The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications

Anonymous ACL submission

Abstract

Task-oriented dialogue systems in healthcare are attracting increased attention, and have been characterized by a diverse range of architectures and objectives. However, although these systems have been surveyed in the medical community from a non-technical perspective, a systematic review from a rigorous computational perspective remains noticeably absent. As a result, many important implementation details of healthcare-oriented dialogue systems remain limited or under-specified, slowing the pace of innovation in this area. To fill this gap, we investigated an initial pool of 4070 papers from well-known computer science, natural language processing, and artificial intelligence venues, identifying 70 papers that satisfied our defined inclusion criteria. We conducted a comprehensive technical review of the included papers, and present our findings along with identified trends and intriguing directions for future research.

1 Introduction

Dialogue systems are intelligent systems designed to converse with humans via natural language. In recent years, these systems have become omnipresent in many individuals' lives, acting as virtual assistants (Hoy, 2018), customer service agents (Xu et al., 2017), or even companions (Zhou et al., 2020). Generally, dialogue systems fall into one of two broadly defined classes: (1) *chatbots*, which are designed to conduct unstructured conversations in open domains; and (2) *task-oriented dialogue systems*, which help users to complete tasks in a specific domain (Jurafsky and Martin, 2009).

In recent years, task-oriented dialogue systems have attracted increased attention in both academic and industrial communities, manifesting in a wide variety of applications (Qin et al., 2019). These systems have the potential to play an important role in health and medical care (Laranjo et al., 2018), and have been adopted by growing numbers of patients, caregivers, and clinicians as AI continues to advance and high-performance hardware becomes more accessible (Kearns et al., 2019). Nonetheless, although much progress has been made in this domain, there remains a translational gap (Newman-Griffis et al., 2021) between cutting-edge, foundational work in dialogue systems and prototypical or deployed dialogue agents in healthcare settings. This limits the proliferation of valuable scientific findings to real-world systems, in turn constraining the potential benefits of fundamental research.

In this work, we move towards closing this gap by conducting a comprehensive, scientifically rigorous analysis of task-oriented dialogue systems designed exclusively for healthcare applications. Our primary contributions are as follows:

- We perform a systematic search through 4070 papers from well-known technical venues and identify 70 papers about task-oriented dialogue systems in the healthcare domain.[1]
- We analyze these systems according to a wide range of factors, including the domain of research, system objective, target audience, language, architecture, system modality, device type, dataset, and system evaluation methods.
- We identify interesting trends and commonalities among the systems described, and uncover key limitations that may serve as intriguing bases for follow-up work.
- We provide practical future suggestions in an effort to streamline the implementation process for interested researchers.

Importantly, we seek to address the limitations of prior systematic reviews by extensively investigating task-oriented dialogue systems from a computational perspective. In the long term, it is our hope

[1] A full listing of these papers is provided in the appendix.

that this survey can stimulate more rapid advancements in the design of future health-related task-oriented dialogue systems, by identifying promising directions and synthesizing prior findings for researchers and system developers in a large but under-explored body of research.

2 Related Work

Dialogue systems in healthcare have been the focus of several recent surveys conducted by the medical and clinical communities (Vaidyam et al., 2019; Laranjo et al., 2018; Kearns et al., 2019). The objective of these surveys has primarily been to investigate the real-world utilization of deployed systems, rather than examining their design and implementation from a technical perspective. Studies examining health-related task-oriented dialogue systems through the lens of artificial intelligence and natural language processing research and practice have been limited. Zhang et al. (2020) and Chen et al. (2017) presented surveys of recent advances and challenges in task-oriented dialogue systems in the general domain. These surveys provide an excellent portrait of the subfield as a whole, but do not delve into aspects that may be of particular interest in healthcare settings (e.g., considering system objectives that double as clinical goals), limiting their usefulness for this audience.

Vaidyam et al. (2019), Laranjo et al. (2018), and Kearns et al. (2019) conducted informative systematic reviews of chatbots or dialogue systems deployed in mental health (Vaidyam et al., 2019) or general healthcare (Laranjo et al., 2018; Kearns et al., 2019) settings. Vaidyam et al. (2019) examined 10 articles, and Laranjo et al. (2018) and Kearns et al. (2019) examined 17 and 46 articles, respectively; all surveys were written for a medical audience. These works discussed characteristics, current applications, and evaluation measures for conversational agents used in health-related settings. Due largely to their focus and target audience (medical researchers and practitioners), these surveys focused primarily on healthcare issues and impact. The surveys covered few articles from artificial intelligence, natural language processing, or general computer science venues.

Montenegro et al. (2019) and Tudor Car et al. (2020) recently reviewed 40 and 47 articles, respectively, covering conversational agents in the healthcare domain. These two surveys are the closest to ours, but differ in several critical ways. First, our focus is on a specific class of conversational agents: task-oriented dialogue systems. The surveys by Montenegro et al. (2019) and Tudor Car et al. (2020) used a wider search breadth, which proved beneficial for providing a broad, high-level overview, but limited their ability to provide extensive technical depth. We also reviewed more papers (70 articles), which were then screened using a more thorough taxonomy constructed as part of the analysis. Some aspects that we considered that differ from these prior surveys include the overall dialogue system architecture, the dialogue management architecture, the system evaluation methods, and the dataset(s) used when developing and/or evaluating the system.

Screening Process	ACM	IEEE	ACL	AAAI	Total
Initial Search	1050	1400	1020	600	4070
Title Screening	151	273	106	55	585
Abstract Screening	32	45	26	8	110
Final Screening	**21**	**31**	**16**	**2**	**70**

Table 1: The number of papers included from each database in each step of the paper screening process.

3 Search Criteria and Screening

We designed search criteria in concert with our goal of filling a translational information gap between basic and applied dialogue systems in the healthcare domain. To do so, we retrieved articles from well-respected computer science, artificial intelligence, and natural language processing databases and screened them for focus on task-oriented dialogue systems designed for healthcare settings. Specifically, our target databases were: (1) ACM,[2] (2) IEEE,[3] (3) the ACL Anthology,[4] and (4) the AAAI Digital Library.[5] ACM and IEEE are large databases of papers published at prestigious conferences and journals across many computer science fields, including but not limited to robotics, human-computer interaction, data mining, and multimedia systems. The ACL Anthology is the premier

[2]https://dl.acm.org/

[3]https://ieeexplore.ieee.org/Xplore/home.jsp

[4]https://www.aclweb.org/anthology/

[5]https://aaai.org/Library/library.php

database of publications within natural language processing, hosting papers from major conferences (e.g., *ACL* or *EMNLP*) and topic-specific venues (e.g., *SIGDIAL*, organized by the Special Interest Group on Discourse and Dialogue). The AAAI Digital Library hosts papers not only from the *AAAI Conference on Artificial Intelligence*, but also from other AI conferences, *AI Magazine*, and the *Journal of Artificial Intelligence Research*. We applied the following conditions as inclusion criteria when identifying papers:

- The main focus of the article must be on the technical design or implementation of a task-oriented dialogue system.
- The system must be designed for health-related applications.
- The article must *not* be dedicated to one specific module of the system's architecture (e.g., the natural language understanding component of a health-related dialogue system).

We followed four steps in our screening process, outlined as follows:

1. **Initial Search:** We applied a predefined research query to the databases to populate our initial list of papers. To generate the research query, we used the keywords "task-oriented," "dialogue system," "conversational agent," "health," and "healthcare." We also used synonyms and abbreviations of those keywords.
2. **Title Screening:** We performed a preliminary screening through the initial list of papers by reading the titles, keeping those that satisfied the inclusion criteria.
3. **Abstract Screening:** We went through the list of papers remaining after the title screening and read the abstracts, keeping those that satisfied the inclusion criteria.
4. **Final Screening:** We read the body of the papers remaining after the abstract screening and kept those that satisfied the inclusion criteria.

We detail the number of papers remaining after each screening step in Table 1. In total, 70 papers (21 from ACM, 31 from IEEE, 16 from ACL, and 2 from AAAI[6]) satisfied the inclusion criteria and were used for further analysis. We survey papers meeting our inclusion criteria according to a wide range of parameters, including domain of research, system objective, target audience, language, overall and dialogue management architecture, system modality and device type, dataset, and system evaluation measures. We present our findings in the following subsections, grouped into thematic categories: ontology (§4), system architecture (§5), system design (§6), dataset (§7), and system evaluation (§8).

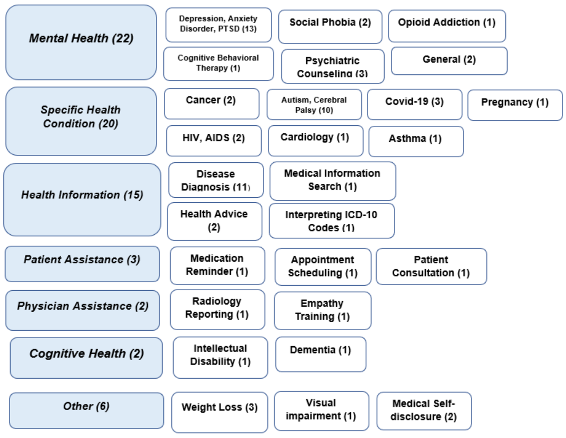

Figure 1: Research domains and corresponding subcategories for the included papers.

4 Ontology

We map each paper to several categories in our ontology, including domain of research (§4.1), system objective (§4.2), target audience (§4.3), and language (§4.4). We present our findings corresponding to each ontological category.

4.1 Domain of Research

Task-oriented dialogue systems offer enormous potential impact on many facets of healthcare in society (Bickmore and Giorgino, 2004). We define a *domain of research* as the healthcare application for which a dialogue system is designed. We identify both broad domains and more specific subcategories thereof, based on the 70 papers surveyed. We outline these domains and corresponding subcategories in Figure 1, along with the number of

[6] Papers about task-oriented dialogue systems published at AAAI often focus on one specific component of the system from a technical perspective, rather than proposing a conversational agent as a whole for a task. Therefore, only two papers from the AAAI Digital Library satisfied the inclusion criteria in this review.

System Objective	# Papers
Diagnosis	7
Monitoring	8
Intervention	13
Counseling	5
Assistance	12
Multi-Objective	25

Table 2: Distribution of system objectives across the surveyed papers. Additional details regarding *multi-objective* papers are provided in the appendix.

Designed for Engagement?	# Papers
Yes	29
No	41

Table 3: Distribution of papers with and without an objective of engagement. The presence of this objective is independent of the primary system objective.

papers belonging to each (in parentheses). Broad domain categories include *mental health*, *specific physical health conditions*, *general health information*, *patient assistance*, *physician assistance*, *cognitive health*, and *other* (comprising several subcategories not easily classifiable to one of the broader domains). Dialogue systems designed for the mental health domain, specific physical health conditions, and general health information proved to be by far the most prevalent, covering a sum total of 57 of the 70 included papers.

4.2 System Objective

Conversational agents seek to generate dialogues that have value to their end-users. We categorized included articles as having one or more of the following objectives:

- **Diagnosis:** The system is designed to diagnose a health condition (e.g., predicting whether the user suffers from cognitive decline).
- **Monitoring:** The system is designed to monitor users' physical, mental, and/or cognitive states (e.g., tracking a user's diet or periodically checking on their mood).
- **Intervention:** The system is designed to address a user's health concern or improve their physical/mental/cognitive state (e.g., teaching children how to map facial expressions to emotions).
- **Counseling:** The system is designed to provide support for users without any direct intervention (e.g., listening to the users' personal, social, or psychological problems and empathizing with them).
- **Assistance:** The system is designed to provide information or guidance to users (e.g., answering questions from users who are filling out forms).
- **Multi-Objective:** The system is designed for more than one of the above objectives.

Table 2 shows the number of papers surveyed having each of the objectives above. We found that many papers (25 of the 70 surveyed) were designed for more than one target objective, and provide additional details in the appendix. Separately, we also considered the role of engagement as an objective of each system. We define the objective of engagement as the act of designing systems that engage users from the specified population in interaction, *with or without* underlying health goals. Engagement may be of particular interest to system designers in healthcare settings since it can be critical in encouraging adoption or adherence with respect to healthcare outcomes (Montenegro et al., 2019). We report our findings in Table 3. Surprisingly, almost 60% of the papers did not focus on designing a dialogue system that specifically sought to engage users in having more interactions.

4.3 Target Audience

When designing any system, narrowing the focus to a core audience helps to develop an effective product (Dell and Kumar, 2016). The final consumers of healthcare systems often fall into three groups: *patients*, *caregivers*, and *clinicians*. Table 4 shows the number of papers focusing on each category. We find that out of 70 task-oriented dialogue systems, 59 are designed specifically for patients.

4.4 Language

Despite remarkable progress in task-oriented dialogue systems in recent years, most such work has been conducted in English and a small set of other high-resource languages (Artetxe et al., 2020). Working on languages beyond English may extend the benefits of health-related dialogue

Target Audience	# Papers
Patients	59
Caregivers	3
Patients & Caregivers	2
Clinicians	11

Table 4: Distribution of the target audiences of the systems described in the surveyed papers.

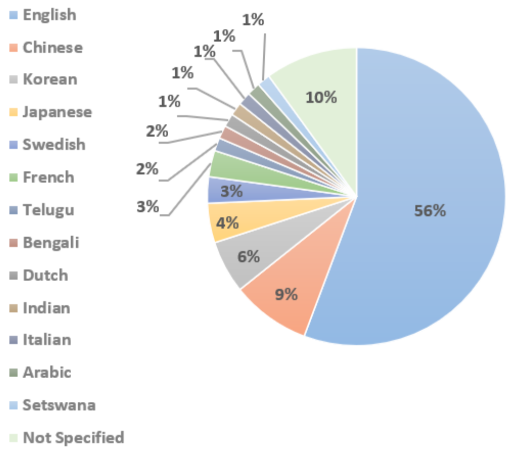

Figure 2: Language diversity across the surveyed systems. A small percentage (10%) of papers do not specify the system's language.

System Architecture	# Papers
Pipeline	58
End-to-End	2
Not Specified	10

Table 5: Distribution of papers describing systems with pipeline or end-to-end architectures, or that do not specify the architecture.

systems more globally. Thus, we investigate language diversity in our systematic review, presenting our findings in Figure 2. As expected, 56% of the systems are designed for English speakers, indicating substantial potential for future growth in generalizing many of these innovations and thereby increasing global access. Encouragingly, several of the included systems did focus on lower-resource languages, including Telugu (Duggenpudi et al., 2019), Bengali (Rahman et al., 2019), and Setswana (Grover et al., 2009).

5 System Architecture

We investigate system architecture from two perspectives. First, we focus on the general architecture of the system as a whole (§5.1), and then if applicable, we examine the architecture of the dialogue management module specifically (§5.2).

5.1 General Architecture

The general architecture of task-oriented dialogue systems often falls into one of two categories: *pipeline* or *end-to-end*. Pipeline architectures typically consist of four key components: *natural language understanding*, *dialogue state tracking*, *dialogue policy learning*, and *natural language generation*. The ensemble of the dialogue state tracking and dialogue policy learning modules is referred to as the *dialogue manager* (Chen et al., 2017). End-to-end architectures use a single encoder-decoder model to train the whole system. This architecture interacts with structured external databases and requires extensive training data (Chen et al., 2017).

We categorized each of the included papers into one of these classes or a third class, "Not Specified," reserved for papers that did not directly specify the general architecture of their developed system. We present our findings in Table 5. Unsurprisingly, only 3% of papers implemented an end-to-end model for their system; this is almost certainly due to the lack of health-related training data in the medical field.

5.2 Dialogue Management Architecture

Dialogue management is an essential component of every pipeline architecture, controlling the dialogue flow and determining which action the system should take next given the current conversation history. We investigated the type of dialogue management architecture in the included papers based on the following classes:

- **Rule-based:** In rule-based approaches, the system interacts with users based on a predefined set of rules. The success of this architecture is conditioned upon its coverage of all relevant cases. Otherwise, the system will not understand the information or intent that the user wants to communicate (Siangchin and Samanchuen, 2019).

- **Intent-based:** Intent-based approaches seek to extract the user's intention from the dialogue, and then perform the relevant action for the user (Jurafsky and Martin, 2009).

- **Hybrid Architecture:** In hybrid architectures, the system is designed using a combina-

Dialogue Management Architecture	# Papers
Rule-based	17
Intent-based	20
Hybrid Architecture	21
Corpus-based	0
Not Applicable	2
Not Specified	10

Table 6: Distribution of dialogue management architectures across the surveyed papers. End-to-end architectures do not have a separate dialogue management module, and are thus listed as *Not Applicable*.

tion of rule-based and intent-based approaches (Jurafsky and Martin, 2009).

- **Corpus-based:** Corpus-based approaches mine the dialogues of human-human conversations and produce responses using retrieval methods (grabbing a response from a corpus) or generative methods (generating a response given the dialogue context) (Jurafsky and Martin, 2009).

When analyzing this component in the included papers, we also add "Not applicable" and "Not Specified" to the above classes. "Not applicable" is assigned to papers that have an end-to-end architecture, and therefore lack a dialogue management module. The results are provided in Table 6. We observe a fairly even mix of rule-based, intent-based, and hybrid architectures.

6 System Design

To evaluate the mechanisms through which humans interact with the surveyed papers, we consider two perspectives: the *modality* through which users interact with the system (§6.1), and the *device* that they use to do so (§6.2).

6.1 Modality

Modality is the mode of sensory input or output used to transfer information between a computer and a human (Karray et al., 2008). The type of modality used can play an important role in dialogue quality and user satisfaction from the interactions (Bilici et al., 2000). We consider the following categories for dialogue system modality:

- **Unimodal:** A system is unimodal if it uses a single modality for information exchange (Karray et al., 2008). The reviewed unimodal dialogue systems in this study belong to one of the following groups:
  - *Text-based interaction*: Users interact with the system by typing.
  - *Spoken interaction*: Users interact with the system by speaking.
  - *Interaction via graphical user interface (GUI)*: Users interact with the system through the use of visual elements.
- **Multimodal:** A system is designated as multimodal if it uses multiple modalities for information exchange (Karray et al., 2008). The reviewed multimodal dialogue systems in this study use a combination of the above unimodal categories.

Unimodal		Multimodal	
Category	**# Papers**	**Category**	**# Papers**
Text	23	Spoken + Text	14
Spoken	25	Spoken + GUI	4
GUI	1	Text + GUI	3

Table 7: Distribution of modality type across the unimodal (49 total, left) and multimodal (21 total, right) systems surveyed.

Multimodal dialogue systems often offer more affordance to users and can result in more robust systems, but implementing a multimodal dialogue system in the medical domain has its own challenges (Sonntag et al., 2009). We find that out of 70 included papers, 49 describe unimodal systems and 21 describe multimodal systems. Table 7 provides more details regarding the distribution of papers in each category.

6.2 Device

Dialogue systems can facilitate interaction between an application and its user via many devices, including mobile and landline telephones and computers (Arora et al., 2013). Traditionally, dialogue systems were linked to telephones to provide a wide range of services (e.g., flight booking (Garvey and Sankaranarayanan, 2012)), but nowadays due to the progress of handheld devices, dialogue systems have increasingly manifested in mobile phones, especially for multimodal systems (McTear, 2010). Conversational agents can also be implemented in the form of avatars that provide lifelike characters

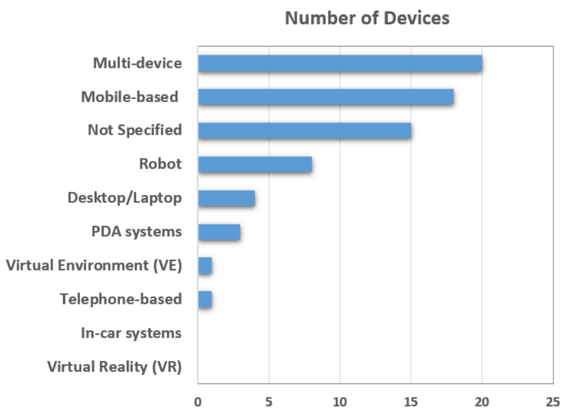

Figure 3: Distribution of device type across the surveyed papers.

Multi-Device Category	# Papers
Desktop/Laptop + Mobile-based	8
Desktop/Laptop + VE	5
Desktop/Laptop + Robot	2
Mobile-based + PDA systems	2
Desktop/Laptop + GUI	1
Desktop/Laptop + PDA systems	1
Mobile-based + VE	1

Table 8: Details regarding the distribution of multi-device systems across the surveyed papers (20 total).

for interaction (Brinkman et al., 2012b; McTear, 2010). When analyzing the included papers in this study, we considered *mobile-based*, *telephone-based*, *desktop/laptop*, *in-car*, *PDA*, *robot-based*, *virtual environment*, and *virtual reality* (including virtual agents and avatars) systems.

We also add one additional category, *multi-device*, to the above labels. Multi-device systems are defined as dialogue systems that use multiple devices for interaction. Figure 3 illustrates the number of papers corresponding to each category. Table 8 provides additional details regarding the multi-device categories. Per the results, multi-device and mobile-based dialogue systems are more popular in the health domain.

7 Dataset

To develop effective dialogue systems that can quickly generate appropriate responses and satisfy user requests without any human intervention, having access to relevant training data is necessary (Serban et al., 2015), and larger quantities of data often lead to better performance. Currently, the dialogue datasets used for training conversational agents are relatively small compared to datasets that are being used for other language-related tasks (Lowe et al., 2017). This is even more pronounced for health-related datasets. It is often hard to access medical data (e.g., corpora of human-human health-care dialogues) due to the risk of data misuse by other parties or the lack of data sharing incentives (Lee and Yoon, 2017).

Knowledge of the underlying data is crucial for developing a full understanding of a system's design and implementation; thus, we checked each included paper for any information regarding the data used during system development. In particular, we focused on dataset size and public data availability, or lack thereof. Unfortunately, out of 70 included papers, only 20 provide details about the quantity and characteristics of the data used (two of the papers provided a link to the dataset, and 18 papers discussed the dataset size).

8 System Evaluation

Finally, a crucial step in developing conversational agents is assessing their performance (Deriu et al., 2019). The ultimate goal when evaluating a dialogue system is to check both its usability and its quality (Hastie, 2012). We broadly categorize the evaluation techniques available for dialogue systems as follows:

- **Human Evaluation:** Prior work on dialogue systems has explored many different approaches to human evaluation. In one popular approach, users are asked to solve a task using a spoken dialogue system and subsequently fill out a questionnaire regarding their experience. In another popular approach, the system is evaluated via feedback from real-world users (Deriu et al., 2019). Broadly speaking, we define human evaluation as any form of evaluation that relies on subjective, first-hand, human user experience.

- **Automated Evaluation:** Automated evaluation provides an objective quantitative measurement of conversational agent quality by analyzing various dimensions of the system from mathematical perspectives (Finch and Choi, 2020). Some of the metrics used for automated evaluation are *BLEU* (Papineni et al., 2002), *Coherence* (Xu et al., 2018), *Entity Accuracy/Recall* (Liu et al., 2018), *Entity Score*

Evaluation Type	# Papers
Human Evaluation	30
Automated Evaluation	8
Human & Automated Evaluation	8
Not Specified	24

Table 9: Distribution of evaluation methods across the surveyed papers.

(Young et al., 2018), *Perplexity* (Chen et al., 2001), and *ROUGE* (Lin, 2004).

We examined how the dialogue systems in each of the included papers were evaluated, and provide our findings in Table 9. We find that nearly half of the papers conducted human evaluations of the described systems; however, a large percentage (34%) did not discuss evaluation at all. In addition to the reported evaluation procedures, we further analyzed papers conducting human evaluations and found that the average number of participants was 26, with a mode of 12 participants.

9 Discussion

When analyzing our findings, several noteworthy trends emerge. First, we found that most task-oriented dialogue systems developed for the health-care domain (83% of surveyed papers) have a pipeline architecture. In pipeline architectures, constituent modules are optimized individually, and the optimization schema does not necessarily improve the overall task performance of the system. In contrast, end-to-end dialogue systems are often trained only on input-output utterances. We speculate that end-to-end architectures could outperform pipeline architectures given sufficient high-quality data, in line with trends seen in other domains, with two caveats: (1) external knowledge sources, a necessary component of many end-to-end architectures, are notoriously complex in many healthcare sub-domains; and (2) for many healthcare applications, interpretable explanations about why the system generated a particular response are critically useful (Ham et al., 2020). Beyond those challenges, developing an end-to-end architecture for task-oriented dialogue systems in the health domain may be further hindered by access limitations to healthcare datasets. A promising future direction could be to generate external health data that could be leveraged in implementing end-to-end architectures. We view these and associated challenges in implementing such systems in healthcare as an intriguing new frontier in translational dialogue systems research.

Additionally, we observed that the target audience of most systems (56% of surveyed papers) in the health domain are English speakers. While developing multilingual dialogue systems, or systems for speakers of low-resource languages specifically, brings up various challenges (López-Cózar Delgado and Araki, 2005), we believe solving this problem could have have tremendous benefit for overburdened healthcare workers in non-English speaking communities, as well as for patients in non-English speaking communities with minimal or unreliable healthcare access. The systems developed by Duggenpudi et al. (2019), Rahman et al. (2019), and Grover et al. (2009) provide case examples for how such systems may be implemented.

Finally, while conducting this systematic review, we also observed that many papers lack important implementation details such as the characteristics of the dataset (71%) and the evaluation methods (34%). This prevents the research community from replicating developed systems and generalizing study findings more broadly. As replication is a crucial part of the scientific process (Walker et al., 2018), we urge researchers in this domain to provide implementation details in their publications and supplemental documentation.

10 Conclusion

In this work, we conducted a systematic technical survey of task-oriented dialogue systems used for health-related purposes, providing much-needed analyses from a computational perspective and narrowing the translational gap between basic and applied dialogue systems research. We comprehensively searched through 4070 papers in computer science, natural language processing, and artificial intelligence databases, finding 70 papers that satisfied our inclusion criteria. We analyzed these papers based on numerous aspects including the domain of research, system objective, target audience, language, system architecture, system design, training dataset, and evaluation methods. It is our hope that interested researchers find the information provided in this review to be a unique and helpful resource for developing task-oriented dialogue systems for healthcare applications.

References

Parham Aarabi. 2013. Virtual cardiologist — a conversational system for medical diagnosis. In *2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 1–4.

Yuna Ahn, Yilin Zhang, Yujin Park, and Joonhwan Lee. 2020. A chatbot solution to chat app problems: Envisioning a chatbot counseling system for teenage victims of online sexual exploitation. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–7, New York, NY, USA. Association for Computing Machinery.

Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K. Schubert, and Ehsan Hoque. 2020. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

Mohammad Rafayet Ali, Taylan Sen, Benjamin Kane, Shagun Bose, Thomas Carroll, Ronald Epstein, Lenhart K. Schubert, and Ehsan Hoque. 2021. Novel computational linguistic measures, dialogue system and the development of sophie: Standardized online patient for healthcare interaction education. *IEEE Transactions on Affective Computing*, pages 1–1.

Masahiro Araki, Kana Shibahara, and Yuko Mizukami. 2011. Spoken dialogue system for learning braille. In *2011 IEEE 35th Annual Computer Software and Applications Conference*, pages 152–156.

Suket Arora, Kamaljeet Batra, and Sarabjit Singh. 2013. Dialogue system: A brief review. *CoRR*, abs/1306.4134.

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

Lekha Athota, Vinod Kumar Shukla, Nitin Pandey, and Ajay Rana. 2020. Chatbot for healthcare system using artificial intelligence. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, pages 619–622.

Saminda Sundeepa Balasuriya, Laurianne Sitbon, Andrew A. Bayor, Maria Hoogstrate, and Margot Brereton. 2018. Use of voice activated interfaces by people with intellectual disability. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, OzCHI '18, page 102–112, New York, NY, USA. Association for Computing Machinery.

R. V. Belfin, A. J. Shobana, Megha Manilal, Ashly Ann Mathew, and Blessy Babu. 2019. A graph based chatbot for cancer patients. In *2019 5th International Conference on Advanced Computing Communication Systems (ICACCS)*, pages 717–721.

Timothy Bickmore and Toni Giorgino. 2004. Some novel aspects of health communication from a dialogue systems perspective. *AAAI Fall Symposium - Technical Report*.

Vildan Bilici, Emiel Krahmer, Saskia Riele, and Raymond Veldhuis. 2000. Preferred modalities in dialogue systems.

Willem-Paul Brinkman, Dwi Hartanto, Ni Kang, Daniel de Vliegher, Isabel L. Kampmann, Nexhmedin Morina, Paul G.M. Emmelkamp, and Mark Neerincx. 2012a. A virtual reality dialogue system for the treatment of social phobia. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, page 1099–1102, New York, NY, USA. Association for Computing Machinery.

Willem-Paul Brinkman, Dwi Hartanto, Ni Kang, Daniel Vliegher, Isabel Kampmann, Nexhmedin Morina, Paul Emmelkamp, and Mark Neerincx. 2012b. A virtual reality dialogue system for the treatment of social phobia. pages 1099–1102.

Jacqueline Brixey, Rens Hoegen, Wei Lan, Joshua Rusow, Karan Singla, Xusen Yin, Ron Artstein, and Anton Leuski. 2017. SHIHbot: A Facebook chatbot for sexual health information on HIV/AIDS. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 370–373, Saarbrücken, Germany. Association for Computational Linguistics.

Leonardo Campillos Llanos, Dhouha Bouamor, Éric Bilinski, Anne-Laure Ligozat, Pierre Zweigenbaum, and Sophie Rosset. 2015. Description of the Patient-Genesys dialogue system. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 438–440, Prague, Czech Republic. Association for Computational Linguistics.

Bo-Wei Chen, Po-Yi Shih, Karunanithi Bharanitharan, Po-Chuan Lin, Jhing-Fa Wang, and Chia-Ming Chen. 2013. Customizable cloud-healthcare dialogue system based on lvcsr with prosodic-contextual post-processing. In *2013 1st International Conference on Orange Technologies (ICOT)*, pages 246–249.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *CoRR*, abs/1711.01731.

Stanley Chen, Douglas Beeferman, and Ronald Rosenfeld. 2001. Evaluation metrics for language models.

Ching-Hua Chuan and Susan Morgan. 2021. Creating and evaluating chatbots as eligibility assistants for clinical trials: An active deep learning approach towards user-centered classification. *ACM Trans. Comput. Healthcare*, 2(1).

Karl Daher, Jacky Casas, Omar Abou Khaled, and Elena Mugellini. 2020. Empathic chatbot response for medical assistance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20, New York, NY, USA. Association for Computing Machinery.

Prathyusha Danda, Brij Mohan Lal Srivastava, and Manish Shrivastava. 2016. Vaidya: A spoken dialog system for health domain. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 161–166, Varanasi, India. NLP Association of India.

Johan Oswin De Nieva, Jose Andres Joaquin, Chaste Bernard Tan, Ruzel Khyvin Marc Te, and Ethel Ong. 2020. Investigating students' use of a mental health chatbot to alleviate academic stress. In *6th International ACM In-Cooperation HCI and UX Conference*, CHIuXiD '20, page 1–10, New York, NY, USA. Association for Computing Machinery.

Nicola Dell and Neha Kumar. 2016. The ins and outs of hci for development. pages 2220–2232.

Orianna Demasi, Yu Li, and Zhou Yu. 2020. A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3623–3636, Online. Association for Computational Linguistics.

Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. Survey on evaluation methods for dialogue systems. *CoRR*, abs/1905.04071.

David DeVault, Kallirroi Georgila, Ron Artstein, Fabrizio Morbini, David Traum, Stefan Scherer, Albert Skip Rizzo, and Louis-Philippe Morency. 2013. Verbal indicators of psychological distress in interactive dialogue with a virtual human. In *Proceedings of the SIGDIAL 2013 Conference*, pages 193–202, Metz, France. Association for Computational Linguistics.

Alessandro Di Nuovo, Josh Bamforth, Daniela Conti, Karen Sage, Rachel Ibbotson, Judy Clegg, Anna Westaway, and Karen Arnold. 2020. An explorative study on robotics for supporting children with autism spectrum disorder during clinical procedures. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 189–191, New York, NY, USA. Association for Computing Machinery.

Francesca Dino, Rohola Zandie, Hojjat Abdollahi, Sarah Schoeder, and Mohammad H. Mahoor. 2019. Delivering cognitive behavioral therapy using a conversational social robot. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2089–2095.

Suma Reddy Duggenpudi, Kusampudi Siva Subrahamanyam Varma, and Radhika Mamidi. 2019. Samvaadhana: A Telugu dialogue system in hospital domain. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 234–242, Hong Kong, China. Association for Computational Linguistics.

Wilmer Stalin Erazo, Germán Patricio Guerrero, Carlos Carrión Betancourt, and Iván Sánchez Salazar. 2020. Chatbot implementation to collect data on possible covid-19 cases and release the pressure on the primary health care system. In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0302–0307.

Ahmed Fadhil and Ahmed Ghassan Tawfiq AbuRa'ed. 2019. Ollobot - towards a text-based arabic health conversational agent: Evaluation and results. In *RANLP*.

Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 236–245, 1st virtual meeting. Association for Computational Linguistics.

Floyd Garvey and Suresh Sankaranarayanan. 2012. Intelligent agent based flight search and booking system. *International Journal of Advanced Research in Artificial Intelligence*, 1(4).

Nancy Green, William Lawton, and Boyd Davis. 2004. An assistive conversation skills training system for caregivers of persons with alzheimer's disease. In *Proceedings of the AAAI 2004 Fall Symposium on Dialogue Systems for Health Communication*.

Aditi Sharma Grover, Madelaine Plauché, Etienne Barnard, and Christiaan Kuun. 2009. Hiv health information access using spoken dialogue systems: Touchtone vs. speech. In *Proceedings of the 3rd International Conference on Information and Communication Technologies and Development*, ICTD'09, page 95–107. IEEE Press.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online. Association for Computational Linguistics.

Helen Hastie. 2012. *Metrics and Evaluation of Spoken Dialogue Systems*, pages 131–150.

Samuel Holmes, Anne Moorhead, Raymond Bond, Huiru Zheng, Vivien Coates, and Michael Mctear. 2019. Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, ECCE 2019, page 207–214, New York, NY, USA. Association for Computing Machinery.

Matthew B. Hoy. 2018. Alexa, siri, cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1):81–88. PMID: 29327988.

Chin-Yuan Huang, Ming-Chin Yang, Chin-Yu Huang, Yu-Jui Chen, Meng-Lin Wu, and Kai-Wen Chen. 2018. A chatbot-supported smart wireless interactive healthcare system for weight control and health promotion. In *2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, pages 1791–1795.

Tae-Ho Hwang, JuHui Lee, Se-Min Hyun, and KangYoon Lee. 2020. Implementation of interactive healthcare advisor model using chatbot and visualization. In *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 452–455.

Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. 2016. Talking with ERICA, an autonomous android. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 212–215, Los Angeles. Association for Computational Linguistics.

Hifza Javed, Myounghoon Jeon, Ayanna Howard, and Chung Hyuk Park. 2018. Robot-assisted socio-emotional intervention framework for children with autism spectrum disorder. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 131–132, New York, NY, USA. Association for Computing Machinery.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Dipesh Kadariya, Revathy Venkataramanan, Hong Yung Yip, Maninder Kalra, Krishnaprasad Thirunarayanan, and Amit Sheth. 2019. kbot: Knowledge-enabled personalized chatbot for asthma self-management. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 138–143.

Takeshi Kamita, Atsuko Matsumoto, Boyu Sun, and Tomoo Inoue. 2020. Promotion of continuous use of a self-guided mental healthcare system by a chatbot. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, CSCW '20 Companion, page 293–298, New York, NY, USA. Association for Computing Machinery.

B. Amir H. Kargar and Mohammad H. Mahoor. 2017. A pilot study on the ebear socially assistive robot: Implication for interacting with elderly people with moderate depression. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 756–762.

Fakhri Karray, Milad Alemzadeh, Jamil Saleh, and Mo Nours Arab. 2008. Human-computer interaction: Overview on state of the art. *International Journal on Smart Sensing and Intelligent Systems*, 1:137–159.

William Kearns, Nai-Ching Chi, Yong Choi, Shih-Yin Lin, Hilaire Thompson, and George Demiris. 2019. A systematic review of health dialog systems. *Methods of Information in Medicine*, 58:179–193.

Liliana Laranjo, Adam Dunn, Huong Ly Tong, A. Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie Lau, and Enrico Coiera. 2018. Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 0.

Choong Lee and Hyung-Jin Yoon. 2017. Medical big data: promise and challenges. *Kidney Research and Clinical Practice*, 36:3–11.

Dongkeon Lee, Kyo-Joong Oh, and Ho-Jin Choi. 2017. The chatbot feels you - a counseling service using emotional response generation. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 437–440.

Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020a. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).

Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020b. "i hear you, i feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. Association for Computational Linguistics.

Peter Ljunglöf, Britt Claesson, Ingrid Mattsson Müller, Stina Ericsson, Cajsa Ottesjö, Alexander Berman, and Fredrik Kronlid. 2011. Lekbot: A talking and playing robot for children with disabilities. In

Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies, pages 110–119, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Peter Ljunglöf, Staffan Larsson, Katarina Heimann Mühlenbock, and Gunilla Thunberg. 2009. TRIK: A talking and drawing robot for children with communication disabilities. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 275–278, Odense, Denmark. Northern European Association for Language Technology (NEALT).

A. Loisel, N. Chaignaud, and J-Ph. Kotowicz. 2007. Designing a human-computer dialog system for medical information search. In *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*, pages 350–353.

Ramón López-Cózar Delgado and Masahiro Araki. 2005. *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment*. Wiley, Chichester, UK.

Ryan Lowe, Nissan Pow, Iulian Serban, Laurent Charlin, Chia-Wei Liu, and Joelle Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue and Discourse*, 8:31–65.

Raju Maharjan, Per Bækgaard, and Jakob E. Bardram. 2019. "hear me out": Smart speaker based conversational agent to monitor symptoms in mental health. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '19 Adjunct, page 929–933, New York, NY, USA. Association for Computing Machinery.

Rohit Binu Mathew, Sandra Varghese, Sera Elsa Joy, and Swanthana Susan Alex. 2019. Chatbot for disease prediction and treatment recommendation using machine learning. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 851–856.

Michael McTear. 2010. Chapter 9 - the role of spoken dialogue in user–environment interaction. In Hamid Aghajan, Ramón López-Cózar Delgado, and Juan Carlos Augusto, editors, *Human-Centric Interfaces for Ambient Intelligence*, pages 225–254. Academic Press, Oxford.

Mahdi Naser Moghadasi, Yu Zhuang, and Hashim Gellban. 2020. Robo: A counselor chatbot for opioid addicted patients. In *2020 2nd Symposium on Signal Processing Systems*, SSPS 2020, page 91–95, New York, NY, USA. Association for Computing Machinery.

Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications*, 129:56–67.

Fabrizio Morbini, David DeVault, Kallirroi Georgila, Ron Artstein, David Traum, and Louis-Philippe Morency. 2014. A demonstration of dialogue processing in SimSensei kiosk. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 254–256, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Fabrizio Morbini, Eric Forbell, David DeVault, Kenji Sagae, David Traum, and Albert Rizzo. 2012. A mixed-initiative conversational dialogue system for healthcare. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 137–139, Seoul, South Korea. Association for Computational Linguistics.

Denis Newman-Griffis, Jill Fain Lehman, Carolyn Rosé, and Harry Hochheiser. 2021. Translational NLP: A new paradigm and general principles for natural language processing research. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4125–4138, Online. Association for Computational Linguistics.

Kyo-Joong Oh, Dongkun Lee, Byungsoo Ko, and Ho-Jin Choi. 2017. A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *2017 18th IEEE International Conference on Mobile Data Management (MDM)*, pages 371–375.

Alexandros Papangelis, Robert Gatchel, Vangelis Metsis, and Fillia Makedon. 2013. An adaptive dialogue system for assessing post traumatic stress disorder. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '13, New York, NY, USA. Association for Computing Machinery.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation.

Falguni Patel, Riya Thakore, Ishita Nandwani, and Santosh Kumar Bharti. 2019. Combating depression in students using an intelligent chatbot: A cognitive behavioral therapy. In *2019 IEEE 16th India Council International Conference (INDICON)*, pages 1–4.

Frano Petric, Damjan Miklic, and Zdenko Kovacic. 2017. Robot-assisted autism spectrum disorder diagnostics using pomdps. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 369–370, New York, NY, USA. Association for Computing Machinery.

Marco Polignano, Fedelucio Narducci, Andrea Iovine, Cataldo Musto, Marco De Gemmis, and Giovanni Semeraro. 2020. Healthassistantbot: A personal health assistant for the italian language. *IEEE Access*, 8:107479–107497.

A. Prange, Margarita Chikobava, P. Poller, Michael Barz, and D. Sonntag. 2017. A multimodal dialogue system for medical decision support inside virtual reality. In *SIGDIAL Conference*.

Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. 2019. Entity-consistent end-to-end task-oriented dialogue system with kb retriever.

Juan C. Quiroz, Tristan Bongolan, and Kiran Ijaz. 2020. Alexa depression and anxiety self-tests: A preliminary analysis of user experience and trust. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp-ISWC '20, page 494–496, New York, NY, USA. Association for Computing Machinery.

Md. Moshiur Rahman, Ruhul Amin, Md Nazmul Khan Liton, and Nahid Hossain. 2019. Disha: An implementation of machine learning based bangla healthcare chatbot. In *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, pages 1–6.

Nudtaporn Rosruen and Taweesak Samanchuen. 2018. Chatbot utilization for medical consultant system. In *2018 3rd Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, pages 1–5.

Sanket Sanjay Sadavarte and Eliane Bodanese. 2019. Pregnancy companion chatbot using alexa and amazon web services. In *2019 IEEE Pune Section International Conference (PuneCon)*, pages 1–5.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *CoRR*, abs/1512.05742.

Bhuvan Sharma, Harshita Puri, and Deepika Rawat. 2018. Digital psychiatry - curbing depression using therapy chatbot and depression analysis. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 627–631.

Tianhao She, Xin Kang, Shun Nishide, and Fuji Ren. 2018. Improving leo robot conversational ability via deep learning algorithms for children with autism. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 416–420.

Naohiro Shoji, Takayo Namba, and Keiichi Abe. 2020. Proposal of spoken interactive home doctor system for elderly people. In *2020 IEEE 9th Global Conference on Consumer Electronics (GCCE)*, pages 421–423.

Noppon Siangchin and Taweesak Samanchuen. 2019. Chatbot implementation for icd-10 recommendation system. In *2019 International Conference on Engineering, Science, and Industrial Applications (ICESI)*, pages 1–6.

Daneil Sonntag and Manuel Moller. 2010. Prototyping semantic dialogue systems for radiologists. In *2010 Sixth International Conference on Intelligent Environments*, pages 84–89.

Daniel Sonntag, Gerhard Sonnenberg, Robert Neßelrath, and Gerd Herzog. 2009. Supporting a rapid dialogue engineering process.

Prakhar Srivastava and Nishant Singh. 2020. Automatized medical chatbot (medibot). In *2020 International Conference on Power Electronics IoT Applications in Renewable Energy and its Control (PARC)*, pages 351–354.

Bo-Hao Su, Shih-Pang Tseng, Yu-Shan Lin, and Jhing-Fa Wang. 2018. Health care spoken dialogue system for diagnostic reasoning and medical product recommendation. In *2018 International Conference on Orange Technologies (ICOT)*, pages 1–4.

Konstantinos Tsiakas, Lynette Watts, Cyril Lutterodt, Theodoros Giannakopoulos, Alexandros Papangelis, Robert Gatchel, Vangelis Karkaletsis, and Fillia Makedon. 2015. A multimodal adaptive dialogue manager for depressive and anxiety disorder screening: A wizard-of-oz experiment. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '15, New York, NY, USA. Association for Computing Machinery.

Lorainne Tudor Car, Dhakshenya Ardhithy Dhinagaran, Bhone Myint Kyaw, Tobias Kowatsch, Shafiq Joty, Yin-Leng Theng, and Rifat Atun. 2020. Conversational agents in health care: Scoping review and conceptual analysis. *J Med Internet Res*, 22(8):e17158.

A. Vaidyam, Hannah Wisniewski, J. Halamka, M. S. Kashavan, and J. Torous. 2019. Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64:456 – 464.

Richard M Walker, Gene A Brewer, M Jin Lee, Nicolai Petrovsky, and Arjen van Witteloostuijn. 2018. Best Practice Recommendations for Replicating Experiments in Public Administration. *Journal of Public Administration Research and Theory*, 29(4):609–626.

Jinping Wang, Hyun Yang, Ruosi Shao, Saeed Abdullah, and S. Shyam Sundar. 2020. Alexa as coach: Leveraging smart speakers to build social agents that reduce public speaking anxiety. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

J. V. Waterschoot, Iris Hendrickx, Arif Khan, E. Klabbers, M. D. Korte, H. Strik, C. Cucchiarini, and

M. Theune. 2020. Bliss: An agent for collecting spoken dialogue data about health and well-being. In *LREC*.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying 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, Melbourne, Australia. Association for Computational Linguistics.

Charles Welch, Allison Lahnala, Veronica Perez-Rosas, Siqi Shen, Sarah Seraj, Larry An, Kenneth Resnicow, James Pennebaker, and Rada Mihalcea. 2020. Expressive interviewing: A conversational system for coping with COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510.

L. Xu, Q. Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. *ArXiv*, abs/1901.10623.

Xinnuo Xu, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018. Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3981–3991, Brussels, Belgium. Association for Computational Linguistics.

Keigo Yabuki and Kaoru Sumi. 2018. Learning support system for effectively conversing with individuals with autism using a humanoid robot. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 4266–4270.

Akihiro Yorita, Simon Egerton, Carina Chan, and Naoyuki Kubota. 2020. Chatbot for peer support realization based on mutual care. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1601–1606.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. 32(1).

Zheng Zhang, Ryuichi Takanobu, Minlie Huang, and Xiaoyan Zhu. 2020. Recent advances and challenges in task-oriented dialog system. *CoRR*, abs/2003.07490.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Multi-Objective System	**# Papers**
Diagnosis + Assistance	7
Diagnosis + Intervention	2
Diagnosis + Monitoring	1
Diagnosis + Counseling	1
Intervention + Monitoring	2
Intervention + Assistance	1
Assistance + Counseling	2
Intervention + Monitoring + Diagnosis	2
Intervention + Monitoring + Assistance	2
Intervention + Monitoring + Counseling	1
Diagnosis + Monitoring + Counseling	1
Diagnosis + Assistance + Intervention	2
Diagnosis + Intervention + Monitoring + Assistance	1

Table 10: Distribution of varying combinations of multiple system objectives across the surveyed papers.

A Multi-Objective Systems

Conversational agents seek to generate dialogues that have value to their end-users. We categorized included articles as having one or more of the following objectives:

- **Diagnosis:** The system is designed to diagnose a health condition (e.g., predicting whether the user suffers from cognitive decline).
- **Monitoring:** The system is designed to monitor users' physical, mental, and/or cognitive states (e.g., tracking a user's diet or periodically checking on their mood).
- **Intervention:** The system is designed to address a user's health concern or improve their physical/mental/cognitive state (e.g., teaching children how to map facial expressions to emotions).
- **Counseling:** The system is designed to provide support for users without any direct intervention (e.g., listening to the users' personal, social, or psychological problems and empathizing with them).
- **Assistance:** The system is designed to provide information or guidance to users (e.g., answering questions from users who are filling out forms).

- **Multi-Objective:** The system is designed for more than one of the above objectives.

In this survey, 25 out of 70 included articles were designed for more than one target objective. We provide additional details describing these multi-objective systems in Table 10.

B Included Papers

In this systematic review, we investigated 4070 papers involving dialogue systems for healthcare applications, identifying 70 papers that satisfied our defined inclusion criteria. We comprehensively analyzed these papers on the basis of their domain of research, system objective, target audience, language, architecture, modality, device type, data, and evaluation methods. We provide aggregated statistics for each of these categories in the main body of the paper. In Table 11 beginning on the following page, we provide a listing of each included paper and its categorization across all included classes. Full references for each included paper can be found in the bibliography.

Paper	DS Arch.	DM Arch.	Mod.	Device	Sys. Obj.	Engagement	Dom. of Research	Target Aud.	Lang.	Eval. Method	Dataset Size
Papangelis et al. (2013)	Pipeline	Intent-based	Multi-Modal	Desk /Lap	Monitoring, Intervention, Diagnosis	Yes	PTSD	Patients	English	Not Specified	Not Specified
Brinkman et al. (2012a)	Pipeline	Rule-based	Speech	Virtual Environment	Monitoring, Diagnosis	No	Social Phobia	Clinicians	English	Human Evaluation	Not Specified
Ali et al. (2020)	Pipeline	Intent-based	Speech	Desk /Lap	Monitoring, Assistance, Intervention	Yes	Autism Spectrum Disorder	Patients	English	Human Evaluation	46 videos
Tsiakas et al. (2015)	Pipeline	Intent-based	Multi-Modal	Desk /Lap, Virtual Environment	Diagnosis, Assistance	Yes	Anxiety Disorders, Depression, PTSD	Patients	English	Human Evaluation	90 speech segments
Wang et al. (2020)	Pipeline	Hybrid	Speech	PDA	Intervention	Yes	Social Phobia	Patients	English	Human Evaluation	Not Specified
Balasuriya et al. (2018)	Pipeline	Hybrid	Speech, GUI	PDA	Monitoring	Yes	Intellectual Disability	Patients	English	Human Evaluation	Not Specified
Chuan and Morgan (2021)	Pipeline	Intent-based	Speech	Desk /Lap	Assistance	No	Clinical Application	Patients	English	Human Evaluation	Not Specified
Grover et al. (2009)	Pipeline	Rule-based	Speech	Telephone	Assistance	No	HIV	Clinicians	Setswana	Human & Automated Evaluation	Not Specified
Petric et al. (2017)	Pipeline	Intent-based	Speech	Robot	Diagnosis	No	Autism Spectrum Disorder	Clinicians	English	Human Evaluation	Not Specified
Javed et al. (2018)	Not Specified	Not Specified	Speech, GUI	Robot	Monitoring	Yes	Autism Spectrum Disorder	Patients	English	Human Evaluation	Not Specified

Di Nuovo et al. (2020)	Not Specified	Not Specified	Speech	Robot	Monitoring	Yes	Autism Spectrum Disorder	Patients, Caregivers	English	Human Evaluation	Not Specified
Quiroz et al. (2020)	Pipeline	Hybrid	Speech	PDA, Mobile	Diagnosis, Intervention	Yes	Depression, Anxiety	Patients	English	Human Evaluation	Not Specified
Maharjan et al. (2019)	Pipeline	Hybrid	Speech	PDA, Mobile	Monitoring	No	Mental Health	Patients	English	Not Specified	Not Specified
Ahn et al. (2020)	Not Specified	Not Specified	Text	Mobile	Intervention, Assistance	Yes	Online sexual exploitation, PTSD	Patients	Korean	Not Specified	Not Specified
Kamita et al. (2020)	Not Specified	Not Specified	Text	Mobile	Intervention	Yes	Cognitive Behavioral Therapy, stress reduction	Patients	Japanese	Human Evaluation	Not Specified
Lee et al. (2020b)	Pipeline	Hybrid	Speech	Mobile	Monitoring	Yes	Health-related Self-disclosure	Patients	English	Human Evaluation	Not Specified
Moghadasi et al. (2020)	Pipeline	Hybrid	Text	Desk /Lap, Mobile	Assistance, Counseling	No	Opioid Addiction	Patients	English	Not Specified	20,494 records
De Nieva et al. (2020)	Pipeline	Hybrid	Text	Mobile	Monitoring, Intervention, Counseling	Yes	Anxiety, Depression	Patients	English	Human & Automated Evaluation	Not Specified
Lee et al. (2020a)	Pipeline	Hybrid	Text	Mobile	Monitoring	Yes	Health-related Self-disclosure	Patients	English	Human Evaluation	Not Specified
Daher et al. (2020)	Pipeline	Rule-based	GUI	Not Specified	Monitoring	No	Empathy for medical Assistance	Patients	English	Human Evaluation	Not Specified
Holmes et al. (2019)	Pipeline	Hybrid	Multi-Modal	Mobile	Assistance	Yes	Weight Loss	Patients	English	Human & Automated Evaluation	Not Specified

Oh et al. (2017)	Pipeline	Intent-based	Multi-Modal	Mobile	Diagnosis, Monitoring, Intervention	Yes	Psychiatric Counseling	Patients	Korean	Not Specified	49,846,477 records
Dino et al. (2019)	Pipeline	Rule-based	Speech	Robot	Intervention	Yes	Depression	Patients	English	Human Evaluation	Not Specified
Patel et al. (2019)	Not Specified	Not Specified	Text	Not Specified	Diagnosis	No	Stress, Depression	Patients	English	Not Specified	7,652 records, ISEAR dataset
Sharma et al. (2018)	Not Specified	Not Specified	Text	Mobile	Diagnosis, Intervention, Assistance	No	Depression	Patients	Not Specified	Not Specified	Not Specified
Belfin et al. (2019)	Pipeline	Intent-based	Multi-Modal	Desk /Lap, Mobile	Assistance	No	Cancer	Patients	English	Not Specified	Not Specified
Yorita et al. (2020)	Pipeline	Rule-based	Multi-Modal	Mobile	Diagnosis, Counseling	No	Stress Management	Clinicians	English	Not Specified	Not Specified
Kargar and Mahoor (2017)	Pipeline	Rule-based	Speech	Robot	Intervention	Yes	Depression	Patients	English	Human Evaluation	Not Specified
Hwang et al. (2020)	Pipeline	Rule-based	Text	Not Specified	Diagnosis, Intervention	No	Medical Assistance	Patients	Korean	Not Specified	Not Specified
Srivastava and Singh (2020)	Pipeline	Rule-based	Text	Not Specified	Diagnosis, Assistance	Yes	Disease Diagnosis	Patients	English	Human Evaluation	Not Specified
Mathew et al. (2019)	Pipeline	Rule-based	Text	Mobile	Diagnosis, Assistance	Yes	Disease Diagnosis	Patients	English	Human Evaluation	Not Specified
Athota et al. (2020)	Pipeline	Rule-based	Multi-Modal	Mobile	Diagnosis, Assistance	No	Disease Diagnosis	Patients	English	Not Specified	Not Specified
Sadavarte and Bodanese (2019)	Pipeline	Hybrid	Multi-Modal	PDA	Assistance	No	Pregnancy	Patients	English	Human Evaluation	Not Specified
Lee et al. (2017)	Pipeline	Hybrid	Text	Mobile	Counseling	Yes	Psychiatric Counseling	Patients	Korean	Not Specified	Not Specified

Rahman et al. (2019)	Pipeline	Hybrid	Text	Not Specified	Diagnosis, Monitoring, Counseling	No	Medical Assistance	Patients	Bengali	Automated Evaluation	4,961 records
Yabuki and Sumi (2018)	Not Specified	Not Specified	Speech	Robot	Intervention	No	Autism Spectrum Disorder	Caregivers	English	Not Specified	Not Specified
Su et al. (2018)	Pipeline	Intent-based	Speech	Not Specified	Diagnosis, Assistance	No	Disease Diagnosis	Patients	Chinese	Automated Evaluation	Not Specified
Shoji et al. (2020)	Not Specified	Not Specified	Speech	Desk /Lap, PDA	Diagnosis	No	Pneumonia	Patients	Not Specified	Automated Evaluation	Not Specified
Polignano et al. (2020)	Pipeline	Hybrid	Multi-Modal	Mobile	Diagnosis, Intervention, Assistance, Monitoring	No	Medical Assistance	Patients	Italian	Human & Automated Evaluation	1,865,700 records
Ali et al. (2021)	Pipeline	Hybrid	Speech	Desk /Lap, Virtual Environment	Intervention	No	Cancer	Clinicians	English	Automated Evaluation	382 transcripts of conversations
Aarabi (2013)	Pipeline	Intent-based	Text	Not Specified	Diagnosis	No	Cardiology	Patients	English	Not Specified	Not Specified
Loisel et al. (2007)	Pipeline	Hybrid	Text	Not Specified	Assistance	No	Medical Assistance	Patients	French	Not Specified	Not Specified
Rosruen and Samanchuen (2018)	Pipeline	Hybrid	Multi-Modal	Desk /Lap, Mobile	Assistance	No	Medical Assistance	Patients	Chinese	Automated Evaluation	Not Specified
Sonntag and Moller (2010)	Pipeline	Intent-based	Multi-Modal	Desk /Lap	Assistance	Yes	Radiology	Clinicians	Not Specified	Human & Automated Evaluation	Not Specified
Kadariya et al. (2019)	Pipeline	Hybrid	Multi-Modal	Mobile	Monitoring, Intervention	Yes	Asthma	Patients	English	Human & Automated Evaluation	Not Specified
Siangchin and Samanchuen (2019)	Pipeline	Hybrid	Text	Mobile	Assistance	No	Medical Assistance	Clinicians	Chinese	Automated Evaluation	Not Specified

Erazo et al. (2020)	Pipeline	Rule-based	Text	Desk /Lap, Mobile	Diagnosis, Assistance	No	COVID-19	Patients	Not Specified	Human Evaluation	Not Specified
Huang et al. (2018)	Pipeline	Hybrid	Multi-Modal	Mobile	Monitoring, Intervention	Yes	Weight Loss	Patients	English, Chinese	Not Specified	Not Specified
Chen et al. (2013)	Pipeline	Rule-based	Speech	Desk /Lap, Mobile	Assistance	No	Medical Assistance	Patients, Caregivers	Chinese	Human Evaluation	MAT 400 dataset
Araki et al. (2011)	Pipeline	Intent-based	Multi-Modal	Desk /Lap	Intervention	No	Visually Impaired	Patients	Japanese	Human Evaluation	Not Specified
She et al. (2018)	End-to-End	Not Applicable	Speech	Robot	Intervention	Yes	Autism Spectrum Disorder	Patients	English	Automated Evaluation	Tager-Flusberg ,Nadig ASD English, and Rollins Corpus
Yabuki and Sumi (2018)	Not Specified	Not Specified	Speech	Robot	Intervention	Yes	Autism Spectrum Disorder	Caregivers	Japanese	Not Specified	Self-Constructed dataset
Wei et al. (2018)	Pipeline	Intent-based	Text	Not Specified	Diagnosis	No	Medical Assistance	Clinicians	Chinese	Automated Evaluation	Self-Constructed dataset
Fadhil and AbuRa'ed (2019)	Pipeline	Intent-based	Multi-Modal	Mobile	Monitoring, Assistance, Intervention	No	Medical Assistance	Patients	Arabic	Human Evaluation	Not Specified
Demasi et al. (2020)	Pipeline	Intent-based	Text	Not Specified	Counseling	No	Mental Health	Patients	English	Human Evaluation	Self-Constructed dataset
Waterschoot et al. (2020)	Pipeline	Intent-based	Speech	Not Specified	Monitoring	No	Mental Health	Patients	Dutch	Not Specified	Self-Constructed dataset
Danda et al. (2016)	Pipeline	Hybrid	Speech	Desk /Lap, Mobile	Diagnosing, Intervention, Assistance	No	Medical Assistance	Patients	Indian	Human & Automated Evaluation	CMU arctic dataset
Duggenpudi et al. (2019)	Pipeline	Rule-based	Text	Not Specified	Assistance	No	Medical Assistance	Patients	Telugu	Human Evaluation	Self-Constructed dataset

Prange et al. (2017)	Pipeline	Rule-based	Multi-Modal	Mobile	Assistance	No	Medical Assistance	Clinicians	Not Specified	Not Specified	475 records
Campillos Llanos et al. (2015)	Pipeline	Intent-based	Multi-Modal	Not Specified	Intervention	No	Medical Assistance	Clinicians	French	Not Specified	Not Specified
Welch et al. (2020)	Pipeline	Intent-based	Text	Not Specified	Counseling, Assistance	Yes	Mental Health	Patients	Not Specified	Human Evaluation	Not Specified
Ljunglöf et al. (2009)	Pipeline	Intent-based	Speech	Desk /Lap, Robot	Intervention	No	Communication Disorders	Patients	Swedish	Human Evaluation	Not Specified
Ljunglöf et al. (2011)	Pipeline	Intent-based	Speech	Desk /Lap, Robot	Intervention	Yes	Communication Disorders	Patients	Swedish	Human Evaluation	Not Specified
Brixey et al. (2017)	Pipeline	Hybrid	Text	Desk /Lap, Mobile	Assistance	No	HIV	Patients	English	Human Evaluation	Self-Constructed dataset
Morbini et al. (2014)	Pipeline	Rule-based	Speech	Desk /Lap, Virtual Environment	Counseling	Yes	Mental Health	Patients	English	Not Specified	Not Specified
DeVault et al. (2013)	Not Specified	Not Specified	Speech	Desk /Lap, Virtual Environment	Diagnosis	No	Mental Health	Clinicians	English	Not Specified	Not Specified
Inoue et al. (2016)	Pipeline	Rule-based	Multi-Modal	Mobile, Virtual Environment	Counseling	Yes	Mental Health	Patients	Not Specified	Not Specified	Not Specified
Morbini et al. (2012)	Pipeline	Intent-based	Text	Desk /Lap, Mobile	Counseling	Yes	PTSD	Patients	English	Not Specified	Not Specified
Xu et al. (2019)	End-to-End	Not Applicable	Text	Not Specified	Diagnosis	No	Disease Diagnosis	Patients	Chinese	Human & Automated Evaluation	Self-Constructed dataset
Green et al. (2004)	Pipeline	Rule-based	Speech	Desk /Lap	Intervention	No	Dementia	Caregivers	English	Human Evaluation	Not Specified