Building Blocks of a Task-Oriented Dialogue System in the Healthcare Domain

Heereen Shim[1,2,3], Dietwig Lowet[1], Bart Vanrumste[2,3] and Stijn Luca[4]
[1]Philips Research, Eindhoven, the Netherlands
[2]Campus Group T, e-Media Research Lab, KU Leuven, Leuven, Belgium
[3]Department of Electrical Engineering (ESAT), STADIUS, KU Leuven, Leuven, Belgium
[4]Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium
{heereen.shim, dietwig.lowet}@philips.com
{bart.vanrumste}@kuleuven.be
{stijn.luca}@ugent.be

Abstract

There has been significant progress in dialogue systems research. However, dialogue systems research in the healthcare domain is still in its infancy. In this paper, we analyse recent studies and outline three building blocks of a task-oriented dialogue system in the healthcare domain: i) privacy-preserving data collection; ii) medical knowledge-grounded dialogue management; and iii) human-centric evaluations. To this end, we propose a framework for developing a dialogue system and show preliminary results of simulated dialogue data generation by utilising expert knowledge and crowdsourcing.

1 Introduction

There has been significant progress in the research field of the dialogue system in past years with the help of large-scale pre-trained language models (LMs) (Vaswani et al., 2017; Radford et al., 2019; Lewis et al., 2020). Pre-trained LMs show a good generalised ability obtained from massive training data collected from the internet and achieve state-of-the-art performance over a wide range of dialogue domains (Zhang et al., 2020). While many studies exist on general purpose dialogues, the research on dialogue systems for healthcare applications is still in its infancy.

There are two major directions in the development of a dialogue system. One direction is to build a chatbot that can have a conversation with a user. This approach mainly focuses on generating appropriate response given user input and dialogue history. Researchers have been working on this direction to create systems to produce more human-like (Adiwardana et al., 2020), consistent (Wolf et al., 2019), and empathetic (Rashkin et al., 2019) responses. The other direction is to build a task-oriented dialogue system that performs a specific task, such as triage or diagnosis within the healthcare domain where researchers focus on developing systems that can detect implicit symptoms or make precise diagnosis/triage result (Middleton et al., 2016; Razzaki et al., 2018; Xu et al., 2019; Wei et al., 2018).

In this study, we consider a dialogue system for a sleep coaching programme for healthy people who would like to optimise their sleep. Motivated by cognitive behaviour therapy for insomnia (CBT-I), we focus on investigating the relationship between how people think, behave, and sleep (Morin et al., 2006). The first step of the coaching programme is a complaints assessment to identify sleep issues and their potential causes and decide the next step (e.g., referring to sleep apnea treatment, providing a sleep education, suggesting a behaviour change programme, etc). During this process, a coaching provider (coach) plays as an active listener, asking questions to probe specific information, while a coaching receiver (user) has more chance to provide complaints and elaborate on these.

Real challenges in the development of a dialogue system, especially a machine learning-based system, come from three fundamental questions: i) how to obtain relevant data; ii) how to develop an automated system; and iii) how to evaluate a system. In this paper, we first analyse existing approaches that address the above questions (Section 2). Then we propose our method to address these questions (Section 3) and show preliminary results and discuss its limitations (Section 4).

The major contributions of this paper are as follows:

- Identifying gaps in existing dialogue systems in the healthcare domain.
- Proposing a framework consisting of three

building blocks.

- Constructing a dataset to illustrate the validity of the proposed method.

2 Related Work

2.1 Data Collection

Obtaining dialogue data is time-consuming and might not be available, especially in the healthcare domain. There are several recent studies on creating a large-scale conversation dataset in the healthcare domain by scrapping dialogues from online websites (Wei et al., 2018; Xu et al., 2019; Zeng et al., 2020). These web-scraping approaches, however, are not scalable and might create potential privacy issues.

To mitigate the scalability issue, some studies leverage domain knowledge to generate simulated dialogue. For example, Liednikova et al. (2020) modelled a typical dialogue flow between doctor-patient in the form of a tree. Then they augmented data by adding similar sentences extracted from an online forum. A drawback of this approach is that access to data sources is required and it might not be available within European countries in the light of the General Data Protection Regulation (GDPR). Contrary to this, Liu et al. (2019) proposed a framework for generating simulated data based on templates, which are logically and clinically verified, and incorporated linguistic knowledge to create diverse augmented data.

Another line of work on collecting dialogue data is to utilise a user simulator. User simulator has been widely used to interact with a dialogue system (Shi et al., 2019). Some of the recent works adapted agenda-based user simulator (Schatzmann and Young, 2009) to create training data for dialogue-based diagnosis systems (Wei et al., 2018; Xu et al., 2019). However, they still utilised web-scrapped data to model user behaviour.

2.2 Dialogue Management

Dialogue management is a component of a dialogue system that processes dialogue context and decides the right next action for the agent to take (Young et al., 2013). For health-related dialogue (e.g., symptom check, triage, diagnosis, etc), the role of dialogue management is to decide what to ask, answer, or inform given the context.

Middleton et al. (2016) casts triage into a sequence of questions and answers. They modelled triage flow as a graph by encoding medical knowledge. This graph plays the role of dialogue management to guide a system to interact with users and make a triage decision. This approach has the following advantages: 1) it alleviates the issue of data collection since they do not rely on machine learning with large-scale data but human expert knowledge; 2) it can reason about its predictions. However, the limitation of this approach is that it requires a lot of expert resources.

Some task-oriented dialogue systems learn how to manage a dialogue flow by reinforcement learning (RL) (Wei et al., 2018; Xu et al., 2019). For example, Wei et al. (2018) framed a dialogue management module as an RL agent with a deep Q-network (Mnih et al., 2015). With this approach, the RL agent can decide the next action (i.e., to inquire about implicit symptoms, to make a diagnosis, etc) based on the current dialogue state. Later, Xu et al. (2019) showed that incorporating a medical knowledge graph and symptom-disease relations can allow an RL agent to ask more relevant implicit symptoms and make a precise diagnosis.

There are also some recent works on developing generative models for an end-to-end dialogue system in the healthcare domain (Liednikova et al., 2020; Zeng et al., 2020) by utilising generative pre-trained LMs (Wolf et al., 2019; Radford et al., 2018, 2019; Lewis et al., 2020; Zhang et al., 2020; Vaswani et al., 2017). However, considering the fact that these generative models are less controllable (Wallace et al., 2019; Sheng et al., 2019), using a pre-trained LM-based generative model for health-related conversation could be risky.

2.3 Evaluation

To evaluate a task-oriented dialogue system, multiple metrics are used; both automatic evaluation metrics and human evaluation metrics. Automatic evaluation metrics include success rate, the average number of turns per dialogue session, matching rate, and average reward for an RL-based system (Li et al., 2017; Wei et al., 2018; Xu et al., 2019). While the automated metrics focus on task completion, human evaluation metrics consider qualitative aspects of the dialogue, such as the quality of dialogue flow, the appropriateness of decision making (diagnosis validity), and dialogue fluency scored by experts (Razzaki et al., 2018; Xu et al., 2019).

However, user perspective has been less considered in evaluating a task-oriented dialogue sys-

tem in healthcare. User-centric metric, such as a user rating score or user preference score (Li et al., 2019), is widely used for evaluating general-purpose dialogue systems (Shi et al., 2019; Shah et al., 2018; Budzianowski and Vulić, 2019; Roller et al., 2020). A user-centric metric can not only be used to assess the performance of a system but debug a system as well. For example, a user might have difficulty understanding the complex language that a system uses or be annoyed by too many questions without a proper explanation. In this case, using proper user-centric metrics can provide an insight into which aspects of a system should be updated.

3 Building Blocks

Here we outline three building blocks of a dialogue system in the healthcare domain and identify open research questions for each building block. To this end, we propose a framework for developing a conversation agent for healthcare-related dialogues.

3.1 Privacy-Preserving Data Collection

As mentioned earlier, the potential privacy issues create challenges in data collection, especially in European countries in the light of GDPR. We identify three potential methods of data collection while safeguarding privacy. The first potential method is to apply appropriate privacy protection techniques to the collected data, such as de-identification that replaces the sensitive information for text (Neamatullah et al., 2008; Meystre et al., 2010; Neubauer and Heurix, 2011). The second potential method is to generate synthetic data by training generative models on the collected data (Guan et al., 2019; Hatua et al., 2019; Pan et al., 2020). The third potential method is to generate simulated data by building a user simulator that can interact with a dialogue system (Wei et al., 2018; Xu et al., 2019; Kao et al., 2018). Applying these three methods, however, entails the following consideration: How much is the risk of information leakage? What is the difference in performance between models trained on de-identified, synthesised, simulated and real data?

3.2 Medical Knowledge-Grounded Dialogue Management

Unlike an open-domain dialogue, healthcare-related dialogue should be grounded in medical knowledge. Two types of knowledge can be included in a dialogue system. The first type of knowledge is the knowledge about dialogue between healthcare professional and healthcare recipient. For example, in the healthcare domain, there exists a typical structure of dialogue that is advised to be followed. Modelling a dialogue structure can guide a system to have an appropriate dialogue flow (Middleton et al., 2016; Razzaki et al., 2018). The second type of knowledge is medical knowledge, including correlations between symptoms and causal relation between symptom and diseases. Incorporating medical knowledge can allow a system to have more appropriate dialogue and make a precise decision (Ni et al., 2017; Ghosh et al., 2018; Chen et al., 2020; Xu et al., 2019). The open questions are: How to efficiently encode expert knowledge into a machine-accessible format (e.g., knowledge graph, knowledge base) and how to incorporate it into a machine learning model? How to maintain the previously built knowledge to keep updated?

3.3 Human-Centric Evaluation

Since a dialogue system is designed to interact with a user, a human evaluation should be is considered as an ideal evaluation. More specifically, two types of human evaluations metrics should be considered to correctly evaluate a dialogue system in the healthcare domain: one from the expert (healthcare professional) perspective and the other from the end-user (healthcare recipient) perspective. Experts from the domain should validate the appropriateness of the dialogue actions made by an agent and assess the quality of the dialogue (Razzaki et al., 2018; Xu et al., 2019). Also, end-user should evaluate a system in terms of satisfaction, usability, and comprehensibility by rating each aspect (Shi et al., 2019; Shah et al., 2018) or deciding the preferred system (Li et al., 2019; Roller et al., 2020). This is associated with the following questions: Which aspects are critical to assess both the functionality and the usability of a system? How can these evaluations be reflected to update a system efficiently?

3.4 A Proposed Framework

Considering the above-mentioned building blocks, we propose a framework for developing a conversational agent in the healthcare domain as illustrated in Figure 1.

Simulated Data Generation The proposed framework generates simulated dialogue data to

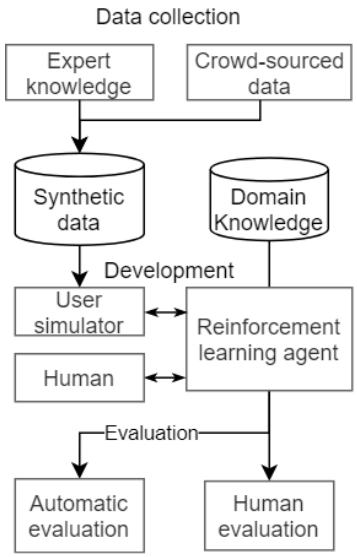


Figure 1: Overview of the proposed framework.

avoid potential privacy issue in data collection. We follow recent works on generating a simulated data set based on the knowledge of user behaviour and the characteristics of dialogue without using real user data (Shah et al., 2018). This consists of two steps: firstly, a template is constructed by exploiting expert knowledge. Secondly, data is augmented by utilising crowdsourcing.

Reinforcement Learning Agent Similar to previous studies (Wei et al., 2018; Xu et al., 2019), we frame a dialogue management module as an RL agent. We propose a two-step training procedure. At the first step, the RL agent is trained with a user simulator, either an agenda-based (Schatzmann and Young, 2009) or a model-based (El Asri et al., 2016; Kreyssig et al., 2018) one. At the second step, the RL agent is further trained by interacting with real-world users.

Model evaluation To evaluate the model, we use both an automatic evaluation metric and a human evaluation metric. Since we consider a task-oriented dialogue system, success rate and matching rate (Xu et al., 2019) are used as automatic metrics. For the human evaluation metric, validity scores by experts (Razzaki et al., 2018) and preference scores by users (Li et al., 2019) are used.

4 Preliminary Results

This section describes an initial approach of generating simulated dialogues based on a template and crowdsourced data. The goal of a dialogue is to assess user complaints related to their sleep and identify all potential behavioural factors that might be associated with the reported complaints.

4.1 Dialogue Template

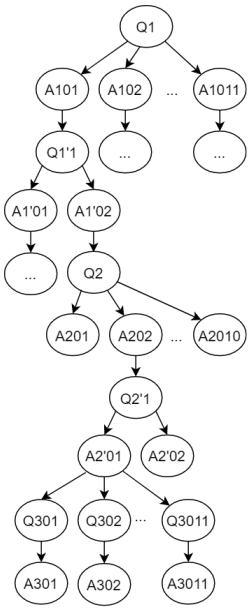
We consulted an expert in the sleep domain to model a dialogue between user and coach in the form of a tree. The dialogue template is structured in three parts of questions and potential answers related to sleep issues, the impacts of sleep issues, and behavioural factors (i.e., habits/lifestyles that might affect sleep quality). More specifically, one open-ended question that is associated with 11 potential answers and two close-ended follow-up questions (i.e., the frequency and the duration of the reported issue) in the sleep issue part, one open-ended question that is associated with 10 potential answers and one close-ended follow-up question (i.e., an enquiry regarding daytime fatigue) in the impact part, and 11 close-ended questions in the behavioural factor part. A subset of the dialogue template and a corresponding dialogue example is shown in Figure 2.

4.2 Crowdsourced Data

Then we collected crowdsourced data via the Amazon MTurk platform. Participants were asked to answer two open-ended questions related to sleep issues and their impacts and check all applicable behavioural factors. Further, the participants are asked to paraphrase the specific sleep conditions (i.e., issues, impacts), if they have ever experienced them, and the selected behavioural factors. The former and the latter data are denoted as the answer data set and the paraphrase data set, respectively. The answer data set are further used to create user goals. Following the previous works (Schatzmann and Young, 2009; Wei et al., 2018; Xu et al., 2019), we create a user goal $G = (E, I)$ consisting of explicit information E, which is reported in the answers to the open-ended questions, and implicit information I, which is the answers to the behavioural factor that can be retrieved via probing questions. Table 1 summarises the size of each data set and the details of each data set are given in Appendix A.

Data set	Goal	Issue	Impact	Habit
Answer	3,015	3,015	3,015	7,961
Paraphrase	-	12,325	7,287	7,961

Table 1: Size of each data set.



(a) Dialogue structure

Coach (Q1)	So, tell me a little bit, what is going on with your sleep?
User (A101)	I lie in bed awake, have trouble falling asleep.
Coach (Q1'1)	How often does it happen? Do you experience that issue more than three times a week?
User (A1'02)	No, less than three times a week.
Coach (Q2)	Tell me how your sleep issues are affecting you?
User (A202)	It affects my performance (e.g. I can't get things done, or I can't deliver the same quality)
Coach (Q2'1)	Do you also experience daytime fatigue?
User (A2'01)	Yes, I feel tired and have less energy or cannot focus.
Coach (Q302)	Do you consume caffeinated drinks, in particular a few hours before going to bed? If so, could you please elaborate it?
User (A302)	I consume caffeinated drinks.

(b) An example of dialogue

Figure 2: A subset of the dialogue template (left) and a corresponding dialogue example (right).

4.3 Dialogue Simulation

The collected crowdsourced data are further used to simulate dialogues. At the beginning of each dialogue, a user goal is sampled from the answer data set. Then a dialogue is simulated based on the dialogue template with a set of handcrafted rules and augmented by using the paraphrase data set. An example of a user goal and the simulated and augmented dialogues are shown in Appendix B.

4.4 Limitations and Future Study

In this paper, we show preliminary results of simulating dialogues based on the dialogue template and crowdsourced data. Our approach aims to augment the size of the simulated dialogue data set by replacing user answers with samples from the separate paraphrase data set. However, there are a few limitations that might be associated with the proposed method. More specifically, the following concerns should be addressed in a future study: First of all, the paraphrased sentences should be diverse and the simulated dialogues should cover all potential dialogue paths. To validate the quality, the paraphrased sentences and the simulated dialogues are required to be accessed by proper measures. Secondly, as Shi et al. (2019) has already pointed out, the RL agent may not generalise enough to real-world dialogues even though it works well with a user simulator. Therefore, there should be the additional step of on-line learning by interacting with real-world users (Shah et al., 2018) to mitigate this issue.

5 Conclusion

In this paper, we analyse recent studies on the development of a dialogue system in the healthcare domain and outline three building blocks, namely: i) privacy-preserving data collection; ii) medical knowledge-grounded dialogue management; and iii) human-centred evaluations. To this end, we propose a framework for developing a dialogue system and show preliminary results of simulated dialogue data generation by utilising expert knowledge and crowdsourcing. In the future study, we foresee working on implementing a user simulator that can interact with a reinforcement learning agent, accessing the quality of the simulated dialogues, and deploying the reinforcement learning agent to interact with both a user simulator and real-world users.

Acknowledgments

We thank anonymous reviewers for providing valuable feedback on this work. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 766139. This article reflects only the author's view and the REA is not responsible for any use that may be made of the information it contains.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2-how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22.

Jun Chen, Xiaoya Dai, Quan Yuan, Chao Lu, and Haifeng Huang. 2020. Towards interpretable clinical diagnosis with bayesian network ensembles stacked on entity-aware cnns. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3143–3153.

Layla El Asri, Jing He, and Kaheer Suleman. 2016. A sequence-to-sequence model for user simulation in spoken dialogue systems. *Interspeech 2016*, pages 1151–1155.

Shameek Ghosh, Sammi Bhatia, and Abhi Bhatia. 2018. Quro: facilitating user symptom check using a personalised chatbot-oriented dialogue system. *Stud Health Technol Inform*, 252:51–56.

Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. 2019. A method for generating synthetic electronic medical record text. *IEEE/ACM transactions on computational biology and bioinformatics*.

Amartya Hatua, Trung T Nguyen, and Andrew H Sung. 2019. Dialogue generation using self-attention generative adversarial network. In *2019 IEEE International Conference on Conversational Data & Knowledge Engineering (CDKE)*, pages 33–38. IEEE.

Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. 2018. Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Florian Kreyssig, Iñigo Casanueva, Paweł Budzianowski, and Milica Gasic. 2018. Neural user simulation for corpus-based policy optimisation of spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 60–69.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 733–743.

Anna Liednikova, Philippe Jolivet, Alexandre Durand-Salmon, and Claire Gardent. 2020. Learning healthbots from training data that was automatically created using paraphrase detection and expert knowledge. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 638–648.

Zhengyuan Liu, Hazel Lim, Nur Farah Ain Suhaimi, Shao Chuen Tong, Sharon Ong, Angela Ng, Sheldon Lee, Michael R Macdonald, Savitha Ramasamy, Pavitra Krishnaswamy, et al. 2019. Fast prototyping a dialogue comprehension system for nurse-patient conversations on symptom monitoring. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 24–31.

Stephane M Meystre, F Jeffrey Friedlin, Brett R South, Shuying Shen, and Matthew H Samore. 2010. Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC medical research methodology*, 10(1):1–16.

Katherine Middleton, Mobasher Butt, Nils Hammerla, Steven Hamblin, Karan Mehta, and Ali Parsa. 2016. Sorting out symptoms: design and evaluation of the'babylon check'automated triage system. *arXiv preprint arXiv:1606.02041*.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Charles M Morin, Richard R Bootzin, Daniel J Buysse, Jack D Edinger, Colin A Espie, and Kenneth L Lichstein. 2006. Psychological and behavioral treatment of insomnia: update of the recent evidence (1998–2004). *Sleep*, 29(11):1398–1414.

Ishna Neamatullah, Margaret M Douglass, H Lehman Li-wei, Andrew Reisner, Mauricio Villarroel, William J Long, Peter Szolovits, George B Moody, Roger G Mark, and Gari D Clifford. 2008. Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1):1–17.

Thomas Neubauer and Johannes Heurix. 2011. A methodology for the pseudonymization of medical data. *International journal of medical informatics*, 80(3):190–204.

Lin Ni, Chenhao Lu, Niu Liu, and Jiamou Liu. 2017. Mandy: Towards a smart primary care chatbot application. In *International symposium on knowledge and systems sciences*, pages 38–52. Springer.

Youcheng Pan, Qingcai Chen, Weihua Peng, Xiaolong Wang, Baotian Hu, Xin Liu, Junying Chen, and Wenxiu Zhou. 2020. Medwriter: Knowledge-aware medical text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2363–2368.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Salman Razzaki, Adam Baker, Yura Perov, Katherine Middleton, Janie Baxter, Daniel Mullarkey, Davinder Sangar, Michael Taliercio, Mobasher Butt, Azeem Majeed, et al. 2018. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. *arXiv preprint arXiv:1806.10698*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Jost Schatzmann and Steve Young. 2009. The hidden agenda user simulation model. *IEEE transactions on audio, speech, and language processing*, 17(4):733–747.

Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gokhan Tur. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Weiyan Shi, Kun Qian, Xuewei Wang, and Zhou Yu. 2019. How to build user simulators to train rl-based dialog systems. 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 1990–2000.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. 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.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. 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.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing

Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

A Crowdsourced Data

We collected two crowdsourced data sets for experiments: The answer data set contains user goals consisting of answers to the two open-ended questions (i.e., sleep issue and the impact of the issue) and one multiple-choice question (i.e., habits/lifestyles). The paraphrase data set contains paraphrased answers related to the sleep conditions (i.e., sleep issue and the impact of the issue) and the selected multiple-choice answers (i.e., habits/lifestyles). The collected data were annotated with class labels as shown in tables 2 to 4. Figure 3 shows label distributions of the collected data sets.

Class	Description
troubleFallingAsleep	Lie in bed awake
troubleStayingAsleep	Wake up frequently
staysUpLate	Stay up late
wakeUpTooEarly	Wake up too early
problemWakingUp	Trouble waking up
sleepsInLater	Sleep in late
snoringBothersMe	Snoring issue 1
snoringBothersOthers	Snoring issue 2
snoringStoppedBreathing	Breathing problem
otherIssue	Other issue
goodSleep	No issue

Table 2: Class labels for sleep issues.

Class	Description
energy	Feel tired or less energy
performance	Affect performance
embarrassedBySnoring	Snoring impact
dryMouth	Cause dry mouth
appearance	Look tired
stressMoodAnxiety	Bad mood
lessPatience	Become less patience
socialImpact	Affect social life
otherHealthImmunity	Affect health
noImpact	No impact

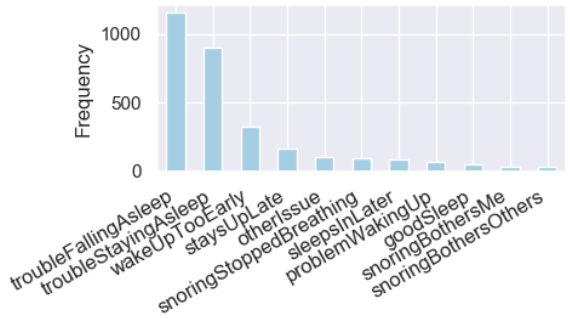
Table 3: Class labels for the impacts of sleep issues.

Class	Description
media	Engage in screen-time
caffeine	Consume caffeine
drinking	Consume drink
alcohol	Consume alcohol drinks
nicotine	Smoke
eating	Eat heavy meals
exercise	Work out/exercise
passivity	Physically not active
napping	Nap during the day
obligationDuties	Too many duties
stressMoodAnxiety	Experience stress

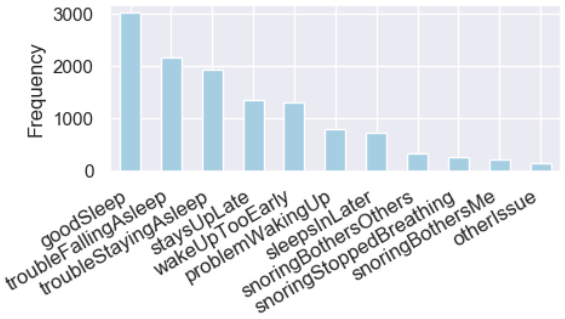
Table 4: Class labels for habits/lifestyles.

B User Goal and Simulated Dialogue

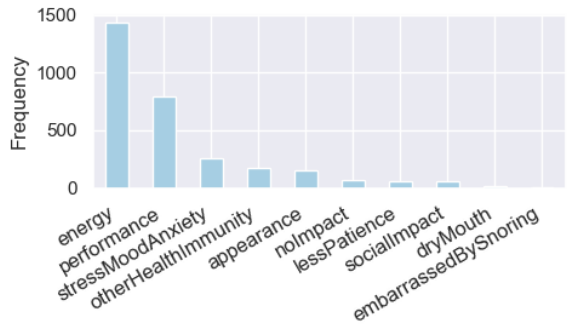
An example of a user goal is shown in Figure 4. To simulate a dialogue, we used the dialogue template with a set of handcrafted rules to select a coach's next question. Each question is followed by the answer by using the sampled user goal. If the question cannot be answered by the user goal, we randomly select an answer either *Yes* or *No*. The simulated dialogue is then paraphrased by replacing user answers with samples from the paraphrase data set. Table 5 illustrates the examples of a simulated dialogue and an augmented dialogue.



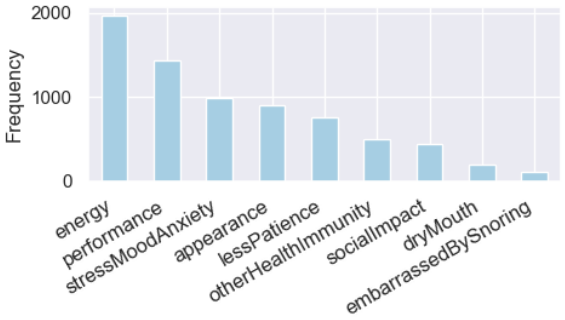
(a) Issue label distribution in the answer data set.



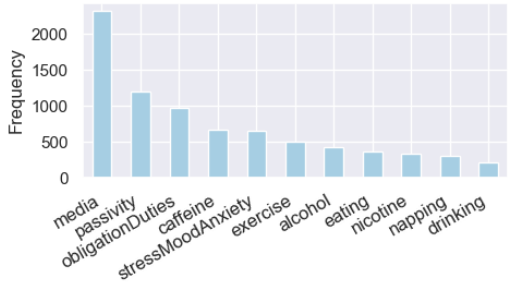
(b) Issue label distribution in the paraphrase data set.



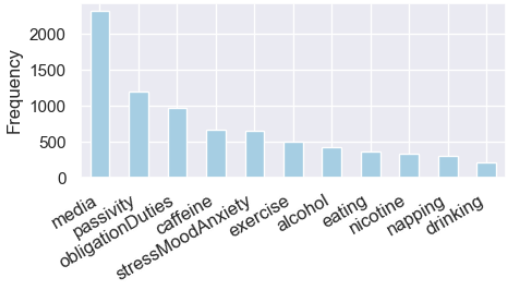
(c) Impact label distribution in the answer data set.



(d) Impact label distribution in the paraphrase data set.



(e) Habit label distribution in the answer data set.



(f) Habit label distribution in the paraphrase data set.

Figure 3: Class label distributions of the collected data sets. Note that the answer data set and the paraphrase data set have identical habit class label distribution but the former contains binary values (i.e., True, False) and the latter contains free-text values (i.e., paraphrased sentences).

	A simulated dialogue
Coach	So, tell me a little bit, what is going on with your sleep?
User	*I just can't get to sleep.*
	I lie in bed awake, have trouble falling asleep.
	I think too much about work issues and need to stop doing that.
	I try to fall asleep, but I just lay there. The sleep doesn't come for me quickly and I have to wait and wait until my body finally falls asleep.
Coach	How often does it happen? Do you experience that issue more than three times a week?
User	Yes.
Coach	How long does your issue last in general? More than 30 minutes?
User	No.
Coach	Tell me how your sleep issues are affecting you?
User	*My exhaustion really affects my work. I'm not sharp like I used to. I feel tortured.*
	I do less because I'm exhausted.
	I need more time to get things done, and I don't have the creativity and energy that I would want to deliver top quality work.
	Because I have not received enough sleep I do not focus as well. This causes my performance to not be as well as it should.
Coach	Do you also experience daytime fatigue?
User	No
Coach	Do you experience stress or mood swings?
User	No
Coach	Do you engage with digital devices/screen, in particular, a few hours before going to bed?
User	Yes
	I'm around screens all the time and it affects my sleep.
	I end up being on my computer working all day and when I'm not working I'm watching TV or on my phone. I do these things immediately before going to bed and while in bed.
	Most of the time leading up to going to bed for us is watching TV. But really this is just about the only time I have to look through facebook, and emails on my phone too. So it's like I'm getting a double whammy of light from these devices.

Table 5: An example of a simulated dialogue based on the dialogue template with a sampled user goal and paraphrased sentences. Italic texts are the source texts extracted from the user goal and underlined italic texts are target sentences sampled from the paraphrased data set. Three randomly sampled paraphrased sentences per user answer are reported.

```
{
 'explicit':{
     'main_issue': 'troubleFallingAsleep
         ',
     'main_issue_text': "I just can't
         get to sleep.",
     'main_impact': 'performance',
     'main_impact_text': "My exhaustion
         really affects my work. I'm not
          sharp like I used to. I feel
         tortured.",
 },
 'implicit': {
     'passivity': False,
     'alcohol': False,
     'nicotine': False,
     'caffeine': False,
     'media': True,
     'exercise': False,
     'drinking': False,
     'eating': False,
     'stressMoodAnxiety': False,
     'obligationDuties': False,
     'napping': False
 }
 }
```

Figure 4: An example of a user goal.