Knowledge-Grounded Dialogue Generation for Medical Conversations: A Survey

1st Xiaoxiao Liu
National Centre for Computer Animation (NCCA)
Bournemouth University,UK
xliu@bournemouth.ac.uk

2nd Jian Chang
NCCA
Bournemouth University,UK
jchang@bournemouth.ac.uk

3rd Jian Jun Zhang
NCCA
Bournemouth University,UK
jzhang@bournemouth.ac.uk

***Abstract*—Applying Artificial Intelligence (AI) techniques such as natural language generation in assisting medical treatment and diagnosis has made distinguished progress. One such technique is dialogue generation. The application of a medical dialogue system in assisting medical treatment has great potential to explore. This paper serves as a survey of digging application of AI techniques in knowledge-grounded dialogue generation for medical conversation systems. Meanwhile, we provide an academic visualization method to present such references.**

***Index Terms*—Dialogue generation, Medical conversation system, Knowledge-grounded dialogue generation**

I. Introduction

The growth of AI technologies in the medical field has gained growing attention over recent years. Natural Language Processing (NLP), as a critical technique in AI, helps computers communicate with humans by processing information and generating responses in natural language. In the medical field, NLP has been applied to the development of clinical decision support systems, Electronic Health Records (EHR) processing, as well as medical chatbots, which can communicate with patients like a human doctor does. However, most medical dialogue systems use rule-based or template-based approaches for response generation, which require large datasets of sentence templates and raise privacy concerns [24]. Therefore, research on automatic response generation systems that do not rely on templates has become a popular topic in building knowledge-grounded medical dialogue systems. The work of [6] contributes the first successful case of Pre-Trained Models (PTM) for NLP, the PTM such as BERT [7], Sequence to sequence PTM [28] has made significant success in many NLP task [26]. In this survey, we aim to provide an overview of recent research in knowledge-grounded dialogue generation for medical conversations, including medical conversation systems, dialogue generation approaches and designs, as well as evaluation metrics and datasets.

In this survey, we adopt data visualization as a method to present a graphical overview of our references. As interest grows in interpretable and trustworthy AI, the discussion on how to apply data visualization methods within AI development has intensified across both data visualization and AI domains. Notable examples of such approaches include the work contributed by Alhabi et al. [3] who provide a comprehensive visualization of an NLP experiment. The research of [29] serves as a literature review on visualizing EHRs and EHR datasets. They also provides an interactive visualization tool for both the EHR references and the EHR dataset. Inspired by the work such as [4], [23], [29], we adopt SurVis as our visualization tool for reference resources. This approach enables readers to quickly gain the patterns within the reference data, further emphasizing the role of data visualization as an effective tool in understanding and interpreting complex data.

The contributions of this unique survey include:

- A survey on knowledge-grounded medical dialogue generation system
- A novel view adopting visualization tool to analyze the references in the survey
- To our best knowledge, we are the first to explore the application of data visualization in the area of medical dialogue generation

II. Medical Conversation Systems

In recent times, NLP researchers have expressed a growing interest in generating medical conversations [19], [30], [31]. This is not just because it can automatically generate diagnostic reports, but also because it streamlines the diagnostic process by asking patients for additional symptoms and data during the conversation [12]. Additionally, it assists healthcare professionals like doctors and nurses in managing complex illnesses by expanding and keeping the knowledge system up to date. According to the different ways of implementation technology, dialogue systems can be divided into two types: end-to-end and pipeline [34]. Although works of [33] attempted to apply end-to-end methods to evaluate medical datasets performing on the LM, recent medical dialogue systems adopted pipeline methods in designing their dialogue system. The tasks to build and improve a medical conversation system mainly focus on natural language understanding (NLU) and dialogue management (DM) [18]. In addition, the medical conversation system can be enhanced by using numerical solutions for natural language understanding (NLU), such as the implementation of named entity recognition (NER) for medical terms [1], entity inference [10], [19], and symptom extraction [9]. Reinforcement learning (RL) driven task-oriented dialogue systems are mostly used for Dialogue Management (DM) [8], [15]. The use of RL was employed in the research

described in reference [30] to improve the effectiveness of automated diagnosis by training the dialogue policy. Researchers working on [31] proposed techniques for integrating knowledge reasoning into DM by utilizing Reinforcement Learning RL.

The work of [5] proposes a novel approach to automatic medical diagnosis by reformulating it as a symptoms Sequence Generation (SG) task. They introduce a simple yet effective model, called Diaformer, based on the Transformer architecture, which incorporates a symptom attention framework to learn the generation of symptom inquiries and disease diagnoses. To address the challenge of implicit symptoms and their disorder in a sequential generation, they also design three orderless training mechanisms. The experiments on three open-resource datasets demonstrate that the proposed model outperforms the baselines on disease diagnosis, with the highest training efficiency. Additionally, the authors show the potential of applying symptom sequence generation for automatic diagnosis by analyzing the prediction of symptom inquiries.

Luo et al. [22] hypotheses using automatic symptom detection as a task in the area of the medical dialogue system, in which a system can be built to chat with patients while detecting and gathering symptoms information automatically. The system is fine-tuned to collect implicit symptoms by asking questions. The decision system is then guided by certain knowledge and collects as much as information. In this way, the system is able to assist the human doctor by collecting as much patient information as possible. Two different models and a training pipeline, along with a knowledge graph for additional inputs are presented accordingly.

Wei et al. [30] describes a task-oriented dialogue system framework for automatic diagnosis, which is based on a dataset collected from an online medical forum. The system, by means of convers with users, collects additional symptoms beyond their self-reports and make a better diagnosis. Experimental results show that additional symptoms extracted from a conversation greatly improve the accuracy for disease identification.

III. Knowledge-Grounded Dialogue Generation

One of the reasons that the medical conversation system cannot replace real human doctors in the diagnosis is that existing conversation systems lack medical knowledge which can lead the system to accurate diagnosis results. Hence, research on incorporating medical knowledge into the dialogue generation system has become challenging. The goal of knowledge-grounded dialogue generation is to create relevant and valuable responses by taking into account both the ongoing conversation and information from external sources [16]. In the medical domain, the PTM has been proved efficient [13], [25]. Existing approaches to diagnosis for dialogue systems mainly rely on knowledge-grounded methods such as [20] and [17].

Xu et al. [31] discusses the development of an End-to-End Knowledge-routed Relational Dialogue System (KR-DS) for medical diagnosis, which integrates a rich medical knowledge graph into the dialogue management process to make the system more cooperative with natural language understanding and natural language generation. The paper introduces a novel Knowledge-routed Deep Q-network (KR-DQN) to manage topic transitions in the dialogue system. The experiments conducted on a public medical dialogue dataset demonstrate the superiority of the KR-DS over state-of-the-art methods by more than 8% in diagnosis accuracy. As a result, the paper shows the effectiveness of the proposed KR-DS on a newly collected medical dialogue system dataset that retains the original self-reports and conversational data between patients and doctors.

Liu et al. [21] put forward to a multi-stage prompting approach to generate knowledgeable responses from a single pre-trained language model (LM). The approach leverages the inherent knowledge stored in the pre-trained LM and its powerful generation ability to address the limitations of existing knowledge-grounded dialogue systems. The proposed method outperforms the state-of-the-art retrieval-based model when combining knowledge relevance and correctness. The paper also demonstrates the effectiveness of scaling the model.

Shalyminov et al. [27] presents the Dialogue Knowledge Transfer Network (DiKTNet), a state-of-the-art approach to goal-oriented dialogue generation that only uses a few example dialogues, none of which need to be annotated. The proposed method reduces the amount of data and annotations necessary for training goal-oriented dialogue systems, which is a central research problem in the field. The proposed approach involves performing unsupervised dialogue representation pre-training on a large source of goal-oriented dialogues in multiple domains and then training DiKTNet using this representation. The paper demonstrates the effectiveness of the proposed method on a benchmark dataset, showing that it outperforms existing few-shot learning approaches by a significant margin.

Kao et al. [12] develops a context-aware hierarchical reinforcement learning scheme for symptom checking, which results in an improvement in the accuracy of symptom checking. The paper highlights the two methods in boosting diagnosis accuracy: by adding a latent layer using anatomical parts, and by introducing ”context” into the model so that the model is empowered by context-aware ability. This research proposes a master model that coordinates anatomical part models to perform symptom checking, resulting in improved diagnostic accuracy with a similar or fewer number of symptom inquiries compared to previous work.

IV. Evaluation Metrics and Datasets

Evaluation Metrics Studies were identified by searching 7 bibliographic databases (eg, MEDLINE and PsycINFO) in addition to conducting backward and forward reference list checking of the included studies and relevant reviews. The studies were independently selected by two reviewers who then extracted data from the included studies. Extracted data were synthesized narratively by grouping the identified metrics into categories based on the aspect of chatbots that the metrics evaluated. Evaluating the medical conversation

systems is difficult due to their highly strict requirements for accuracy in diagnosis results. Research around medical conversation systems adopts both automatic evaluation and human evaluation to measure their results generation from dialogue systems, such as [31], [33]. The reason for this is that the sentences generated by the dialogue system are different in automatic evaluation metrics from in human evaluation benchmarks.

In the work of [2], the authors conducted a review of 65 health chatbot studies that were evaluated using 27 technical metrics. The metrics were related to chatbots as a whole, response generation, response understanding, as well as esthetics. The technical metrics of health chatbot studies were diverse, with survey designs and global usability metrics dominating. The lack of standardization and paucity of objective measures make it difficult to compare the performance of health chatbots and could inhibit the advancement of the field. They suggest that researchers more frequently include metrics computed from conversation logs and recommend the development of a framework of technical metrics with recommendations for specific circumstances for their inclusion in chatbot studies.

Datasets Datasets for creating medical dialogue systems are difficult to access. On one hand, most of the dataset required is in the format of doctor-patient conversations which involves the privacy of both sides. On the other hand, accessing medical-related datasets normally costs a lot of money. We collected 4 public medical datasets which can be applied to fine-tune language models to accomplish medical conversation generation. In the Table I, every dataset is accessible with a hyperlink on resource name. Each resource is categorized by according to the content and format of the data. An application column is also introduced to link the resource to the possible module in this research.

TABLE I
THE ANALYSIS OF DATASET COLLECTED FROM EXISTING PUBLISHED WORK.

Resource Name	Category	Citation	Application
MedDialog	Dialogue	[33]	DG
Respiratory Dialogue Dataset	Dialogue	[11]	DG
ReMeDi	Knowledge	[32]	DM
COVID19 Medical Dialogue	Dialouge	[35]	DG
KaMed	Dialogue	[14]	DG
MedicalChatbot-HRL	Knowledge	[30]	DM

MedDialog [33] is a large medical conversation corpus that contains a Chinese conversation dataset and a English conversation dataset. In the English corpus, 0.26 million doctor-patient conversations, 0.51 million utterances, and 44.53 million tokens are contained. The dataset refers to more than 90 types of diseases. The dataset collected from online communities, which leads to the conversation contains only several rounds between doctors and patients.

Fareez et al. [11] provides a medical conversation dataset containing 272 mp3 audio files and the corresponding transcript text files, which are categorized into six disease categories: respiratory, gastrointestinal, cardiovascular, musculoskeletal, dermatological, and miscellaneous. This dataset is in a simulation scenario in which a group of medical students are taking Objective Structured Clinical Examination (OSCE) and resident doctors are querying them following an overall format of elucidating the needed pertinent information.

ReMeDi [32] is another medical dataset that comprises 96,965 conversations between medical professionals and patients. It includes 1,557 conversations that have been labeled in detail and covers 843 types of diseases, 5,228 medical entities, and 3 specialties of medical services in 40 domains.

The work of [35] also contains an English medical dialogue dataset related COVID-19 and other kinds of pneumonia. The dataset contains 603 consultations. This dataset focuses on COVID-19 only in which patients are worried about their symptoms and query on the online platform about their situation.

Wei et al. [30] contribute a dialogue dataset that comes from the pediatric department in a Chinese online healthcare community. The researchers also annotate the datasets in terms of symptom extraction techniques and symptom normalization. Their experiments find out that extracting additional symptoms from conversational data can significantly enhance the accuracy of disease identification.

V. VISUALIZATION OF THE MEDICAL DIALOGUE SYSTEM REFERENCES

To get a better statistical view, we adopt SurVis [4], an effective visualization tool for analyzing reference data, to provide a visualization for our references. Detailed interactive data visualization of the SurVis screenshot can be seen in Figure 1. The application of the reference visualization tools provides an intuitive vision for the references according to time scale, authors, and titles of the references. The SurVis visualization of the references resources can be accessed here: https://med-survis.github.io/MeDialog-Survis.github.io/

VI. CONCLUSION

This paper serves as a survey that provides an overview of the recent research on medical dialogue system. Three main categories are presented and discussed. In addition, a visualization of all the references is provided in the hope that readers get an academic visualization view of the references on the medical dialogue system. In the future, we would like to extend our research by incorporating the newest technologies such as the application of large language models in the medical dialogue system domain, and the relevant application of how the large language model incorporates medical knowledge.

REFERENCES

[1] Macarious Abadeer. Assessment of distilbert performance on named entity recognition task for the detection of protected health information and medical concepts. In *CLINICALNLP*, 2020.

[2] Alaa Abd-Alrazaq, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, Kerstin Denecke, et al. Technical metrics used to evaluate health care chatbots: scoping review. *Journal of medical Internet research*, 22(6):e18301, 2020.

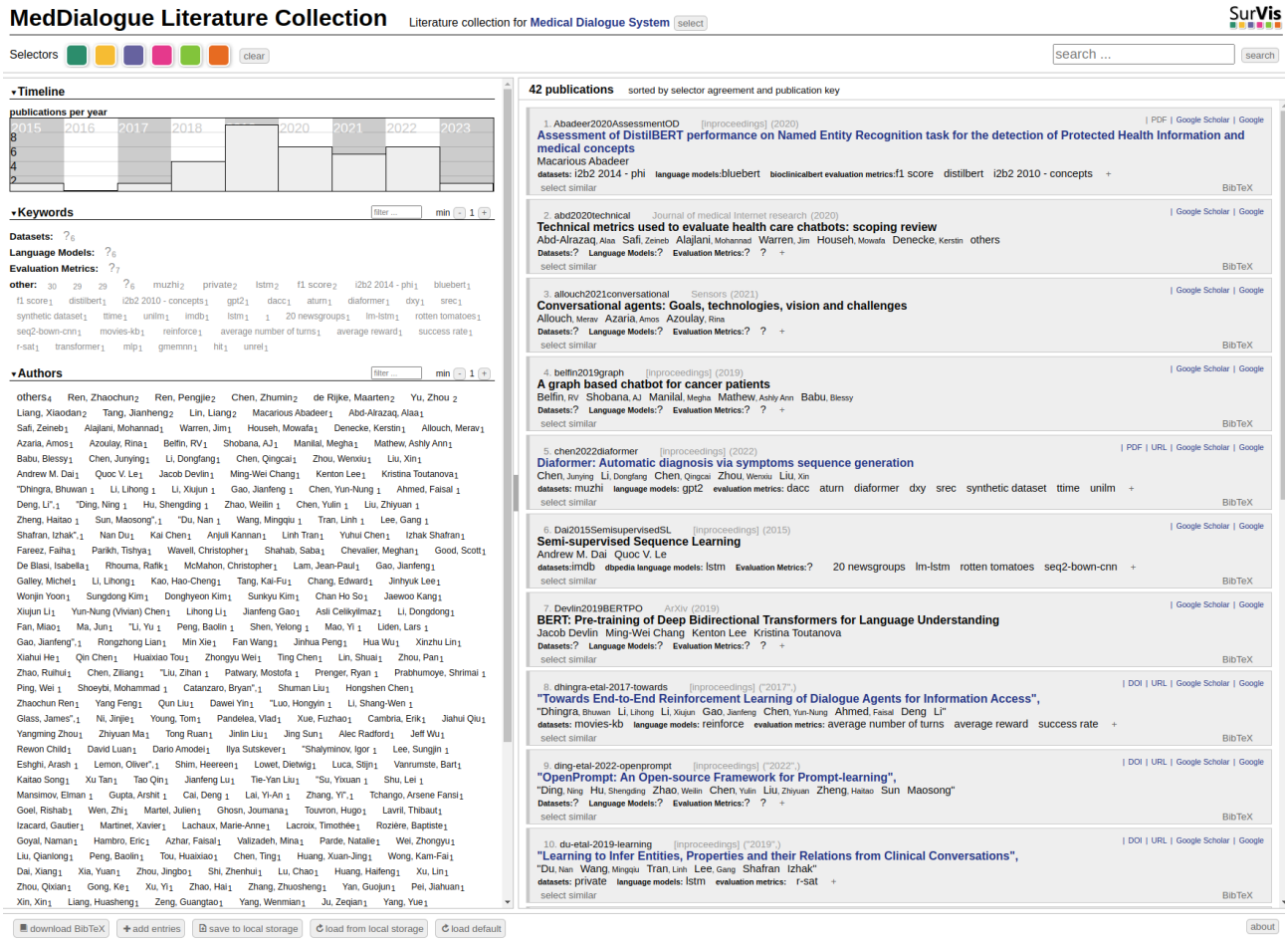

Fig. 1. The interactive Survis visualization for the references in this paper.

[3] Mohammad Alharbi, Matthew Roach, Tom Cheesman, and Robert S Laramee. Vnlp: Visible natural language processing. *Information Visualization*, 20(4):245–262, 2021.

[4] Fabian Beck, Sebastian Koch, and Daniel Weiskopf. Visual analysis and dissemination of scientific literature collections with survis. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):180–189, 2015.

[5] Junying Chen, Dongfang Li, Qingcai Chen, Wenxiu Zhou, and Xin Liu. Diaformer: Automatic diagnosis via symptoms sequence generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4432–4440, 2022.

[6] Andrew M. Dai and Quoc V. Le. Semi-supervised sequence learning. In *NIPS*, 2015.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[8] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–495, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL: https://aclanthology.org/P17-1045, doi:10.18653/v1/P17-1045.

[9] Nan Du, Kai Chen, Anjuli Kannan, Linh Tran, Yuhui Chen, and Izhak Shafran. Extracting symptoms and their status from clinical conversations. In *ACL*, 2019.

[10] Nan Du, Mingqiu Wang, Linh Tran, Gang Lee, and Izhak Shafran. Learning to infer entities, properties and their relations from clinical conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4979–4990, Hong Kong, China, November 2019. Association for Computational Linguistics. URL: https://aclanthology.org/D19-1503, doi:10.18653/v1/D19-1503.

[11] Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313, 2022.

[12] Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[13] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240, 2020.

[14] Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. Semi-supervised variational reasoning for medical dialogue generation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 544–554, 2021.

[15] Xiujun Li, Yun-Nung (Vivian) Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. End-to-end task-completion neural dialogue systems. In *IJCNLP*, 2017.

[16] Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle,

United States, July 2022. Association for Computational Linguistics. URL: https://aclanthology.org/2022.naacl-main.15, doi:10.18653/v1/2022.naacl-main.15.

[17] Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. Learning to select knowledge for response generation in dialog systems. In *IJCAI*, 2019.

[18] Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, and Liang Lin. Graph-evolving meta-learning for low-resource medical dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13362–13370, 2021.

[19] Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *EMNLP*, 2019.

[20] Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. Knowledge diffusion for neural dialogue generation. In *ACL*, 2018.

[21] Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1317–1337, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL: https://aclanthology.org/2022.findings-acl.104, doi:10.18653/v1/2022.findings-acl.104.

[22] Hongyin Luo, Shang-Wen Li, and James Glass. Knowledge grounded conversational symptom detection with graph memory networks. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 136–145, Online, November 2020. Association for Computational Linguistics. URL: https://aclanthology.org/2020.clinicalnlp-1.16, doi:10.18653/v1/2020.clinicalnlp-1.16.

[23] Liam McNabb and Robert S Laramee. Survey of surveys (sos)-mapping the landscape of survey papers in information visualization. In *computer graphics forum*, volume 36, pages 589–617. Wiley Online Library, 2017.

[24] Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, pages 1–101, 2022.

[25] Jiahui Qiu, Yangming Zhou, Zhiyuan Ma, Tong Ruan, Jinlin Liu, and Jing Sun. Question answering based clinical text structuring using pre-trained language model. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1596–1600, 2019.

[26] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[27] Igor Shalyminov, Sungjin Lee, Arash Eshghi, and Oliver Lemon. Data-efficient goal-oriented conversation with dialogue knowledge transfer networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1741–1751, Hong Kong, China, November 2019. Association for Computational Linguistics. URL: https://aclanthology.org/D19-1183, doi:10.18653/v1/D19-1183.

[28] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. In *ICML*, 2019.

[29] Qiru Wang and Robert S Laramee. Ehr star: the state-of-the-art in interactive ehr visualization. In *Computer Graphics Forum*, volume 41, pages 69–105. Wiley Online Library, 2022.

[30] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, 2018.

[31] Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353, 2019.

[32] Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. Remedi: Resources for multi-domain, multi-service, medical dialogues. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3013–3024, 2022.

[33] Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, 2020.

[34] Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027, 2020.

[35] Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, Yichen Zhang, Qingyang Wu, Zhou Yu, Kun Xu, Eric Xing, and Pengtao Xie. On the generation of medical dialogs for COVID-19. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 886–896, Online, August 2021. Association for Computational Linguistics. URL: https://aclanthology.org/2021.acl-short.112, doi:10.18653/v1/2021.acl-short.112.