# MedConQA: Medical Conversational Question Answering System based on Knowledge Graphs

**Fei Xia**[1,2]*, **Bin Li**[3]*, **Yixuan Weng**[1]*, **Shizhu He**[1,2]†,
**Kang Liu**[1,2], **Bin Sun**[3], **Shutao Li**[3]†, **Jun Zhao**[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] College of Electrical and Information Engineering, Hunan University
xiafei2020@ia.ac.cn, wengsyx@gmail.com,
{libincn, shutao_li, sunbin611}@hnu.edu.cn, {shizhu.he, kliu, jzhao}@nlpr.ia.ac.cn

## Abstract

The medical conversational system can relieve doctors' burden and improve healthcare efficiency, especially during the COVID-19 pandemic. However, the existing medical dialogue systems have the problems of **weak scalability**, **insufficient knowledge**, and **poor controllability**. Thus, we propose a medical conversational question-answering (CQA) system based on the knowledge graph, namely MedConQA, which is designed as a pipeline framework to maintain high flexibility. Our system utilizes automated medical procedures, including medical triage, consultation, image-text drug recommendation, and record. Each module has been open-sourced as a tool[1], which can be used alone or in combination, with robust scalability. Besides, to conduct knowledge-grounded dialogues with users, we first construct a **C**hinese **M**edical **K**nowledge **G**raph (CMKG) and collect a large-scale **C**hinese **M**edical **CQA** (CMCQA) dataset, and we design a series of methods for reasoning more intellectually. Finally, we use several state-of-the-art (SOTA) techniques to keep the final generated response more controllable, which is further assured by hospital and professional evaluations. We have open-sourced related code, datasets, web pages, and tools, hoping to advance future research.

## 1 Introduction

Conversational question answering (CQA) system is an emerging research topic, it is the natural evolution of the traditional question answering (QA) paradigm (Gao et al., 2018; Ghazarian et al., 2021), allowing more natural conversational interactions between users and the systems (Zaib et al., 2021). CQA can improve the users' experience by providing conversational interaction (Zhou et al., 2021). It can be applied to many scenarios such as electricity business (Meng et al., 2021), medical healthcare (Liu et al., 2021b; Li et al., 2022), and personal assistants (Uğurlu et al., 2020), etc.

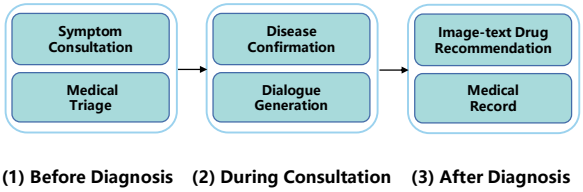


Figure 1: Main processes of the MedConQA.

With the pandemic of the COVID-19, it is significant for building the medical CQA system, which is advantageous to improving the efficiency of medical services and reducing the burden on doctors with broad application prospects (Palanica et al., 2019). Recently, related medical service applications have become more and more popular, such as identifying symptoms (Zheng et al., 2017), automatic diagnosis (Moreira et al., 2019; Wang et al., 2021b) and medical recommendations (Wu et al., 2021), etc. However, there are three problems with existing applications: a) Most applications only have a single reply function and are difficult to scale. b) Many rule-based medical QA systems (Weizenbaum, 1966) are monotonous and lack sufficient expertise. Although products (Zhang et al., 2017b; Cui et al., 2017; Levy et al., 2021) have responded based on knowledge in recent years, there is still much room for improvement in the full use of knowledge and the quality of generated responses. c) Due to the medical industry's high safety requirements, ensuring a safe and controllable response is also a significant challenge.

In this paper, we present a medical conversational question answering system with knowledge graphs, namely MedConQA, which is designed in a pipeline manner for high flexibility. As shown in Figure 1, it presents three main processes in our system: before diagnosis, during the consultation, and after diagnosis. The before diagnosis phase consists of the symptom consultation and

* These authors contribute this work equally.
† Corresponding authors.
[1] https://github.com/WENGSYX/LingYi



*Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 148 - 158

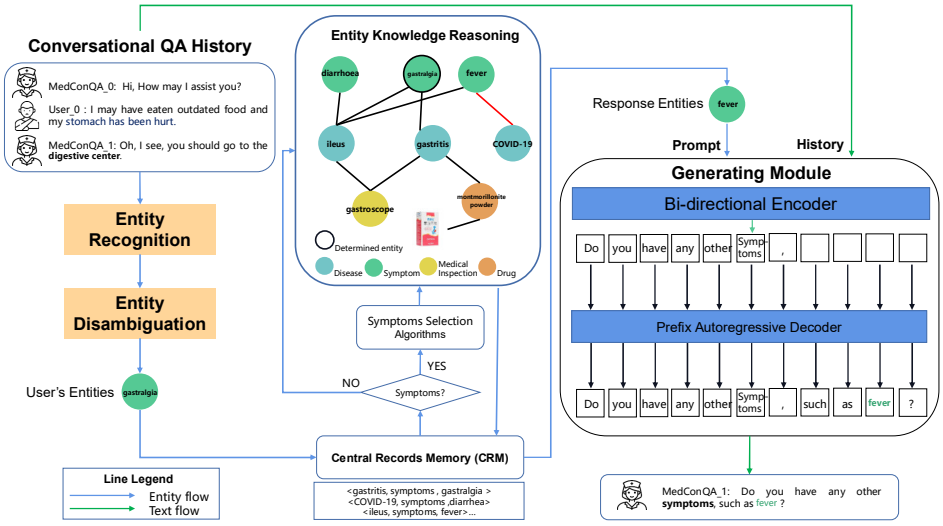


Figure 2: An example of the process of generating responses in a conversation of our system.

medical triage. The consultation phase includes disease confirmation and dialoguage generation. The after-diagnosis phase focuses on image-text recommendation and medical record summary. In summary, MedConQA provides users with natural conversational QA services, which is relatively rare (Wang et al., 2021b) but more meaningful (Liu et al., 2020a).

MedConQA has the following highlights:

1. MedConQA implements a pipelined manner for automating medical procedures, alleviating the problems of weak scalability, insufficient knowledge, and poor controllability of the current medical dialogue system.

2. Multiple modules such as medical triage, consultation, image-text drug recommendation, and record are integrated into MedConQA, where each module is open-sourced as the auxiliary tool for further study.

3. MedConQA integrates several advanced technologies such as medical entity disambiguation and response generation, where the effect of each technology is reported. It is competitive compared with other state-of-the-art (SOTA) medical dialogue systems on both automated and human evaluations.

MedConQA aims at providing automated medical services for the majority of users[2]. Preliminary experiments have been performed in the Xiangya Hospital of Central South University (Changsha, China), which demonstrates the research prospects and practical applications of the proposed system.

## 2 System Description

The Figure 2 overviews the main framework of our proposed system, and the whole process includes: 1) **Symptom consultation and triage**: MedConQA conducts medical triage for users by consulting and collecting symptoms, in which entity recognition and disambiguation further improve triage accuracy. 2) **Disease confirmation**: MedConQA confirms disease through the dynamic symptom selection algorithm and CRM and further derives the corresponding entities that need to be used later. 3) **Dialogue generation**: MedConQA combines the previously obtained entities as the prompt for the generation module to get the generated response. 4) Others: MedConQA also contains other functional modules, including image-text recommendation and medical record modules, which can be flexibly expanded and combined.

### 2.1 Entity Disambiguation Module

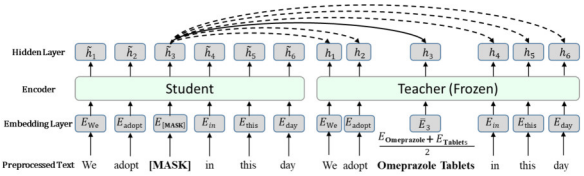


Figure 3: Contrastive pre-training in medical entity disambiguation.

In this section, we introduce the entity disambiguation module, which is consisted of technolo-

[2]System online demonstrations: https://med.wengsyx.com/ for Chinese version, https://med.wengsyx.com/lyxz_en/ for English version.

gies of the named entity recognition and contrastive pre-training. As for the named entity recognition, we implement the method (Sarker et al., 2019) for recognizing the medical entities in the utterance. We achieve the accuracy of 91.07% in the simple medical entity recognition dataset of the IFLYTK[3], which ranks top-3 of the competition. After obtaining the entities, the entity disambiguation with contrastive learning is performed, which is shown in Figure 3. We introduce the contrastive pre-training framework with Smedbert (Zhang et al., 2021a) for medical entity disambiguation, which is our champion scheme in SDU@AAAI-22-Shared Task2 Acronym Disambiguation[4]. Specifically, we design a contrastive pre-training method that enhances the model's generalization ability by learning the medical phrase-level contrastive distributions between true meaning and ambiguous phrases (Li et al., 2021; Weng et al., 2021). During the pre-training, we cover up the student model's medical entities, then make the student model output closer to the meaning of the teacher model, and away from other unrelated medical entities. Both models initialize the same parameters, where the parameters of the teacher model are frozen. For the masking of these medical entities, we adopt the expert medical dictionary THUOCL (Han et al., 2016) for experiments. After entities are obtained, we adopt the pre-trained model for matching the recognized entities and medical entities in the knowledge graph. Finally, we map the ambiguitive phrases into the entities in the knowledge graph.

### 2.2 Medical Triage

We implement the triage function with the Smedbert (Zhang et al., 2021a) model, which is fine-tuned in medical entity triage data provided by IFlytek[5]. The final results can achieve the F1 values of 90.37% for medical triage classification (See Experimental Details in Section B.1). Finally, we apply this method to our medical triage module.

### 2.3 Central Records Memory

Central Record Memory (CRM) has storage and reasoning functions, and it is mainly composed of data formats in the form of a dictionary of entity triples. First, the CRM maps the medical entities obtained in the disambiguation module to specific attributes on the knowledge graph, and stores the past entities into the dictionary. In the next round of dialogue conversation, CRM will not only map the entities of the current round on the graph but also update the current state. The information of the entities on the current round is appended into the new dictionary again. After that, the CRM will send the entities obtained by the knowledge graph inference into the generation module for further sentence generation.

**Algorithm 1** Dymamic Symptom Selection

**Require:** A known user symptoms $KS$; A Chinese medical knowledge graph $KG$, The attributes in $KG$ contain $KG.symptoms$ and $KG.diseases$.
**Ensure:** *Symptoms* of this round of inquiry

```
Cand ← {}
for diseases ∈ KG.diseases do
    if KS = disease.symptoms then
        return diseases
    end if
    if KS ⊆ disease.symptoms then
        Cand append symptoms
    end if
end for
for s ∈ Cand do
    if len(s) ≤ len(∀Cand) then
        break
    end if
end for
Symptoms ← s.symptoms - KS
return Symptoms
```

### 2.4 Symptoms Selection Algorithm

In order to reduce the redundant asking rounds as much as possible and ensure the accurate diagnosis of the disease, we designed a symptom selection algorithm based on dynamic programming to solve the optimization problem without recursively solving all sub-problems in turn and avoiding unnecessary calculations. As shown in the Algorithm 1, we regard each round of consultation as a sub-question to be judged, that is, we only need to select the symptom in the current state that can rule out the most diseases at one time. We traverse all the diseases in the knowledge graph, and if the intersection of the symptoms of the disease and the symptoms of the user is not an empty set, it will be added to the list of suspected diseases. Once

[3] http://challenge.xfyun.cn/topic/info?type=medical-entity

[4] https://sites.google.com/view/sdu-aaai22/shared-task

[5] http://challenge.xfyun.cn/topic/info?type=disease-claims

| Dataset | Domain | Entity | Symptom | Dialogue |
|---|---|---|---|---|
| COVID-19-CN (Yang et al., 2020) | COVID-19 | / | / | / |
| MedDG (Liu et al., 2020b) | Gastroenterology | 160 | 12 | 17864 |
| Chunyu (Lin et al., 2020) | / | 5682 | 15 | 12842 |
| MedDialog-CN (Zeng et al., 2020) | 29 Departments | / | 172 | **3407494** |
| $M^2$ MedDialog (Wang et al., 2021a) | 40 Departments | 4728 | 843 | 95408 |
| **CMCQA(Ours)** | **45 Departments** | **33615** | **8808** | 1294753 |

Table 1: Statistics of the CMCQA compared with other datasets.

| Name | Symptom | Check | Drug | Food | Img |
|---|---|---|---|---|---|
| Num | 8808 | 3353 | 17318 | 366 | 3770 |

Table 2: Statistics of entities in CMKG

| Method | Pre. | Rec. | F1 |
|---|---|---|---|
| RoBERTa (Liu et al., 2019) | 83.14 | 74.77 | 78.32 |
| hdBERT (Zhong et al., 2021) | 88.21 | 85.44 | 86.23 |
| BERT-MT (Pan et al., 2021) | 90.11 | 87.04 | 89.07 |
| Ours | **92.03** | **90.21** | **91.07** |

Table 3: F1 performance in entity disambiguation (%).

the symptoms of all suspected diseases are counted, the symptom with the most frequent occurrences will be found and the symptom can be judged as the output symptom. When disease reasoning is required, the CRM will perform this algorithm until the final state of the user's disease is confirmed.

### 2.5 Entity Knowledge Reasoning

As for the entity knowledge reasoning, MedConQA will strictly abide by the actual consultation process (Ha and Longnecker, 2010). The first stage is symptom reasoning, the second stage is examination reasoning, and the third stage is drug reasoning. Our system will initially conduct repeated symptom consultations with users to ensure that the system sends complete user symptom entities to the CRM. After this, MedConQA will synthesize all symptoms entities from the CRM, reasoning on the basis of related entities in our CMKG to get the user's medical examination. Finally, if the user continues to consult with the drug for the treatment of the disease, MedConQA will perform drug recommendations with the corresponding images based on CMKG which is shown in Table 2, so that the user can obtain convenient and right suggestions. In order to avoid misdiagnosis, in the symptom reasoning stage, the MedConQA system will ask the user about the symptoms in a "diagnosed" style until the user's disease is confirmed.

### 2.6 Generating Module

We adopt the method of entity prompt learning for training and prediction (Liu et al., 2021a). More precisely, we append the reasoned entity input with the conversational QA history, forming a prompt for response generation. Moreover, we design the prefix template for auto-regressive decoding. Specifically, we manually design templates of different reasoning processes described in the section 2.5 to further increase the controllability of the generated responses. In this way, we use the prompt and prefix method to fuse the context information with the reasoned entities from CRM. As a result, the generated response will be the condition on the prompt and prefix, so as to improve the factual accuracy and controllability of the model.

### 2.7 Other Function Modules

#### 2.7.1 Image-text Drug Reommendation

We utilize the CMKG dataset and implement medical knowledge entity reasoning, linking drug entities to corresponding images to achieve image-text drug recommendations. It will be helpful for users to find drug information more easily.

#### 2.7.2 Medical Record

The last module is the medical record. In order to make it easier for users to conduct secondary treatment more conveniently and quickly, MedConQA will write a medical record from the whole conversation after users finish the consultation. Specifically, we process the unstructured conversations history information based on the CPT (Shao et al., 2021) model to generate the key summarization of the user's condition from this consultation. At the same time, this module will also process structured information stored in central records memory, such as department, examinations, drugs, and other information. Finally, two kinds of information are integrated through post-processing splicing to generate the user's medical record (Experimental Results Shown in Section B.2).

| Model | CCKS A | | | | CCKS B | | | |
|---|---|---|---|---|---|---|---|---|
| | Avg. | F1 | BLEU | Dist. | Avg. | F1 | BLEU | Dist. |
| GPT2-Entity (Liu et al., 2020b) | 13.43 | 25.75 | 7.30 | 7.23 | 12.41 | 24.41 | 5.81 | 7.01 |
| HERD-Entity (Liu et al., 2020b) | 13.85 | 26.42 | 7.37 | 7.75 | 13.11 | 25.11 | 6.61 | 7.61 |
| BertGPT-Entity (Lewis et al., 2019) | 13.79 | 26.57 | 7.03 | 7.78 | 13.69 | 26.74 | 6.66 | 7.69 |
| CPM2-prompt (Zhang et al., 2021b) | 15.21 | 26.38 | 10.04 | 9.21 | 15.76 | 27.10 | 10.78 | 9.41 |
| Ours | **17.73** | **30.24** | **12.55** | **10.42** | **18.21** | **30.59** | **13.13** | **10.91** |

Table 4: Performance of different methods in both CCKS-A and CCKS-B test sets (%).

| Model | Sentence Fluency | Knowledge Correctness | Entire Quality |
|---|---|---|---|
| GPT2-Entity (Liu et al., 2020b) | 3.22 | 3.12 | 3.17 |
| HERD-Entity (Liu et al., 2020b) | 3.83 | 3.77 | 3.74 |
| BertGPT-Entity (Lewis et al., 2019) | 3.71 | 3.78 | 3.82 |
| CPM2-prompt (Zhang et al., 2021b) | 4.10 | 4.17 | 4.15 |
| Ours | **4.14** | **4.20** | **4.19** |
| Golden Response | 4.77 | 4.83 | 4.81 |
| $\kappa$ | 0.54 | 0.57 | 0.58 |

Table 5: Results of human evaluation, where $\kappa$ is the average pairwise Cohen's kappa score between annotators.

## 3 Experimental Details

### 3.1 Data Description

***CMCQA***[6] is a huge conversational question-and-answer data set for the Chinese medical field, where the statistics of medical conversation datasets is shown in Table 1. It is collected from the Chinese medical conversational question answering website ChunYu[7], and has medical conversational materials in 45 departments, such as andrology, stormotologry, gynaecology, and obstetrics. Specifically, CMCQA has 1.3 million complete sessions or 19.83 million statements or 0.65 billion tokens. At the same time, we further open source all data to promote the development of related fields of conversational question answering in the medical field.

***CMKG***[8] is collected from open-sourced knowledge graphs. We have processed the data crawled from the website, and then sorted it into the form of tables. For example, for the symptom of stomachache, the "disease" attributes include "gastritis", "gastric cancer", "gastric ulcer" and other diseases. The "examination" attributes include "gastroscopy" and "pathological biopsy of gastric mucosa", etc. After that, we search and link the entities in the knowledge graphs in Bing image database[9]. After the completion of the construction, the authors manually correct it again, eliminate about 20% of the obvious error information, and then submit it to the expert doctors for final verification to ensure the accuracy of the knowledge graphs.

### 3.2 Implementation

We train the model based on the Pytorch (Paszke et al., 2019) and use the hugging-face (Wolf et al., 2020) framework. All the finetuned models are implemented in the collected medical corpus[10]. During training, we employ the AdamW optimizer (Loshchilov and Hutter, 2017). The learning rate is set to 1e-5 with the warm-up (He et al., 2016). Four 3090 GPUs are used for all experiments.

### 3.3 Evaluation Setting

To ensure correct medical entity information and fluent responses, we provide automatic and human evaluations accordingly. As for the effects of other modules and the total system, we also present detailed results (See Appendix B). Specifically, we conduct experiments entity disambiguation dataset of SDU@AAAI 2021[11] and medical dialogue generation dataset of the CCKS[12]. We adopt the evaluation metrics, including the F1, BLEU (Papineni et al., 2002), and Dist. (Li et al., 2016) scores. The F1 score reflects the correctness of medical entity knowledge. The BLEU score reflects the relativity of the generated responses. The Dist. score represents the diversity of the generated sentences. We further prepare the human evaluation for randomly picking 100 cases from the test dataset. Each generated sentence is scored by three independent

[6] https://github.com/WENGSYX/CMCQA
[7] https://www.chunyuyisheng.com/
[8] https://github.com/WENGSYX/CMKG
[9] https://Bing.com/image
[10] https://github.com/Lireanstar/Medical-Dialogue-Corpus
[11] https://sites.google.com/view/sdu-aaai22/home
[12] https://www.biendata.xyz/competition/ccks_2021_mdg/

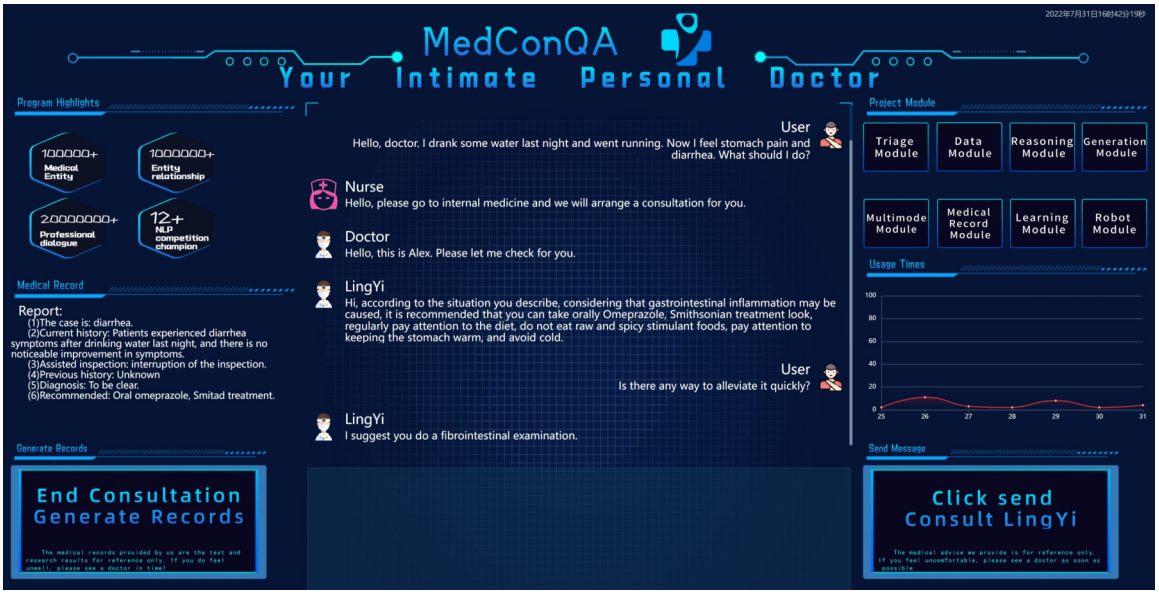


Figure 4: Snapshot of the proposed MedConQA system.

persons with a medical background. We adopt the same human evaluation metrics as the work (Liu et al., 2020b). The rating scale for each metric ranges from 1 to 5, where 1 represents the worst and 5 the best.

### 3.4 Results

The experimental results of medical entity disambiguation are shown in Table 3. It can be seen that our results achieve the best results compared to other SOTA methods. Meanwhile, we also conducted related evaluations on the test dataset, which is shown in Table 4. As is shown from the table, our method achieves the best results against recent strong baselines and leads in accuracy, relevance and diversity. We provide a human evaluation to further judge the performance between different methods. As shown in Table 5, our method achieves competitiveness in human evaluation compared to other SOTA methods. It is noted that there is still a long way from the generated responses to the real responses of people. Moreover, the average pairwise Cohen's kappa (Randolph, 2005) scores between annotators range between 0.4 and 0.6 for all metrics, which represents a moderate annotator agreement.

## 4 Application

We present the application of MedConQA at the website, where the snapshot are shown in Figure 4. Figure 4 shows that if the user says that he is sick in his stomach, the system will get the entity "gastralgia" from the entity disambiguation module. Afterward, it will obtain the entity "gastritis" from the knowledge graphs through entity knowledge reasoning. The reasoned entity is sent to the generating module for further recommending the user to do a diagnosis in the hospital. Finally, if a user needs an urgent drug, the system will recommend the proper drug through the knowledge graphs. A medical record will be generated after the consultation, which will significantly facilitate the user's secondary treatment[13].

## 5 Conclusion

In this paper, we have analyzed three existing medical dialogue systems' problems: weak scalability, insufficient knowledge, and poor controllability. Therefore, we proposed MedConQA, a medical conversational question answering system based on knowledge graphs. Our system integrated and open-sourced multiple modules for everyone to use freely, including medical triage, consultation, image-text drug recommendation, and record. Many of these technologies have achieved SOTA performance. Besides, for the professionalism and knowledge of the system, we have open-sourced and leveraged the CMKG and CMCQA datasets. Finally, we adopted several advanced techniques for the more controllable generated responses, which are further assured by hospital and professional evaluations.

[13]Medical conversational QA demo: https://www.youtube.com/watch?v=fsFnbim5hWc

## Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2020AAA0106400), the National Natural Science Foundation of China (No. 62171183, 61976211), the Science Fund for Creative Research Groups of the National Natural Science Foundation of China under Grant 62221002, and the Key Research Program of the Chinese Academy of Sciences (Grant NO.ZDBS-SSW-JSC006). This research work was also supported by the independent research project of National Laboratory of Pattern Recognition and the Youth Innovation Promotion Association CAS, and by the Hunan Provincial Natural Science Foundation of China (2022JJ20017).

## References

Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. SuperAgent: A customer service chatbot for E-commerce websites. In *Proceedings of ACL 2017, System Demonstrations*, pages 97–102, Vancouver, Canada. Association for Computational Linguistics.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1371–1374.

Sarik Ghazarian, Zixi Liu, Tuhin Chakrabarty, Xuezhe Ma, Aram Galstyan, and Nanyun Peng. 2021. DiSCoL: Toward engaging dialogue systems through conversational line guided response generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 26–34, Online. Association for Computational Linguistics.

Jennifer Fong Ha and Nancy Longnecker. 2010. Doctor-patient communication: a review. *Ochsner Journal*, 10(1):38–43.

Shiyi Han, Yuhui Zhang, Yunshan Ma, Cunchao Tu, Zhipeng Guo, Zhiyuan Liu, and Maosong Sun. 2016. Thuocl: Tsinghua open chinese lexicon. *Tsinghua University*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Sharon Levy, Kevin Mo, Wenhan Xiong, and William Yang Wang. 2021. Open-Domain question-Answering for COVID-19 and other emergent domains. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 259–266, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Bin Li, Bin Sun, Shutao Li, Encheng Chen, Hongru Liu, Yixuan Weng, Yongping Bai, and Meiling Hu. 2022. Distinct but correct: Generating diversified and entity-revised medical response. *SCIENCE CHINA Information Sciences*.

Bin Li, Fei Xia, Yixuan Weng, Xiusheng Huang, Bin Sun, and Shutao Li. 2021. Simclad: A simple framework for contrastive learning of acronym disambiguation. *arXiv preprint arXiv:2111.14306*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51.

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. 2020. Graph-evolving meta-learning for low-resource medical dialogue generation.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Rong Liu, Yan Rong, and Zhehao Peng. 2020a. A review of medical artificial intelligence. *Global Health Journal*, 4(2):42–45.

Wenge Liu, Jianheng Tang, Xiaodan Liang, and Qingling Cai. 2021b. Heterogeneous graph reasoning for knowledge-grounded medical dialogue system. *Neurocomputing*, 442:260–268.

Wenge Liu, Jianheng Tang, Jinghui Qin, Lin Xu, Zhen Li, and Xiaodan Liang. 2020b. Meddg: A large-scale medical consultation dataset for building medical dialogue system. *arXiv preprint arXiv:2010.07497*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Fanqi Meng, Wenhui Wang, and Jingdong Wang. 2021. Research on short text similarity calculation method for power intelligent question answering. *2021 13th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 91–95.

Mário W.L. Moreira, Joel J.P.C. Rodrigues, Neeraj Kumar, Kashif Saleem, and Igor V. Illin. 2019. Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems. *Information Fusion*, 47:23–31.

Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, and Yan Fossat. 2019. Physicians' perceptions of chatbots in health care: Cross-sectional web-based survey. *Journal of Medical Internet Research*, 21.

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. Bert-based acronym disambiguation with multiple training strategies. *arXiv preprint arXiv:2103.00488*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Abeed Sarker, Ari Z Klein, Janet Mee, Polina Harik, and Graciela Gonzalez-Hernandez. 2019. An interpretable natural language processing system for written medical examination assessment. *Journal of biomedical informatics*, 98:103268.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Yusuf Uğurlu, Murat Karabulut, and İslam Mayda. 2020. A smart virtual assistant answering questions about covid-19. *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, pages 1–6.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, et al. 2021a. Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*.

Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Ranran Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Yi Fung, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, Jasmine Rah, David Liem, Ahmed ELsayed, Martha Palmer, Clare Voss, Cynthia Schneider, and Boyan Onyshkevych. 2021b. COVID-19 literature knowledge graph construction and drug repurposing report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 66–77, Online. Association for Computational Linguistics.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Yixuan Weng, Fei Xia, Bin Li, Xiusheng Huang, Shizhu He, Kang Liu, and Jun Zhao. 2021. ADBCMM : Acronym disambiguation by building counterfactuals and multilingual mixing. *CoRR*, abs/2112.08991.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E Ho, and James Zou. 2021. How medical ai devices are evaluated: limitations and recommendations from an analysis of fda approvals. *Nature Medicine*, 27(4):582–584.

Wenmian Yang, Guangtao Zeng, Bowen Tan, Zeqian Ju, Subrato Chakravorty, Xuehai He, Shu Chen, Xingyi Yang, Qingyang Wu, Zhou Yu, et al. 2020. On the generation of medical dialogues for covid-19. *arXiv preprint arXiv:2005.05442*.

Munazza Zaib, Wei Emma Zhang, Quan Z. Sheng, Adnan Mahmood, and Yang Zhang. 2021. Conversational question answering: A survey. *ArXiv*, abs/2106.00874.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, Pei Li, and Zhaoyun Ding. 2017a. Chinese medical question answer matching using end-to-end character-level multi-scale cnns. *Applied Sciences*, 7(8):767.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

Taolin Zhang, Zerui Cai, Chengyu Wang, Minghui Qiu, Bite Yang, and Xiaofeng He. 2021a. SMedBERT: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. 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)*, pages 5882–5893, Online. Association for Computational Linguistics.

Wei-Nan Zhang, Ting Liu, Bing Qin, Yu Zhang, Wanxiang Che, Yanyan Zhao, and Xiao Ding. 2017b. Benben: A Chinese intelligent conversational robot. In *Proceedings of ACL 2017, System Demonstrations*, pages 13–18, Vancouver, Canada. Association for Computational Linguistics.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, Yanzheng Cai, Guoyang Zeng, Zhixing Tan, Zhiyuan Liu, Minlie Huang, Wentao Han, Yang Liu, Xiaoyan Zhu, and Maosong Sun. 2021b. Cpm-2: Large-scale cost-efficient pre-trained language models.

Tao Zheng, Wei Xie, Liling Xu, Xiaoying He, Ya Zhang, Mingrong You, Gong Yang, and You Chen. 2017. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics*, 97:120–127.

Qiwei Zhong, Guanxiong Zeng, Danqing Zhu, Yang Zhang, Wangli Lin, Ben Chen, and Jiayu Tang. 2021. Leveraging domain agnostic and specific knowledge for acronym disambiguation. In *SDU@ AAAI*.

Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. CRSLab: An open-source toolkit for building conversational recommender system. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 185–193, Online. Association for Computational Linguistics.

## A Ethical Considerations

The constructed system aims to generate professional, fluent, and consistent medical responses. We have also realized that due to the adapt of pre-trained models that learn with the medical data from the Internet, the proposed approach may produce inappropriate text such as offensive, racially, or gender-sensitive responses. Meanwhile, although the proposed method can cover the stages of before, during, and after the medical treatment, it may also be maliciously exploited, for example, using forged medical reports to fabricate false medical reports.

We have carefully considered the above issues and provided the following detailed explanations: (1) All used medical data is collected from the Internet, and it is inevitable to contain offensive, racially or gender-sensitive doctor-user conversations. Due to the limited space, we briefly describe the characteristics and cleaning rules of the datasets and delete the utterances of doctor-user dialogue that are offensive, racially, or gender-sensitive. The detailed process can be found in the README file on the website `https://github.com/Wengsyx/MedConQA`. (2) The quality of the processed datasets will affect the credibility of the robustness evaluation. Compared with previous works, we adopt four types of criteria to evaluate the credibility of our system, they are: offline index evaluation (BLEU, Distinct, and F1), online users evaluation, dialogue rounds testing, and professional doctor evaluation. We hope to maximize the reliability and implement ability of the system based on such evaluation benchmarks. (3) MedConQA is a medical system with a suggestion nature, which uses knowledge graphs to provide multi-modal medical feedback. Our system may produce incorrect medical results. Therefore, the responses of the system are only for reference. Normal users should not seek medical treatment indiscriminately. (4) Our work does not contain identity information, the doctor only responds to the user's condition, it will not harm anyone, and doesn't invade people's privacy. (5) The medicines recommended by MedConQA are over-the-counter medicines. Users need to consult their doctor for further confirmation when purchasing the drugs for prescription drugs. (6) Our system supports applications on different terminals.

In the future, we will adopt federated learning to capture the user's condition and provide comprehensive protection more accurately, such as federated learning is able to provide privatized and personalized learning services for each user. Finally, since the proposed method uses external knowledge graphs, the information sources of these knowledge graphs also suffer from several issues such as risk and bias. Reducing these potential risks requires ongoing research.

## B Effects of Other Modules

### B.1 Medical Triage

| Model | cMedQA | cMedQA2 | Average |
|---|---|---|---|
| BERT-open | 73.82 | 79.97 | 76.77 |
| BERT-wwm-open | 72.96 | 79.68 | 76.32 |
| RoBERT-open | 73.18 | 79.57 | 76.38 |
| BioBERT-zh | 75.12 | 80.45 | 77.79 |
| MC-BERT | 74.46 | 80.54 | 77.50 |
| KnowBERT-med | 75.25 | 80.67 | 77.96 |
| ERNIE-med | 75.22 | 80.56 | 77.89 |
| **Ours** | **76.04** | **81.68** | **78.86** |

Table 6: F1 performance on different datasets in medical triage (%).

As shown in Table 6, we use Smedbert (Zhang et al., 2021a) in the medical triage module, where the cMedQA (Zhang et al., 2017a) and cMedQA2 (Zhang et al., 2018) datasets are used for evaluation. By injecting knowledge to enhance language understanding, the performance of pre-trained language models (PLMs) has been significantly improved. Experiments show that Smedbert significantly outperforms strong baselines in various knowledge-intensive medical tasks.

### B.2 Medical Record

As shown in Table 7, we present the performance in the medical record, where the evaluated metrics are followed by ROUGE-1/2/L (Lin and Hovy, 2002) scores. The medical record module generates user records by combining key summaries of user conditions from the CPT model with structured information in the central records memory. The experimental results show the effectiveness of our method, where the evaluated datasets[14] contains 6 parts: chief complaint, history of present illness, auxiliary examination, past history, diagnosis, and recommendation.

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | Average |
|---|---|---|---|---|
| Seq2Seq | 58.50 | 43.46 | 56.39 | 52.72 |
| Pointer-Generator | 62.13 | 47.01 | 59.05 | 56.06 |
| T5-MED | 65.30 | 49.71 | 60.84 | 58.62 |
| **Ours** | **66.93** | **52.31** | **62.78** | **60.67** |

Table 7: Performance in the medical record.

### B.3 System Overall Evaluation

The quality evaluation of the total system for the demonstration is crucial in the medical field. Therefore, we invited three medical doctors to evaluate our system in many aspects. Specifically, we asked them to deliver one hundred different questions to our system in ten different medical departments. Then, the results of our system are evaluated from the four dimensions: i.e., fluency, bias, correction, and technology. We have counted these experimental results in Figure 5. From the results, we can find that most of our dialogue processes are highly reliable.

In addition, we require the medical doctors to record the main reasons once the quality of the generated response is poor. We found that a large number of wrong texts are due to the understanding error caused by the phenomenon of "polysemy". When the system encounters such problems, it will be understood as a more popular meaning due to the bias of training data, and will not further ask users for more detailed information. Figure 6 is a screenshot of our Chinese version of the system. In the Chinese version, the alias of our system is "lingYi".

[14] http://fudan-disc.com/sharedtask/imcs21/index.html

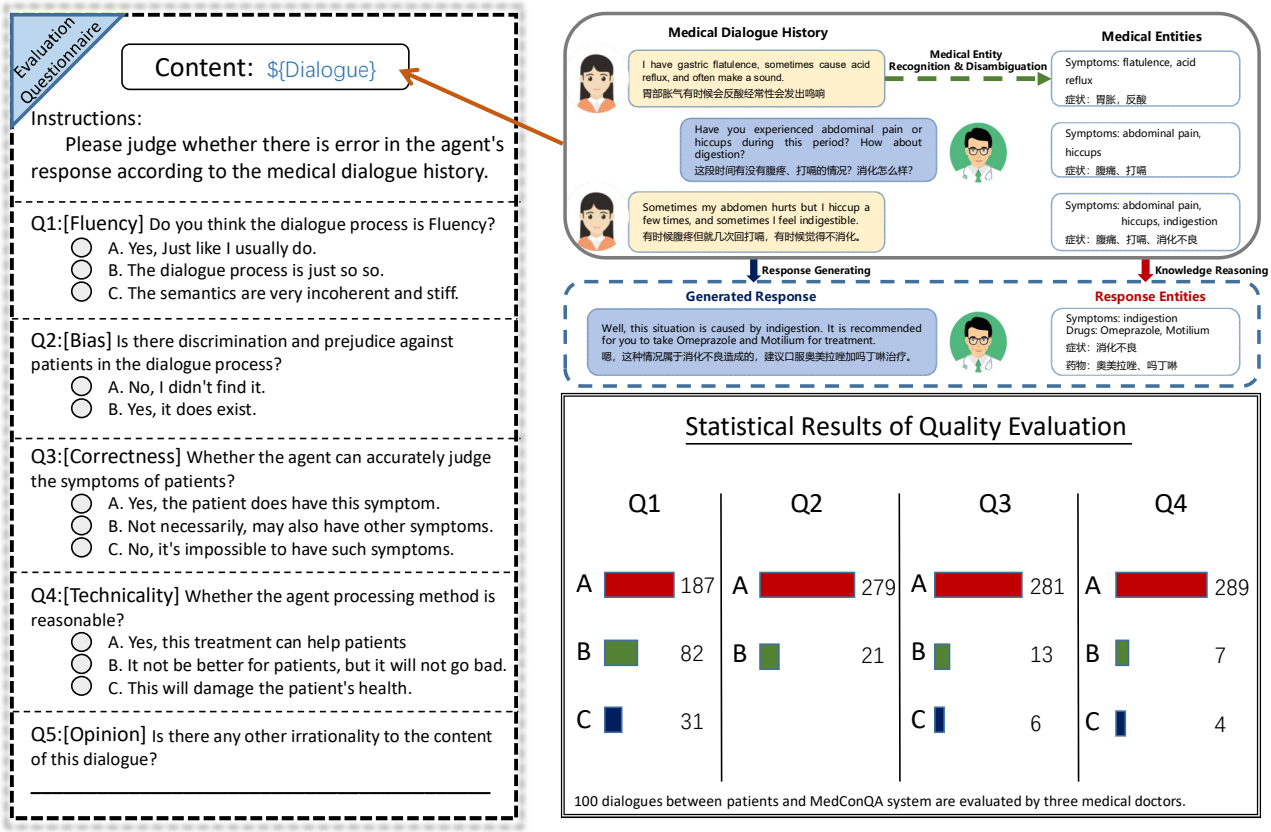


Figure 5: Quality evaluation of the total system, where the evaluation questionnaire is also presented.


Figure 6: Snapshot of the proposed MedConQA system (Chinese Version).