledge-Enhanced Interactive Matching Network for Multi-turn Response Selection in Medical Dialogue Systems. InDatabase Systems for Advanced

Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part III 2022 Apr 8 (pp. 255-262). Cham: Springer International Publishing.

[16] Yang Z, Xu W, Chen R. A deep learning-based multi-turn conversation modeling for diagnostic Q&A document recommendation. Information Processing & Management. 2021 May 1;58(3):102485.

[17] Zeng D, Peng R, Jiang C, Li Y, Dai J. CSDM: A context-sensitive deep matching model for medical dialogue information extraction. Information Sciences. 2022 Aug 1;607:727-38.

[18] Bokaei, Mohammad Hadi, Hossein Sameti, and Yang Liu. "Summarizing meeting transcripts based on functional segmentation." IEEE/ACM Transactions on Audio, Speech, and Language Processing 24, no. 10 (2016): 1831-1841.

[19] Liu W, Tang J, Liang X, Cai Q. Heterogeneous graph reasoning for knowledge-grounded medical dialogue system. Neurocomputing. 2021 Jun 28;442:260-8.

[20] Rani S, Jain A. Optimizing healthcare system by amalgamation of text processing and deep learning: a systematic review. Multimedia Tools and Applications. 2023 May 15:1-25.

[21] Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C.D., Liang, P.S. and Leskovec, J., 2022. Deep bidirectional language-knowledge graph pretraining. Advances in Neural Information Processing Systems, 35, pp.37309-37323.

[22] Wang Y, Li Z, Chen P, Zeng L, Liu A, Xiong N, Huo P, Yu Q. Learning to Embed Knowledge for Medical Dialogue System. InIntelligent Robotics: Third China Annual Conference, CCF CIRAC 2022, Xi'an, China, December 16–18, 2022, Proceedings 2023 Feb 18 (pp. 158-170). Singapore: Springer Nature Singapore.

[23] Abacha, Asma Ben, Wen-wai Yim, Yadan Fan, and Thomas Lin. "An empirical study of clinical note generation from doctor-patient encounters." In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2283-2294. 2023.

[24] Lin S, Zhou P, Liang X, Tang J, Zhao R, Chen Z, Lin L. Graph-evolving meta-learning for low-resource medical dialogue generation. InProceedings of the AAAI Conference on Artificial Intelligence 2021 May 18 (Vol. 35, No. 15, pp. 13362-13370).

[25] Chen W, Li Z, Fang H, Yao Q, Zhong C, Hao J, Zhang Q, Huang X, Peng J, Wei Z. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. Bioinformatics. 2023 Jan;39(1):btac817.

[26] Wang DQ, Feng LY, Ye JG, Zou JG, Zheng YF. Accelerating the integration of ChatGPT and other large‐scale AI models into biomedical research and healthcare. MedComm‐Future Medicine. 2023 Jun;2(2):e43.

[27] Radford Alec, Narasimhan Karthik, Salimans Tim, and Sutskever Ilya. 2018. Improving Language Understanding by Generative Pre-training. Technical Report. OpenAI.

[28] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog. 2019 Feb 24;1(8):9.

[29] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. Advances in neural information processing systems. 2020;33:1877-901.

[30] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. and Schulman, J., 2022. Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, pp.27730-27744.

[31] Biswas SS. Role of chat GPT in public health. Annals of Biomedical Engineering. 2023 Mar 15:1-2.

[32] Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. JMIR Medical Education. 2023 Mar 6;9(1):e46885.

[33] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 320–335, 2022.

[34] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130b: An open bilingual pre-trained model. In The Eleventh International Conference on Learning Representations (ICLR), 2023

[35] Xiong H, Wang S, Zhu Y, Zhao Z, Liu Y, Wang Q, Shen D. DoctorGLM: Fine-tuning your Chinese Doctor is not a Herculean Task. arXiv preprint arXiv:2304.01097. 2023 Apr 3.

[36] Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. InHealthcare 2023 Mar 19 (Vol. 11, No. 6, p. 887). MDPI.

[37] Kenton JD, Toutanova LK. Bert: Pre-training of deep bidirectional transformers for language understanding. InProceedings of naacL-HLT 2019 Jun 2 (Vol. 1, p. 2).

[38] Liu W, Tang J, Cheng Y, Li W, Zheng Y, Liang X. MedDG: An Entity-Centric Medical Consultation Dataset for Entity-Aware Medical Dialogue Generation. InNatural Language Processing and Chinese Computing: 11th CCF International Conference, NLPCC 2022, Guilin, China, September 24–25, 2022, Proceedings, Part I 2022 Sep 24 (pp. 447-459).

[39] Chen B, Cherry C. A systematic comparison of smoothing techniques for sentence-level BLEU. InProceedings of the ninth workshop on statistical machine translation 2014 Jun (pp. 362-367).

[40] Lin, S., Zhou, P., Liang, X., Tang, J., Zhao, R., Chen, Z. and Lin, L., 2021, May. Graph-evolving meta-learning for low-resource medical dialogue generation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 15, pp. 13362-13370).

[41] Zhao Y, Li Y, Wu Y, Hu B, Chen Q, Wang X, Ding Y, Zhang M. Medical Dialogue Response Generation with Pivotal Information Recalling. InProceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2022 Aug 14 (pp. 4763-4771).

[42] Liang, Ke, Sifan Wu, and Jiayi Gu. "MKA: A Scalable Medical Knowledge-Assisted Mechanism for Generative Models on Medical Conversation Tasks." Computational and Mathematical Methods in Medicine 2021 (2021).

[43] Tang C, Zhang H, Loakman T, Lin C, Guerin F. Terminology-aware medical dialogue generation. InICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023 Jun 4 (pp. 1-5). IEEE.

[44] Varshney D, Zafar A, Behera NK, Ekbal A. Knowledge grounded medical dialogue generation using augmented graphs. Scientific Reports. 2023 Feb 27;13(1):3310.

[45] Li, Bin, Encheng Chen, Hongru Liu, Yixuan Weng, Bin Sun, Shutao Li, Yongping Bai, and Meiling Hu. "More but correct: Generating diversified and entity-revised medical response." https://deepai.org/publication/more-but-correct-generating-diversified-and-entity-revised-medical-response. Aug 03, 2021

[46] Li D, Ren Z, Ren P, Chen Z, Fan M, Ma J, de Rijke M. Semi-supervised variational reasoning for medical dialogue generation. InProceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval 2021 Jul 11 (pp. 544-554).

[47] Xiong L, Guo Y, Chen Y, Liang S. How can entities improve the quality of medical dialogue generation?. In2023 2nd International Conference on Big Data, Information and Computer Network (BDICN) 2023 Jan 6 (pp. 225-229). IEEE.

[48] Dou, C., Jin, Z., Jiao, W., Zhao, H., Tao, Z. and Zhao, Y., 2023. Plug-and-Play Medical Dialogue System. arXiv preprint arXiv:2305.11508.