



Process-Aware Dialogue System With Clinical Guideline Knowledge

Meng Wang, School of Computer Science, Wuhan University of Science and Technology, Key Laboratory of Rich-Media Knowledge Organization and Service, Digital Publishing Content, National Press and Publication of the People's Republic of China, China*

 <https://orcid.org/0000-0002-2540-9991>

Feng Gao, School of Computer Science, Wuhan University of Science and Technology, Key Laboratory of Rich-Media Knowledge Organization and Service, Digital Publishing Content, National Press and Publication of the People's Republic of China, China

 <https://orcid.org/0000-0002-2396-1360>

Jinguang Gu, School of Computer Science, Wuhan University of Science and Technology, Key Laboratory of Rich-Media Knowledge Organization and Service, Digital Publishing Content, National Press and Publication of the People's Republic of China, China

ABSTRACT

Task-oriented dialogue systems aim to engage in interactive dialogue with people to ultimately complete specific tasks. Typical application domains include ticket booking, online shopping, and healthcare providing. Medical dialogue systems can interact with patients, provide initial clinical advice, and improve the efficiency and quality of healthcare services. However, current medical dialogue systems lack the ability to utilize domain knowledge. This paper extracts regular domain knowledge as well as medical process knowledge from clinical guidelines to improve the performance of dialogue systems. Regular knowledge is used to generate accurate responses for a given input, and process knowledge is used to steer the conversation. The authors divide the task of multi-turn conversation generation into four sub-tasks and propose a four-layer knowledge-based process-aware dialogue model that incorporates the domain knowledge to generate responses. Results indicate that the approach can lead medical conversations actively while providing accurate responses.

KEYWORDS:

Clinical Guidelines, Generation-Based Models, Hierarchical Model, Representation Learning, Process Knowledge, Retrieval-Based Model, Task-Oriented Dialogue System, Topic Shifting

INTRODUCTION

Task-oriented Dialogue Systems (TDSs) have recently attracted increasing interest. TDSs aim to use human-machine conversations to help users complete specific tasks efficiently. Incorporating deep learning techniques in dialogue systems can significantly increase the accuracy and timeliness of the generated responses (Zhao et al., 2020; Wu et al., 2019). Furthermore, the knowledge-driven dialogue systems using domain knowledge also improve the quality of responses (Zhang et al., 2020; Zhou et al., 2020). Therefore, many researchers are focusing on the impact of domain knowledge on dialogue systems.

DOI: 10.4018/IJWSR.304392

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

In the clinical domain, a dialogue system for clinical diagnosis converses with patients to obtain additional symptoms and make a diagnosis automatically, which has significant potential to simplify the diagnostic procedure and reduce the cost of collecting information from patients (Tang et al., 2016; Wang et al., 2021). More importantly, the clinical guidelines (CGs) documents provide clinical knowledge about diagnostic indicators of disease, pathogenesis, relevant drugs, prognosis and so forth, which are the natural source of knowledge for generating responses in the clinical dialogue system. CGs also provide process-related knowledge, such as how a disease develops or how a treatment plan spans over a period. Such process-related knowledge can be used to steer the conversation towards a specific goal, just like how human doctors control the topic shifts based on their expert knowledge.

Figure 1. Application of process knowledge and triple knowledge in the conversation

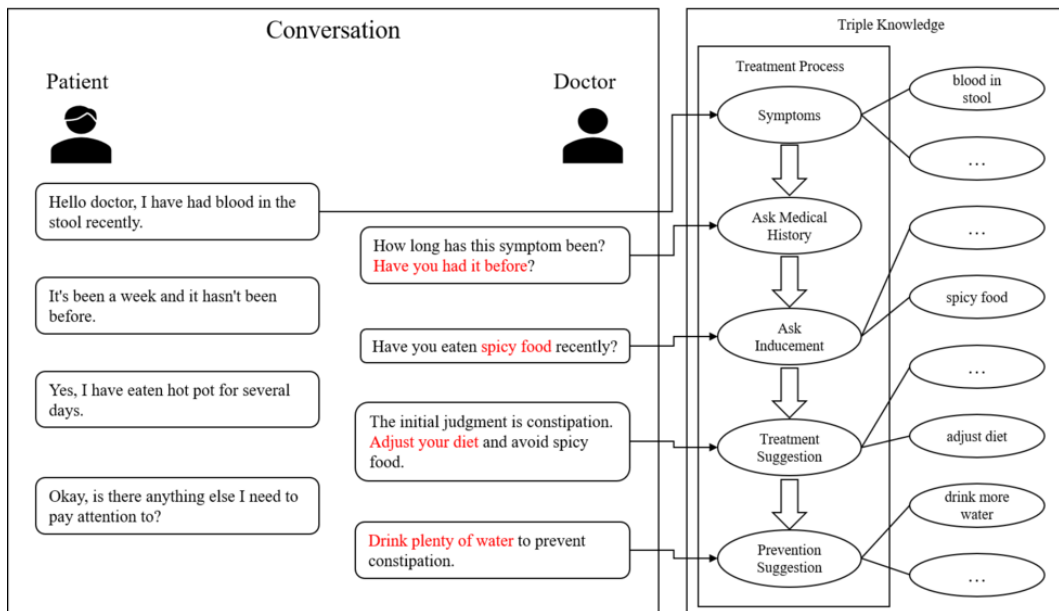


Figure 1 illustrates an exemplifying scenario. The left part shows the conversation, and the right part depicts the relevant process and knowledge grounding. The authors correlate the key part (marked red) to a triple or a topic on the right in the picture. In this example, most parts of the conversations are initiated by the doctor with a question, and it is observable that the questions are chosen based on the process knowledge as well as the answers provided by the patient. Figure 1 shows that the physician always leads the dialogue and controls the topic shifting, such as “Symptoms -> Ask Medical History -> Ask Inducement -> Treatment Suggestion -> Prevention Suggestion.” Therefore, the strategy of topic shifting is of great importance in the consultation task.

Based on the analysis of real-world conversation records and clinical guideline documents, it can be concluded that the treatment procedures are essential references for the topic selection in the consultation process. In addition, reasonable topic selection and utilization of knowledge are essential to the task of clinical consultation; however, current dialogue systems cannot realize this. Thus, it remains a challenge for automated diagnosis to allow the TDS to communicate with the patient as guided by the treatment process. To address this issue, in this paper, the authors first divided the generation task into four sub-tasks, and propose a Process Aware Hierarchical Decision model (PAHD model). The PAHD model leverages regular knowledge to improve the accuracy of responses as well

as process knowledge to control the topic shifting. Towards selecting reasonable topic and triple, the PAHD model is trained by optimizing rewards. Finally, the authors conducted extensive experiments with some benchmark models, including selecting process, topic, triple with RNN- and CNN- based models, generating sentence by generation- and retrieval-based models. Evaluations demonstrate PAHD's effectiveness in terms of conversational coherence and knowledge accuracy, compared to state-of-the-art baselines. The main contributions of the paper are summarized as follows:

- 1 This work divide the task of multi-turn conversation generation into four sub-tasks: treatment process selection, topic selection, triple selection and sentence generation. Following this strategy, authors propose the PAHD model.
- 2 With the help of Clinical Guideline, authors introduce explicit explainable topic shifting policy, which is convenient to design topic related-rewards to optimize planning; they also introduce medical triple to guide response generation for better coherence and informativeness.
- 3 The authors not only assessed the accuracy of the responses, but also proposed some metrics to assess the reasonableness of the responses. Finally, the results of a large number of controlled and ablation experiments show that the method proposed in this paper has a better strategy in topic shifting to actively guide the conversation and obtain more accurate responses.

RELATED WORK

Task-oriented dialogue systems have achieved good results in the general open domain, such as booking (Peng et al., 2018), shopping (Yan et al., 2017), and search (Wen et al., 2017). Task-oriented dialogue system can be implemented in two ways: pipeline- and end-to-end- based models (Chen et al., 2017). The former consists of three main modules: Natural Language Understanding (NLU), Dialogue Management (DM), and Natural Language Generation (NLG) (Zhao et al., 2020); the latter just gets direct responses based on user input. Furthermore, more and more interest is being shown in using knowledge to generate appropriate and informative responses (Ghazvininejad et al., 2018; Zhou et al., 2018; Zhou et al., 2020). This section will introduce the work on knowledge-driven and clinical dialogue system.

Knowledge-Driven Dialogue System

Recently, researchers recognized that the knowledge base is critical to providing accurate responses (Ghazvininejad et al., 2018; Moon et al., 2019; Tuan et al., 2019). Various models use different knowledge annotation, storage, and embedding methods.

Zhou et al. (2018) automatically obtained the knowledge utilized in the conversation, such as while the input contains entities that are the head of the triple and the response contains entities that are the tail of the triple, then the triple is the knowledge grounding of this dialogue. What's more, they used the TransE model to represent the knowledge triple. Finally, they used the Commonsense knowledge aware Conversational Model (CCM) to retrieve relevant knowledge graphs from a knowledge base and then encode the graphs with a static graph attention mechanism, which augments the semantic information of the post. Zhou et al. (2020) provided a Chinese conversation dataset Knowledge-driven Conversation (KdConv) with knowledge annotation by humans and a structured knowledge graph. They validated, through extensive experiments, that the knowledge grounding has significantly contributed to improving the accuracy of sentences. Xu et al. (2020) found that previous neural models of open domain conversation generation did not have effective mechanisms to control the topic of the chat, which tended to generate poorly coherent conversations.

The aforementioned approaches have some limitations (Wang et al, 2021). The approach of Zhou et al. (2020) assumes a high degree of overlap between conversation content and knowledge, which lacks variability. Xu et al. (2020) considered topic shifting as moving from one entity to another in the knowledge graph, however, which is not applicable to task-oriented dialogue systems.

Clinical Task-Oriented Dialogue System

Previous research has mainly focused on individual modules in pipeline-based models. For example, Wei et al. (2018) attempted to address automatic diagnosis issues using the Deep Q-Network. These works facilitated the development of technologies in the medical dialogue domain. Xu et al. (2019) proposed a novel knowledge-routed Deep Q-network (KR-DQN) and promoted the performance of the DM module. They improved the rationality of decision-making for medical dialogue, which incorporates external probabilistic symptoms related to the framework of reinforcement learning.

For end-to-end implementations, Zeng et al. (2020) built large-scale medical dialogue datasets. (MedDialog), and pre-trained several dialogue generation models, including Transformer, GPT, and BERT-GPT, and they also studied the transferability of models trained on MedDialog to low-resource medical dialogue generation tasks. Liu et al. (2019) designed a dialogue comprehension system and proposed a framework inspired by nurse initiated clinical symptom monitoring conversations to construct a simulated human–human dialogue dataset. They constructed some templates and developed strategies for template selection to improve the effectiveness of the dialogue system, which are designed for sentence generation (e.g., yes/no response, detailed response). However, all of them ignored the significance of other important information such as the attributes of the symptom, tests, and medicine. Liu et al. (2020) also built and released a large-scale, high-quality medical dialogue dataset, annotated by five entities and related to 12 types of common gastrointestinal diseases, named MedDG. Moreover, they divided the sentence-generation task into two sub-tasks: topic prediction and sentence generation. However, a weakness with these methods is that all of them ignored the domain knowledge-CGs. Specifically, they ignored the diagnosis and treatment process.

To address these issues, this paper proposes a new neural model that uses the treatment processes in clinical guideline documents to guide the conversation as well as to improve the accuracy of sentence generation using conventional medical knowledge.

CLINICAL GUIDELINES-BASED DIALOGUE SYSTEM

CG documents are the gold standard in evidence-based medicine and are the definitive guidance documents for diagnosis. In a CG document, knowledge of a disease’s diagnostic and therapeutic process is usually presented in a flow chart, which is defined as **Process Knowledge** in this paper. A process contains a list of topics, such as “symptom -> medical history -> inducement -> treatment -> prevention.” It represents the process of topic shifting during the conversation, illustrated by Figure 1. Furthermore, the process is not constant. All the branch processes can be obtained from the CGs’ flow chart. In addition, **Triple Knowledge** can be acquired by annotating the CGs file, which can represent medical background knowledge. For example, “Colonoscopy is a method of checking for constipation” can be represented by colonoscopy, type, inspection method.

Here are explicit definitions and examples of some terms:

Entity: An entity is an individual with practical significance in CG. Such as *lactulose*, *blood in the stool*.

Topic: A topic is a word with an abstract meaning that is extracted by classifying all the entities. Such as *symptom*.

Process Knowledge: A process is an ordered list containing more than one topic. Such as “*symptom -> treatment -> prevention*.”

Triple Knowledge: A triple contains three items: subject, relation, and object, in which subject and object are composed of entities, subject represents head node, and object represents tail node. Such as (colonoscopy, type, inspection method).

Conventional dialogue systems generate response statements based on the input text. Some models are only fed with the patient's statement; some models are fed with all historical dialogue utterances, which can capture contextual textual information and contextual information. However, sentences generated based on historical conversational information alone are not interpretable and have low accuracy. Zhou et al. (2020) proposed a knowledge-driven model, which adds the triple knowledge satisfied by the sentences to the input texts. This approach improved the expressiveness but still failed to achieve the effect of guided dialogue, which is far from sufficient for task-oriented dialogue systems in the clinical domain.

Inspired by previous works (Xu et al., 2020; Zhou et al., 2020; Zeng et al., 2020), this paper divides the dialogue generation task into four sub-tasks to utilize process and triple knowledge to enhance the performance of task-oriented dialogue system: 1) select the appropriate process for the conversation based on historical sentences and topics; 2) select the topic for doctor/agent based on the process and historical sentences and topics; 3) select triple knowledge based on historical sentences and the topics; 4) generate the sentence based on historical sentences and triple knowledge.

PROBLEM DEFINITION

This paper proposed a four-layer, knowledge-based PAHD model according to the four sub-tasks mentioned above. There are three selection layers and a generation layer, such as Process Selection (PS) layer, Topic Selection (ToS) layer, Triple Selection (TrS) layer, and Sentence Generation (SG) layer. Detailed symbol definitions are as follows:

Formally, the collection of all process knowledge extracted from clinical guidelines is $S_{Process} = \{Process_0, \dots, Process_n\}$. $S_{Topic} = \{Topic_0, \dots, Topic_n\}$ represents the collection of all topics. The term $S_{Triple} = \{Triple_0, \dots, Triple_n\}$ refers to the collection of all medical triple knowledge.

Set t moment as the sentence generation moment. $H_{Sentence} = \{Sentence_0, \dots, Sentence_{t-1}\}$ represents the historical sentences, whose vector-matrix can be denoted by W_{hs} ; $H_{Topic} = \{Topic_0, \dots, Topic_{t-1}\}$ denotes the historical topics, whose vector-matrix can be symbolized by W_{hto} .

At moment t , the selection result of the PS layer is $Process_t$, and its vector-matrix is W_{pt} ; the result of the ToS layer is $Topic_t$, and its vector-matrix is W_{to_t} ; the result of the TrS layer is $Triple_t$, and its vector-matrix is W_{tr_t} ; the result of the SG layer is $Sentence_t$, and its vector-matrix is W_{st} . The four sub-tasks can be expressed using the following equations (1)-(4).

$$Process_t = f_{ps}(W_{hs}, W_{hto}), Process_t \in S_{Process} \quad (1)$$

$$Topic_t = f_{tos}(W_{hs}, W_{pt}, W_{hto}), Topic_t \in S_{Topic} \quad (2)$$

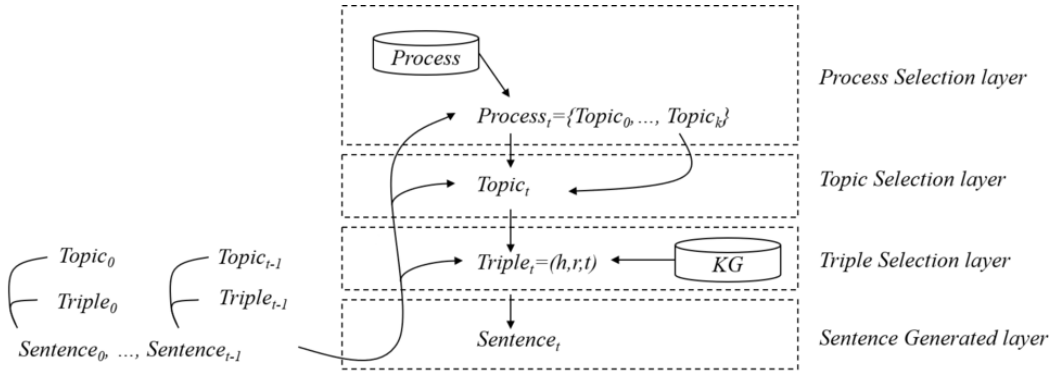
$$Triple_t = f_{trs}(W_{to_t}, W_{hs}), Triple_t \in S_{Triple} \quad (3)$$

$$Sentence_t = f_{sg} \left(W_{tr}, W_{hs} \right) \quad (4)$$

PAHD MODEL

The PAHD is composed of a four-layer neural model. Figure 2 gives the overall structure of it and the flow of data. More specific details will be discussed as follows.

Figure 2. Overall structure

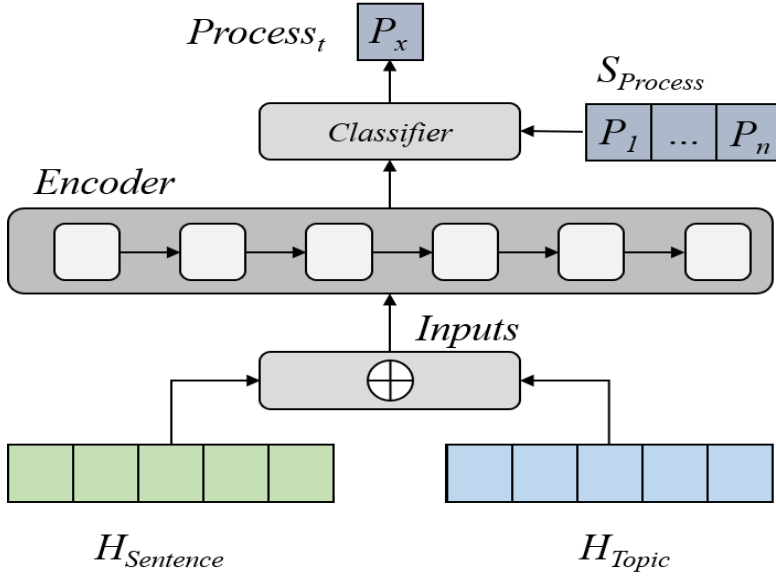


PROCESS SELECTION

Model Design

The process selection model is made up of a neural dialogue encoder in conjunction with a single-layer classifier. It aims to choose the appropriate process for the conversation based on the input information. The selection result is taken from $S_{Process}$. The inputs of the encoder are historical sentences $H_{Sentence}$ and historical topics H_{Topic} , and the output, a single-dimensional context vector $Hidden_{SP}$, represents a summary of the dialogue history. The authors then input $Hidden_{SP}$ into classifier to obtain the most probable process $Process_i$. We have conducted experiments using different encoders such as LSTM (Wang & Nyberg, 2015), GRU (Chung et al., 2014), and TextCNN (Kim, 2014). It is shown in Figure 3, and can be defined by Equation (5).

Figure 3. PS layer



$$Process_t = \text{softmax}\left(\frac{\exp((I_t^{ps})^T v_{pt})}{\sum_{i=1}^{NP} \exp((I_t^{ps})^T v_{pi})}\right) \quad (5)$$

Where $I_t^{ps} = W_{ps} = [W_{hs}, W_{hto}]$ is the input of the PS layer, and $v_{pt} = [W_{pt}]$ denotes the result of PS layer. The amount of $S_{Process}$ collection is NP . This is a classification task, and the loss function can be defined $Loss = -\sum_{i=1}^{NP} y_i \cdot \log \hat{y}_i$, where \hat{y}_i is the i -th scalar value in the model output, and y_i is the corresponding target value.

Rewards & Metrics for PS

It is necessary to consider whether the process chosen is correct or reasonable. Therefore, this paper not only uses accuracy as a metric for evaluating PS models but also the metrics of consistency of purpose and consistency of process. Correspondingly, if the result is reasonable, the model will be rewarded.

Purpose Consistency (PuC). Processes can be divided into groups by consistent purpose. For example, $Process_i = [Symptom, MedicalHistory, Treatment]$ and

$Process_j = [Symptoms, MedicalHistory, Inducement, Treatment]$ belongs to the same group as their final goal is Treatment. If the selection result belongs to the group with the correct answer, the score is 1.

Process Consistency (PrC). There are many branches in the flowchart, so there may be more than one corresponding process when the conversation is not finished. Thus, it is necessary to compare the effective topics in the process to calculate their consistency, which means they are shown before the t moment. First, set $P_{correct} = \{Topic_0, \dots, Topic_n\}$ denotes the correct process of the conversation; $P_{effective_t} = \{Topic_0, \dots, Topic_x\}$ denotes the effective child process at each t moment of $P_{correct}$, and

$P_{effective_t} \subseteq P_{correct}$. Choose the first x topics from $Process_t$ as $P_{t_x} = \{Topic_0, \dots, Topic_x\}$. If $P_{t_x} = P_{effective_t}$, the score is 1, otherwise, it is 0.

The values of PuC and PrC are used as reward factors in the training process for the PS layer model. During the training process, the loss function of the model was modified so that the selection results satisfying the PuC and PrC cases received smaller loss values, and the modified loss function is shown in Equation (6), where $PuC = 0$ or 1 and $PrC = 0$ or 1 .

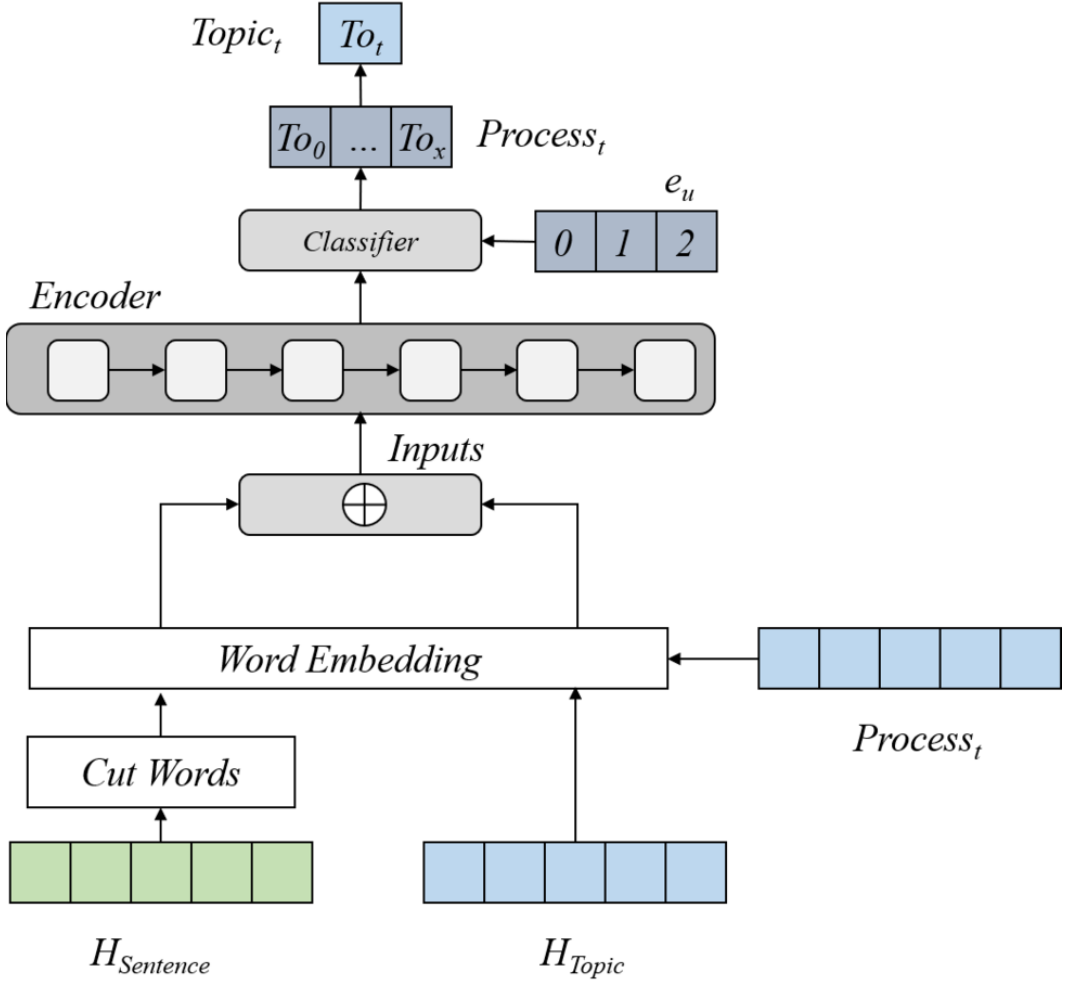
$$Loss = -\sum_{i=1}^N y_i * \log \hat{y}_i - \mu * PuC - \lambda * PrC \quad (6)$$

Topic Selection

Although Liu et al. (2020) attempted to select suitable dialogue entities based on historical dialogue tasks, their approach was to select suitable entities from a large entity pool and failed to achieve a high index in terms of accuracy. In addition, due to the lack of guidance from process knowledge, the selected topic entities failed to achieve the role of guiding the dialogue when the text features were not obvious. For example, when there are a large number of stop words or meaningless sentences, it is often difficult to select a reasonable topic. However, in consultation tasks, the shifting of topics is determined by the consultation process in the CG document, and it is more likely that the patient will only respond “yes” or “no” during the consultation task. Therefore, this paper restricts the selection of the topic to the process, which makes the selection of the topic more accurate and meaningful on the one hand, and enables the model to play a guiding role in the dialogue on the other.

At the ToS layer, the task objective is to select a suitable topic. Similarly, the ToS layer has the same structure as the PS layer. The difference is that it does not directly select the topic with the highest probability from the topic dataset, but it selects the topic from $Process_t$. $t_u = 0 / 1 / 2$ indicates that the goal topic is *None*, unchanged topic ($Topic_x$), or the following topic ($Topic_{x+1}$) in the process, where $Topic_x$ means the last effective topic. Briefly, at this layer, the usual topic prediction task is turned into a classification problem. Constraining the selection of topic in a reasonable treatment process can improve the hits of topics. Figure 4 illustrates the ToS layer, and it can be defined by Equation (7).

Figure 4. ToS layer



$$Result_{ToS} = softmax(\frac{\exp((I_t^{tos})^T v_{tot})}{\sum_{i=1}^{NTo} \exp((I_t^{tos})^T v_{toi})})$$

$$Topic_t = \begin{cases} None, & \text{if } Result_{ToS} = 0 \\ Topic_x, & \text{if } Result_{ToS} = 1 \\ Topic_{x+1}, & \text{if } Result_{ToS} = 2 \end{cases} \quad (7)$$

Where $I_t^{tos} = W_{tos} = [W_{hs}; W_{hto}; v_{pt}]$ means the input of ToS layer. $v_{tot} = [t_u]$, and $NTo = 3$. This is also a multi-label classification task, thus its loss is similar to the PS's.

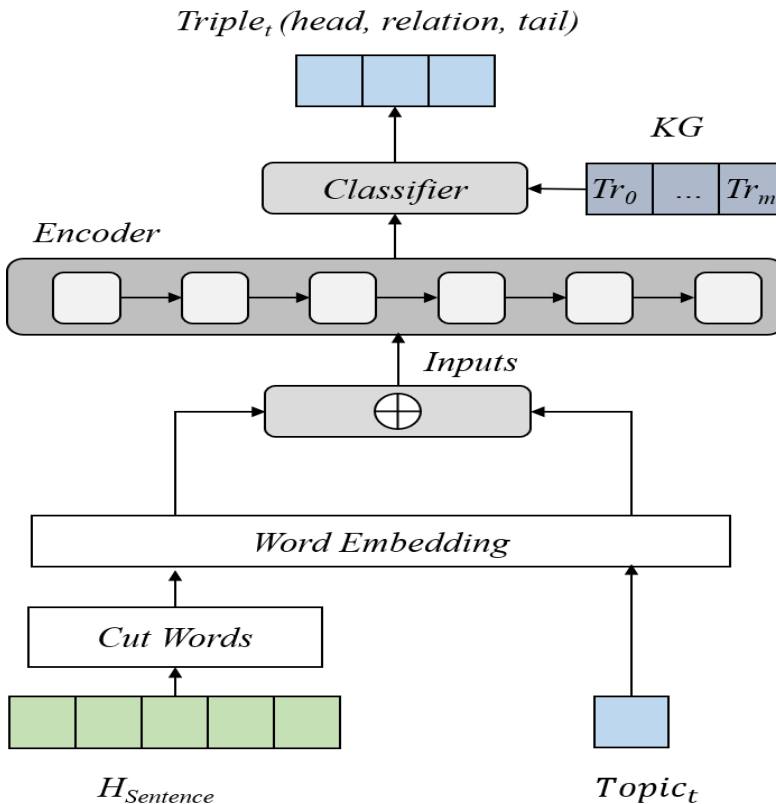
TRIPLE SELECTION

Model Design

Though the topic words can generate sentences closely aligned to the topic, there is still a need for improvement in the quality of sentence generation. After analyzing the dataset, it can be found that even if the topics are the same, there may be differences in the content of the conversation, mainly due to the inconsistent triple knowledge the doctor uses about the topic. For example, the same topic *Medication* sometimes requires *Corkage* and sometimes *Laxative*, which also leads to inconsistent response statements. The triple selection layer (TrS) is designed to address this problem, with a goal of selecting the appropriate triple $Triple_t$ from the triple knowledge base S_{Triple} , according to the historical dialogue statements $H_{Sentence}$ and the topic $Topic_t$. Ideally, the selection triple's head node or tail node is the topic $Topic_t$, so this paper designs the reward mechanism for the TrS layer model.

At the TrS layer, the task objective is to select the latent triple knowledge, which head node or tail node is $Topic_t$. The structure is also composed of an encoder and a classifier, like the PS layer. Figure 5 illustrates the structure, which can be defined by Equation (8).

Figure 5. TrS layer



$$Triple_t = softmax(\frac{\exp((I_t^{trs})^T v_{trt})}{\sum_{i=1}^{NTr} \exp((I_t^{trs})^T v_{tri})}) \quad (8)$$

Where $I_t^{trs} = [W_{hs}; W_{tot}]$ denotes the input of TrS layer, and where $v_{trt} = [W_{tri}]$, NTr denotes the quantity of triple knowledge. And the loss function is similar to the PS's.

Rewards & Metrics for TrS Layer

It is necessary to consider the reasonableness of the triple selection result. Therefore, this paper not only uses accuracy for evaluation but also uses a novel evaluation indicator for judging triple selection reasonableness.

Hit Head or Tail of Triple (HTT). While the head or tail of $Triple_t$ is equal to $Topic_t$, the score is 1.

The loss function is redefined as Equation (9).

$$Loss = -\sum_{i=1}^N y_i * \log \hat{y}_i + \delta * HTT \quad (9)$$

Sentence Generation

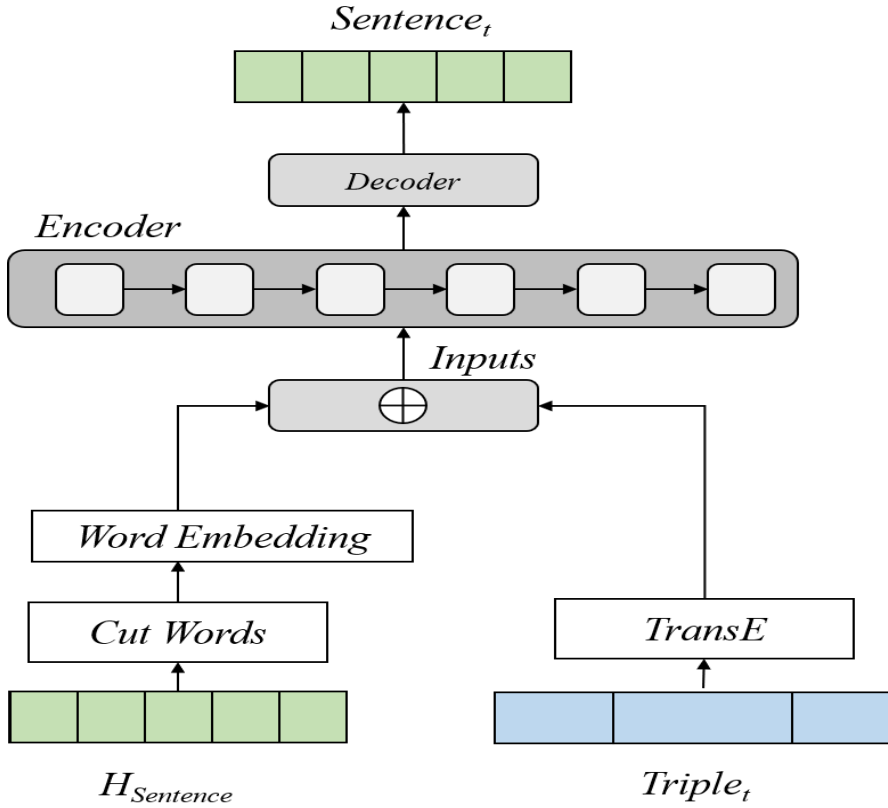
The generation-based model concatenates the triple's embedding from TransE and the historical sentences as new input text of Seq2Seq and HRED. The retrieval-based model selects a sentence from candidate collection by the triple-sentences dict, in which key is a triple and value are a set of candidate sentences.

For a retrieval-based model, the training task is to predict whether a candidate is the correct response. For a generation-based model, the training task is to directly generate sentence by decoder.

Figure 6 illustrates the SG layer and is defined by Equation (10), where $I_t^{sg} = [W_{hs}; W_{tri}]$ denotes the input of the SG layer. And the generation loss is the average of negative log likelihood of the target

sequence $\{y_t^*\} (1 \leq t \leq T)$: $Loss = \frac{1}{T} \sum_{t=1}^T \log P(y_t^*)$.

Figure 6. SG layer



$$Sentence_t = softmax(\exp((I_t^{sg})^T)) \quad (10)$$

EXPERIMENT

This paper chose constipation disease for the experiment, obtained the clinical guideline document on chronic constipation from the Chinese Medical Association organization, and collected the dataset of the CCKS 2021 Chinese medical conversation generation competition with embedded entities¹. Moreover, the datasets are filtered and only retained data related to constipation disorders. This paper evaluates both generation- and retrieval-based models on the corpus to validate the approach. The experimental part analyzes the performance of different models, the error transmission of hierarchical models, and the influence of process knowledge and triple knowledge.

DATASET

The dataset includes a CG for constipation disease and patient–doctor conversations. Primarily, five annotators were involved in the annotation process to obtain the process knowledge and triple knowledge from the CGs. They first discussed and designed a knowledge graph of constipation. Then summarize the diagnosis and treatment processes in all branches. Initially, hundreds of processes are

obtained based on the flowchart. Finally, it can be found that only about 70 processes were commonly used while annotating the dialogues. Thus, those processes rarely utilized are discarded. Table 1 shows some statistics.

Table 1. Data statistics

	Amount	Example
Class	42	Symptom, Crowd, Medicine
ObjectProperty	8	usingMedicine, type
DataProperty	5	Cautions
Individual	150+	Lactulose, BloodInStool
Triple	700+	(Medication, usingMedicine, Lactulose)
Process	70	[Symptom, Medical History, Inducement, Treatment, Prevention]
Topic	32	Symptom

As the existing dialogue dataset lacked process annotation and triple knowledge annotation, four annotators trained in constipation disorders were invited to complete the annotation task. The steps are: 1) mark the response statements of doctors with triple knowledge; 2) mark one most important topic of each sentence; and 3) mark the process of the whole conversation. Table 2 summarizes the data statistics. The proportion of train data is 80%, and test is 20%.

Table 2. Corpus statistics

	Group	Sentence	Turn	Process len	Sentence with to	Sentence with tr
Total	2.1K	35K	17K	-	18.3K	7.1K
Avg(/g)	-	16.6	8.4	5.35	8.7	3.3

"/g" means per group; "with to" means with topic; "with tr" means with triple knowledge.

AUTOMATIC EVALUATION

This paper uses several custom metrics to evaluate the quality of the generated response, in addition to frequently used evaluation metrics. For the PS layer, the authors use Acc_p (Accuracy of Process), PuC , PrC to validate the selection results of the model. For the ToS layer, the authors use the metrics P_t , R_t , and $F1_t$. For TrS layer, the authors utilize the Acc_t (Accuracy of Triple) and HTT_t metrics to validate the outcomes. $BLEU - 1 / 2 / 3 / 4$ (Papineni et al., 2002) is a popular metric to compute the k -gram overlap between a generated sentence and a reference. $Distinct - 1 / 2 / 3 / 4$ (Li et al., 2016) is also provided to evaluate the diversity of generated responses. *Perplexity* (Meister & Cotterell, 2021) is also commonly used to evaluate the results of language models. All three are used to evaluate the effectiveness of the utterances generated by the SG layer.

IMPLEMENTATION DETAILS

Both the word embedding and hidden dimensions are set to 300. The authors use the Adam optimizer, with a minibatch size of 16, and set the initial learning rate to 0.001. The authors use BERT as a search-based model, and follow the configurations in the origin paper. The researchers train the model with Tesla P100-PCIE graphic cards, Intel(R)Xeon(R) CPU E5-2660 v4 @ 2.00GHz CPU, and 64G memory.

AUTOMATIC EVALUATION

PS Layer Results

Firstly, the models of the PS layer were analyzed using two control datasets, one with additional knowledge (historical topics) introduced, which were additionally marked with “+HTo”, and one without any additional information added. The result is shown in Table 3.

Table 3. PS layer results

Model	Acc_p	PuC	PrC
LSTM	27.13	55.71	67.44
GRU	28.44	58.12	68.55
TextCNN	31.27	61.34	70.12
LSTM+HTo	58.55	75.32	86.94
GRU+HTo	66.73	85.49	89.07
TextCNN+HTo	69.77	84.28	92.17

According to the results in Table 3, it can be seen that the TextCNN model has the best performance among all the baseline models. All the baseline models were improved by the additional historical topics, and the TextCNN model still performed best. The performance capability of each baseline model was improved because the model learned feature information of historical topics during the training process, and the historical topics were closely related to the process. However, for the Acc_p metric, this paper has conducted extensive analysis of the experimental data and results. After analysis, it could be found that when a large number of “yes,” “no,” “don’t know,” and other meaningless phrases appeared in the historical conversation, the selection task tends to become more difficult and extremely easy to interfere with the process selection task. However, the results are still more desirable for the rationality of the process (PuC , PrC). It remains evident from Table 3 that these models perform slightly less efficiently than PrC in terms of PuC due to the historical dialogue and historical topics containing a large number of features that maintain process consistency and less obvious features of purpose consistency.

ToS Layer Results

Similar to the PS layer, the model that introduced additional knowledge (process knowledge) was marked as “+P.” The difference is that the model without process knowledge directly selects candidate

topic from the topic database, and the input text information only contains historical dialogue data without historical topic information. The result is shown in Table 4.

Table 4. ToS layer results

Model	P_t	R_t	$F1_t$
LSTM	75.14	70.15	72.59
GRU	77.33	71.48	74.29
TextCNN	76.18	75.47	75.82
LSTM+P	85.19	85.17	85.09
GRU+P	90.80	88.83	89.78
TextCNN+P	89.88	91.15	90.51

According to the results in Table 4, it can be seen that among all the baseline models, the TextCNN model and the GRU model have comparable performance capabilities, and both outperform the LSTM model. Secondly, all the baseline models received a large improvement after inputting the features of the process text to the classifier, among which the GRU model excelled in P_t , and the TextCNN model excelled in R_t and $F1_t$ metrics. The experimental results table shows that restricting the candidate set of topics to process knowledge significantly improves the accuracy of topic prediction.

TrS Layer Results

For the TrS layer, two sets of controlled experiments were set up. Similar to the PS layer, the model with additional knowledge (topic) was marked as “+To,” while the input text of the model without additional knowledge was only historical dialogue data, without historical topics and topic, and the experimental results are shown in Table 5.

Table 5. TrS layer results

Model	Acc_t	HHT_t
LSTM	47.31	54.17
GRU	47.67	58.11
TextCNN	48.79	59.04
LSTM+To	72.76	86.17
GRU+To	70.58	89.44
TextCNN+To	79.69	90.47

According to Table 5, all baselines have poorer results and lower metric values. After introducing additional knowledge (topic), all baseline models were improved in both and improved in both Acc_t and HTT_t , and HTT_t improved more. After analyzing the results, it can be concluded that the models learn the feature values of the topic from the input information, which is easier for filtering out some triples. However, it is more difficult for the model to achieve complete matching. By analyzing the data and the results, some of the triples are used less frequently or even appear only once, such as $(Medication, useOfMedication, X)$, where X is a specific infrequently used medication.

SG Layer Results

For the SG layer, two sets of data are designed in this paper for comparison experiments. One group contains only historical dialogue information, while the other group introduces additional ternary knowledge information (marked as “+Tr”). The results of the experiments are shown in Table 6. Since the sentences of the BERT model are from a human corpus, calculating perplexity is not meaningful and, therefore, not shown. The authors also compared with the existing, more advanced model CCM (Zhou et al., 2018) and the methods of Zhou et al. (marked as KdConv) (2020).

Table 6. SG layer results

Model	$BLEU - 1 / 2 / 3 / 4$				$Distinct - 1 / 2 / 3 / 4$				$Perplexity$	
HRED	14.53	6.58	3.65	2.17	2.13	7.68	15.12	22.81	53.92	
Seq2Seq	14.46	6.67	3.61	2.11	2.43	9.06	17.87	27.21	37.94	
CCM	15.22	7.47	4.15	2.87	2.74	9.17	18.44	27.58	44.85	
BERT	65.04	58.44	54.81	52.51	3.67	26.30	46.56	54.59	-	
KdConv-BERT	65.25	59.17	55.33	54.64	3.99	26.79	47.03	54.67	-	
HRED+Tr	17.43	8.12	5.23	3.97	2.97	9.12	18.74	27.64	48.11	
Seq2Seq+Tr	17.34	8.54	5.17	3.48	3.46	12.74	19.37	29.23	34.29	
BERT+Tr	67.19	61.03	56.12	54.77	4.91	27.88	48.15	54.89	-	

As shown in Table 6, the performance capability of all the models was improved with the introduction of additional triad knowledge. Among the baseline models, the BERT model performed the best; in other words, the retrieval-based model outperformed the generation-based model. However, the improvement for the BERT model was smaller for all models using the methods in this paper, as the difference in features between individual sentences and triadic knowledge in the shallow network of BERT was not significant, so the improvement was not significant for the BERT model. The generation-based model improved in both $BLEU - k$, that is, the $Distinct - k$ metrics, after using the methods in this paper.

In addition, according to Table 6, the method in this paper can generate more accurate and diverse responses than other state-of-the-art models. For example, Seq2Seq+Tr and HRED+Tr perform better than CCM; BERT+Tr performs better than KdConv-BERT.

Error Transmission Analysis

The effect of error transfer in the two methods is investigated separately, where the model using hierarchical decision making is labelled “+PA (Pross Aware).” The results of the experiments are shown in Table 7.

Table 7. Joint selection experiment results

Model	P_t	R_t	$F1_t$	Acc_t	HTT_t
LSTM+PA	74.29	81.54	77.74	58.29	71.23
GRU+PA	85.35	79.59	82.37	60.35	75.14
TextCNN+PA	83.07	86.76	84.88	60.62	76.43

As can be seen in Table 7, for all baseline models, there is a decrease in all metrics, but still higher than the original model. The TextCNN model has the least impact and the least degradation in performance. After analyzing the data, it was found that the selection of topic was very difficult when stop words or meaningless sentences such as “I don’t know” and “I don’t remember” were present in the historical conversation messages. Although good results were achieved in terms of process consistency or purpose consistency, the topic at the next step may not be consistent, thus leading to reduced decision-making ability at the TrS layer when selecting triple knowledge. The experimental results show that error transfer is present, but the results are still better than the baseline model.

MANUAL EVALUATION

Three annotators were invited to perform a human evaluation of the generated response statements. There are two metrics to evaluate the response, which are as follows:

Fluency is defined as a metric to assess the fluency and naturalness of sentences. If there are grammatical errors or the sentences are poorly stated and difficult to understand, mark 0. If there are grammatical errors but the sentences are easy to understand, mark 1. If sentences are fluent and plausibly produced by a human, mark 2.

Coherence is defined as a metric to assess the relevance and reasonableness of sentences. If it is irrelevant to the context, mark 0. If it is relevant to the last sentence, but not the whole history, mark 1. If it is both relevant and reasonable, mark 2.









For the human measures *Fluency* and *Coherence*, 40% of the data generated by the model was randomly selected for manual evaluation, and the results are shown in Table 8. The hierarchical decision model is labelled as “+PA (Pross Aware).”

Table 8. Manual evaluation results

Model	Fluency	Coherence
HRED	0.87	0.51
Seq2Seq	1.03	0.67
BERT	1.67	1.06
HRED+PA	1.05	1.05
Seq2Seq+PA	1.17	1.18
BERT+PA	1.68	1.46

As can be seen from the table, the BERT model outperformed the other models in these two metrics. Because the candidate utterances are drawn from a human corpus, the retrieval-based model scores higher in terms of fluency. However, in terms of plausibility, it performed imperfectly and therefore could not receive a full score. In addition, all the original models had low coherence scores, with HRED and Seq2Seq scoring below 1.0 and the BERT model scoring slightly above 1.0. For the original models, if only historical dialogue data were considered, they would face a loss of contextual coherence as the number of dialogue rounds increased and historical dialogue information grew. However, due to the introduction of process and triadic knowledge, the models based on this paper’s approach are improved in both of these aspects.

Figure 7. Example of each model

Patient	Doctor	HRED	Seq2Seq	BERT	Seq2Seq+PA	HRED+PA	BERT+PA
							
I usually eat vegetables and fruit and drink water but I don't have a bowel movement.	How long have you had these symptoms?	Hello, how long has this been going on?	What's wrong with eating?	How long?	Have you had it before?	How long has this Have you had it before?	Have you had before?
3, 4 days	Do you usually eat a lot?	That's OK.	No.	eat something that is well digested.	spicy, irritating food food?	Do you eat spicy foods?	It has to do with your diet.
I will eat a lot at meal times and usually have a small snack if I'm hungry.	You should adjust your diet and relax your mind.	Do not eat spicy and stimulating food	You can take some laxatives.	Adjust your diet and relax your mind.	take care of your diet and adjust.	Eating regularity regularity.	Adjust your diet and relax your mind.

ABLATION STUDY

In the ablation experiments, for the PAHD model, the experimental group that introduced Process + Topic is labelled “+PT (Process & Topic),” and the experimental group that introduced triple knowledge is labelled “+OT (Only Triple).” The results of the experiments are shown in Table 9. The BERT model is not discussed further as the improvements to the BERT model are not significant.

Table 9. Ablation study results

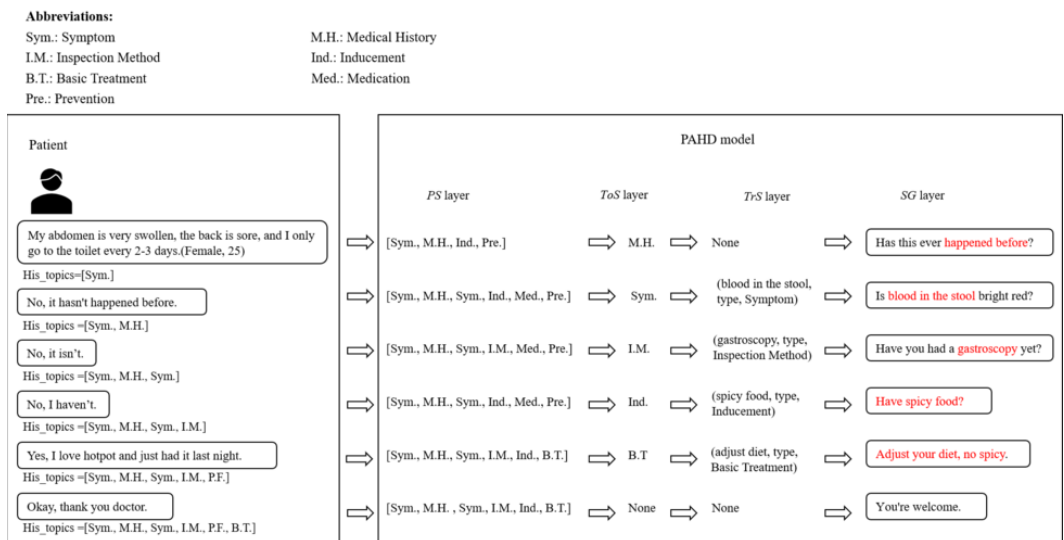
Model	$BLEU - 1 / 2 / 3 / 4$				$Distinct - 1 / 2 / 3 / 4$				Perplexity
HRED+PT	16.74	7.65	4.34	2.64	2.05	7.37	16.87	25.14	51.24
Seq2Seq+PT	16.32	7.87	4.22	2.87	2.47	10.34	17.61	27.65	35.11
HRED+OT	15.97	6.23	3.17	2.18	2.53	8.23	17.54	26.58	50.89
Seq2Seq+OT	15.01	6.12	3.13	2.06	3.22	11.64	18.39	28.46	35.23

According to Table 9, models that introduce process knowledge and make selections about topic (+PT) perform better in terms of the $BLEU - k$ metric. Analysis of the experimental results revealed that such models had good decision-making capabilities in terms of topic shifting and consequently improved the accuracy and fluency of the generated utterances. The model that introduced only triple knowledge (+OT) outperformed the other models and the baseline model in terms of $Distinct - k$, that is, sentence diversity. The analysis of the experimental results showed that the sentences generated by introducing triple knowledge were more diverse. In summary, additional process knowledge and triple knowledge can improve the model's performance in generating sentences.

CASE STUDY

Figure 8 illustrates the details. As the names of the topic mentioned in the process are too long, it uses their abbreviations and provides a reference.

Figure 8. Case study



Most processes are started with *Symptom* topic. For the first turn, the selection result of the PS layer is $Process_1$, though overall, it does not match the process of the conversation going on later.

However, the $P_{effective} = [Symptom]$ is appropriate for the information obtained from the first round of dialogue. At the second turn, the “patient” uses the stop word “no.” Therefore, the system needs to lead the conversation. Thus, the model chooses the appropriate diagnosis process and selects the topic from it so that the conversation can continue with purpose. It also can be found that the process selection results of each turn are not consistent; however, their $P_{effective}$ is approximately consistent. Nevertheless, there exist some mistakes. For example, the $P_{effective}$ of the result decided of the fourth turn is an error: it omitted the *InspectionMethod* topic. What’s more, the sentences generated by models do not conform to human speech, such as the generation of the fifth turn.

CONCLUSION

In this paper, the authors present a new approach to improving the performance of dialogue systems in the clinical domain using CG documents. Firstly, this paper extracted regular domain knowledge as well as medical process knowledge from CGs. Then, this paper divided the sentence-generation task into four sub-tasks and proposed the PAHD model. In this paper, researchers conducted extensive experiments in which they implemented some baseline models and designed some appropriate metrics to measure their selection outcomes. The experimental results show that the approach proposed in this paper has a better strategy in topic-shifting control to guide clinical consultation conversations actively and provide accurate responses.

However, there still exist some limitations: 1) there are few medical conversation datasets with annotation information. and 2) a multimodal data-based dialogue system can better help physicians diagnose. Therefore, the authors will work in the future to make the dialogue model more intelligent.

CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

FUNDING AGENCY

This work is supported by the National Natural Science Foundation of China [No. U1836118], the Open Fund of Key Laboratory of Content Organization and Knowledge Services for Rich Media Digital Publishing [ZD2021-11/01].

The Open Access Processing fee for this article was covered in full by the authors.

REFERENCES

- Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explorations*, 19(2), 25–35. doi:10.1145/3166054.3166058
- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling*. arXiv preprint arXiv:1412.3555.
- Ghazvininejad, M., Brockett, C., Chang, M. W., Dolan, B., Gao, J., Yih, W. T., & Galley, M. (2018, April). A knowledge-grounded neural conversation model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 14006–14014.
- Kim, Y. (2014, October). Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751). Doha: Association for Computational Linguistics. doi:10.3115/v1/D14-1181
- Li, J., Galley, M., Brockett, C., Gao, J., & Dolan, W. B. (2016, June). A diversity-promoting objective function for neural conversation models. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. doi:10.18653/v1/N16-1014
- Lin, X., He, X., Chen, Q., Tou, H., Wei, Z., & Chen, T. (2019, November). Enhancing dialogue symptom diagnosis with global attention and symptom graph. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5033–5042. doi:10.18653/v1/D19-1508
- Liu, W., Tang, J., Qin, J., Xu, L., Li, Z., & Liang, X. (2020). *Meddg: A large-scale medical consultation dataset for building medical dialogue system*. arXiv preprint arXiv:2010.07497.
- Liu, Z., Lim, H., Suhaimi, N. F. A., Tong, S. C., Ong, S., Ng, A., Lee, S., Macdonald, M. R., Ramasamy, S., Krishnaswamy, P., Chow, W. L., & Chen, N. (2019, June). Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2, 24–31. doi:10.18653/v1/N19-2004
- Meister, C., & Cotterell, R. (2021, August). Language model evaluation beyond perplexity. 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. 5328–5339). Academic Press.
- Moon, S., Shah, P., Kumar, A., & Subba, R. (2019, July). Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 845–854. doi:10.18653/v1/P19-1081
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- Peng, B., Li, X., Gao, J., Liu, J., Chen, Y. N., & Wong, K. F. (2018, April). Adversarial advantage actor-critic model for task-completion dialogue policy learning. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6149–6153. doi:10.1109/ICASSP.2018.8461918
- Tang, K. F., Kao, H. C., Chou, C. N., & Chang, E. Y. (2016, December). Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning. *Proceedings of NIPS Workshop on Deep Reinforcement Learning*.
- Tuan, Y. L., Chen, Y. N., & Lee, H. Y. (2019, November). DyKgChat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1855–1865. doi:10.18653/v1/D19-1194
- Wang, D., & Nyberg, E. (2015, July). A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 707–712). doi:10.3115/v1/P15-2116

- Wang, J., Wu, L., Wang, H., Choo, K. K. R., & He, D. (2020). An efficient and privacy-preserving outsourced support vector machine training for internet of medical things. *IEEE Internet of Things Journal*, 8(1), 458–473. doi:10.1109/JIOT.2020.3004231
- Wang, X., Li, C., Zhao, J., & Yu, D. (2021, May). NaturalConv: A Chinese dialogue dataset towards multi-turn topic-driven conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16), 14006–14014.
- Wen, T. H., Vandyke, D., Mrkšić, N., Gasic, M., Barahona, L. M. R., Su, P. H., Ultes, S., & Young, S. (2017, April). A network-based end-to-end trainable task-oriented dialogue system. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 438–449). doi:10.18653/v1/E17-1042
- Wu, L., Quan, C., Li, C., Wang, Q., Zheng, B., & Luo, X. (2019). A context-aware user-item representation learning for item recommendation. *ACM Transactions on Information Systems*, 37(2), 1–29. doi:10.1145/3298988
- Xu, J., Wang, H., Niu, Z., Wu, H., & Che, W. (2020, April). Knowledge graph grounded goal planning for open-domain conversation generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5), 9338–9345. doi:10.1609/aaai.v34i05.6474
- Xu, L., Zhou, Q., Gong, K., Liang, X., Tang, J., & Lin, L. (2019, July). End-to-end knowledge-routed relational dialogue system for automatic diagnosis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 7346–7353. doi:10.1609/aaai.v33i01.33017346
- Yan, Z., Duan, N., Chen, P., Zhou, M., Zhou, J., & Li, Z. (2017). Building task-oriented dialogue systems for online shopping. *31st AAAI Conference on Artificial Intelligence*.
- Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., Zhang, R., Fang, H., Zhu, P., Chen, S., & Xie, P. (2020, January). MedDialog: Large-scale medical dialogue dataset. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. doi:10.18653/v1/2020.emnlp-main.743
- Zhang, H., Liu, Z., Xiong, C., & Liu, Z. (2020, July). Grounded conversation generation as guided traverses in commonsense knowledge graphs. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2031–2043. doi:10.18653/v1/2020.acl-main.184
- Zhao, Y., Wang, Z., Wang, P., Tim, Y., Zhang, R., & Yin, K. (2020, October). A survey on task-oriented dialogue systems. *Chinese Journal of Computers*, 43, 1862–1896.
- Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., & Zhu, X. (2018, July). Commonsense knowledge aware conversation generation with graph attention. *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, 4623–4629. doi:10.24963/ijcai.2018/643
- Zhou, H., Zheng, C., Huang, K., Huang, M., & Zhu, X. (2020, July). KdConv: A Chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7098–7108. doi:10.18653/v1/2020.acl-main.635

ENDNOTE

¹ https://www.biendata.xyz/competition/ccks_2021_mdg/data/

Meng Wang is currently a graduate student at School of Computer Science, University of Wuhan University of Science and Technology. His research interests include Data mining, Knowledge Graphs and Task-oriented Dialogue Systems.

Feng Gao obtained his PhD degree in computer science from National University of Galway, Ireland in 2016, and has been working in Wuhan University of Science and Technology since 2016. His main research area is Knowledge Graph/Semantic Web.

Jinguang Gu obtained his PhD degree from Wuhan University in 2005. He is currently a professor and doctoral supervisor in Wuhan University of Science and Technology. His main research areas are Distributed Computing and Knowledge Graphs. He has published more than 50 papers and chaired more than 10 projects at provincial and ministerial levels.