TASK-ORIENTED DIALOGUE SYSTEMS: PERFORMANCE VS QUALITY-OPTIMA, A REVIEW

Ryan Fellows[1*], Hisham Ihshaish[1], Steve Battle[1],
Ciaran Haines[1], Peter Mayhew[1,2], J.Ignacio Deza[1, 3]

[1] Computer Science Research Centre (CSRC),
University of the West of England (UWE), Bristol, United Kingdom
[2]GE Aviation, Cheltenham, United Kingdom
[3] Universidad Atlántida Argentina, Mar del Plata, Argentina

ABSTRACT

Task-oriented dialogue systems (TODS) – designed to assist users to achieve a goal – are continuing to rise in popularity as various industries find ways to effectively harness their capabilities, saving both time and money. However, even state-of-the-art TODS have not yet reached their full potential. TODS typically have a primary design focus on completing the task at hand, so the metric of task-resolution should take priority. Other conversational quality attributes that may point to the success, or otherwise, of the dialogue, are usually ignored. This can harm the interactions between the human and the dialogue system leaving the user dissatisfied or frustrated. This paper explores the role of conversational quality attributes within dialogue systems, looking at if, how, and where they are utilised, and examining their correlation with the performance of the dialogue system.

KEYWORDS

Dialogue Systems, Chatbot, Conversational Agents, AI, Natural Language Processing, Quality Attributes.

1. INTRODUCTION

Dialogue systems, by nature, are typically either chat-oriented or task oriented [1]. Chat-oriented, or conversational, dialogue systems have the objective of relaying contextually appropriate and stimulating responses [2], whereas task-oriented dialogue systems (TODS), or transactional systems, are designed to assist a user in completing their goals. Examples include finding transport times, booking tickets or customer support [3].

Over recent years, the adoption of TODS has surged significantly, as companies recognise their potential in alleviating the resource requirements inherent in human-based dialogue services. A prediction by market research firm Grand View Research estimates that the global chatbot market will reach \$1.23 billion by 2025 [4, 5].

The literature exploring TODS performance generally focuses on benchmarking against human-generated supervised feedback, such as that of task-resolution [6, 7]; a measure that encapsulates the dialogue system's success rate in resolving a task or set of tasks. A direct correlation is assumed between task resolution and the performance of the dialogue system as a whole.

David C. Wyld et al. (Eds): SIPP, NLPCL, BIGML, SOEN, AISC, NCWMC, CCSIT - 2022
pp. 69-87, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.121306

Whilst task-resolution is a priority – as the journey of the whole conversation is considered – performance and user experience cannot be disregarded, as they have the potential to hinder adoption of the system, independently of its performance. For this reason, in addition to task-resolution within TODS performance evaluation studies, more in particular compared to other types of dialogue systems, user satisfaction is commonly considered as another performance metric, as an indicator of system efficiency [8, 9] or usability [7].

The user satisfaction metric assumes a relative usability or efficiency for a dialogue system on the basis of how its users are satisfied. These are usually approximated by two approaches: either by means of laboratory experiments, eliciting human judgment on system outputs and behaviour relative to a predefined set of interaction parameters (e.g. number of turns [10], dialogue duration [11]). Or through modelling satisfaction, whereby the aim is to create models that provide ratings of performance similar to those which humans would do. The ratings based on human judgment are then used as target labels to learn an evaluation model based on objectively measurable performance attributes [12].

Comparing the performance of dialogue systems is a non-trivial task. This is due to the wide range of domains in which the systems are deployed, and the criteria they are evaluated against. Interactions are also subjective. What might be an optimal response for one individual, could be completely unsuitable for another, with performance being gauged on that specific individual's communicative preferences.

This paper explores quality attributes that describe different qualities of conversational interactions between a system and the user, besides task outcome. We analyse conversational quality attributes in TODS and explore how they are utilised, and to what effect. To accomplish this, a literature survey is undertaken to examine current considerations to conversational quality attributes used in conjunction with dialogue systems.

Throughout this paper, adherence will be made to a real-world locally collected corpus of interactions between University students and staff and University helpdesk assistants. This dataset consists of 600 email threads and 5697 subsequent emails - which are made up of a sender direction (incoming or outgoing), subject, body and a time stamp. Interactions consist of a range of issues which students and staff are in need of resolving. This GDPR compliant dataset will be referred to as the *ITS helpdesk* dataset throughout this paper.

The rest of the paper is organised as follows: Section 2 explores TODS conversational quality attributes and surveys their application to study and evaluate dialogue systems. This section is broken down into sub-sections consisting of individual quality attributes. Further discussion and conclusions are provided in Section 3.

2. CONVERSATIONAL QUALITY ATTRIBUTES

In a real world, human-to-human, task-oriented interaction, a conversation would likely not be deemed successful if only the task was resolved. If the advisor, in this situation, was friendly, personable and efficient in their manner, the advisee would be significantly more likely to have a positive experience and return in the future. However, if the advisor was rude or did not convey information competently, the advisee would most likely be left frustrated or even angry, leaving with a bad impression. Of course, interactions with a human do not translate perfectly to interactions with machines, yet findings from real world communication can be extrapolated and applied to virtual communication.

In most circumstances, a TODS should elicit a positive user experience while seeking to resolve tasks in the most effective way possible. Accordingly, the evaluation of TODS performance generally seeks to optimise two main qualities: task-resolution and dialogue efficiency.

This section surveys the state-of-the-art developments on conversational quality attributes in the context of TODS, and highlights some of the most prominent attributes addressed in the literature around TODS performance.

2.1. Task Resolution

Task-resolution, or goal completion, is one of the most accessible metrics — and arguably can be the easiest to derive given a well-defined user goal as well as a predefined function to quantify a resolved, unresolved or somewhere between, task — to evaluate the success of a TODS. The main purpose of a TODS is to assist a user with a specific task in an automated fashion. Therefore, the success of a dialogue system in fulfilling information requirements established by user goals is an indicator of a dialogue system's performance.

Practically, task-resolution (or success) is used to test dialogue systems success in providing not only the correct information, but also all user requested information — addressing as such the components for a given user-task: a set of constraints (target information, or information scope) and a set of requests (all required information) [13]. This in fact is consistent with the established understanding in Psychology around the notion of 'conversation', that is, it is understood that when individuals engage in conversation, there is a mutual understanding of the goals, roles and behaviours that can be expected from the interaction [14, 15]. Therefore, the 'performance' of the dialogue has to be evaluated on the basis of their mutual understanding and expectations.

In its simplest form, however, this metric can be quantified as a Boolean — binary task success (BTS) — value indicating whether a task or set of tasks has been resolved or not. Using this metric, organisations can capture useful statistics over a number of interactions to derive how effective their dialogue system is at solving tasks, in comparison to interactions with human assistance or even other dialogue systems.

One of the more inherent challenges of task-resolution, as a performance metric, is knowing whether the task in question has been resolved. Especially so as the different users may have different goals, or intrinsically multiple goals, and these may even change in response to system behaviour throughout the course of interaction. On top of this, different users may have varying definitions of success, for example, a domain-specific expert user may deem a task resolved with less detailed information acquired compared to a novice user.

Typically, an interaction with a dialogue system will end when a user terminates the conversation, however this doesn't necessarily imply that their goals have been met. Some dialogue systems opt to explicitly elicit 'task completion' in some form: "*has your request been resolved?"* or "*is there anything else I can help you with?"*, others attempt to use some form of classifier to infer when a task has been resolved through a machine learning and NLP model (eg. [16, 17]). This requires a structured definition of goals and a mechanism to measure success relative to that goal. In this fashion, much of the work on automating the evaluation of task success has largely focused on the domain-specific TODS. This is usually an easier task as such systems can be highly scripted, and task success can be specifically defined – especially so in traditional dialogue systems, such as the Cambridge Restaurant System [18] and the ELVIS email assistant [19] — where the relevant ontology defines intents, slots and values for each slot of the domain.

However, a structured definition of goals will usually bind dialogue systems to a specific class of goals, constraining their ability to adapt to the diversity and dynamics of goals pertinent in human-human dialogue [20]. To address the shortcomings in adaptability and transferability encountered in single-domain systems, research into domain-aware, or multi-domain, dialogue systems has attracted noticeable attention in recent years [21, 22]. This saw the introduction of the concept of the domain state tracker (DST), which accumulates the input of the turn along with the dialogue history to extract a *belief* state: user goals/intentions expressed during the course of conversation. User intentions are then encoded as a discrete set of dialogue states, i.e., a set of slots and their corresponding values, as shown in e.g., [23, 24]. As a result, the multiple user intentions are subsequently evaluated, whether objectively met or otherwise - Please refer to Figure 1 in [25] for a detailed characterisation of DSTs.

Reinforcement learning systems aim to find the optimal action that an automated agent can take in any given circumstance, by either maximizing a reward function or minimizing a cost function. With a dialogue system as the agent, the given circumstance is the belief state held by the DST, the reward function is linked to task-resolution, and the actions are the system's output slots and values. Dialogue systems will inevitably encounter problems; examples include incorrectly identifying a word, or a user changing their goal. A system could assign confidence levels to the belief states, track multiple belief states, and include a plan to recover the conversational thread after the errors are noticed.

Casting the conversation as a partially observable Markov decision process (POMDP) allows for these uncertainties to be encoded [31]. A POMDP is defined as a tuple {S, A, τ, R, O, Z, λ, b0}where S is a set of states describing the environment; A is a set of actions that may be taken by the agent; τ is the transition probability $P(s' \mid s,a)$; R defines the expected reward r(s, a); O is a set of verifiable observations the agent can receive about the world; Z defines an observation probability, $P(o' \mid s',a)$; λ is a geometric discount factor $0 \leq \lambda \leq 1$; and b0 is an initial belief state b0(s).

A POMDP dialogue system tracks multiple parallel belief states, selecting actions based on the belief state that is most likely. When misunderstandings occur, the current belief state can be made less likely, allowing the system to move to a new belief state. Because the belief states' probabilities are tracked alongside the expected action rewards and the chance that an action will transition as expected, a POMDP is able to effectively plan how to manage a dialogue. This framework allows a TODS to track multiple possible user goals, to plan error checking of user utterances, and to use context to potentially identify when the dialogue system has misunderstood the user intention. However, converting this potential benefit into practice is not trivial. Such systems are known to require a significant amount of training, as the state - action space can be very large even for single domains, and uncertainty in the task resolution may weaken the agent's learning [26].

In general, task-resolution is commonly quantified as the result of a performance metric in which user satisfaction is maximised. The PARADISE framework [27], which is frequently used as a baseline for task success evaluation throughout literature, values user satisfaction as a weighted linear combination of task success measures side by side with dialogue costs (reported in Sec. 2.5). These measures can be objective, which entail features such as word error rate [28], automatic speech recognition (ASR), word-level confidence score [29], number of errors made by the speech recognizer [30] and time to fire, task completion rate, and accuracy metrics as used in [31], or subjective such as intelligibility of synthesized speech [32] and perception tests [33].

Table 1 breaks down the threads within the ITS helpdesk dataset into the task resolution percentage via topic. In this example, threads have been classified into groups of topics using the

unsupervised topic modelling algorithm of latent Dirichlet allocation (LDA) to provide a baseline example of topic categorisation. Threads have now been contextualised to some degree which can then allow further analysis in conjunction with the objective measures that will follow this Section.

Table 1. Breakdown of Task resolution statuses of ITS Helpdesk threads.

Topic	Topic Keywords	Number of Threads	Task Resolution Percentage
1	Person, Need, Would, Email, Work	89	Resolved: 82% Unresolved: 14.6% N/A: 3.4%
2	Student, Person, Access, Look, Module	19	Resolved: 82% Unresolved: 14.6% N/A: 3.4%
3	Access, File, Document, Try, Help	24	Resolved: 82% Unresolved: 14.6% N/A: 3.4%
4	Order, Laptop, Could, Login, Generic	7	Resolved: 82% Unresolved: 14.6% N/A: 3.4%
5	Person, System, Group, Purchase, User	20	Resolved: 82% Unresolved: 14.6% N/A: 3.4%
6	Screen, Mark, Drive, Room, File	17	Resolved: 82% Unresolved: 14.6% N/A: 3.4%
7	Folder, Course, Number, Upload, Video	21	Resolved: 82% Unresolved: 14.6% N/A: 3.4%
8	Person, Add, Address, Email, Staff	311	Resolved: 82% Unresolved: 14.6% N/A: 3.4%
9	Desktop, Office, Slow, Urgently, Computer	92	Resolved: 82% Unresolved: 14.6% N/A: 3.4%

2.2. Usability and Dialogue Efficiency

Usability attributes, such as user satisfaction, learnability, efficiency, etc, are the foundation of the design of ‘successful’ dialogue systems, as these are ultimately created for the user, and for the user to achieve their intended, and occasionally variable, goal(s). While such attributes should ultimately be the criteria to evaluate a dialogue system, they are well-known to be subjective, and subsequently hard to measure. This is why much literature on evaluating dialogue systems tends to deal with quantifiable performance metrics, like task-resolution rate or elapsed time of the interaction. It has been proposed, however, that an agent's competence in objectively measurable dialogue does not necessarily induce a better user experience, and subsequently a better overall usability [34]. In fact, the different metrics may even prompt contentious interpretations, or simply contradict each other [35].

Although usability ratings are notoriously hard to interpret, especially if the system is not equipped to infer and keep track of user goals, the successful encapsulation of such values can provide insight that explicit metrics struggle to capture. From the study of Malchanau et al, usability experts rated examined questions from a 110 item questionnaire and derived an

evaluation of their agreement of usability concepts. This led to a collection of 8 attributes they saw as key factors: task completion and quality, robustness, learnability, flexibility, likeability, ease of use and usefulness (value) of an application [34]. This questionnaire was used to evaluate a dialogue system designed for training purposes, in which the overall system usability was determined by the quality of agreements reached, by the robustness and flexibility of the interaction, and by thequality of system responses.

Additionally, these different metrics may in fact have an inconsistent statistical interpretation to different designers. In the same way human evaluation will provide different outcomes based on the subjective criteria, the same can be said for metrics of usability which are difficult to consistently quantify [35].

2.3. User Sentiment

Because of the insights sentiment analysis reveals about the more concise bodies of text on social media, the field of sentiment analysis has seen a take-up of use over recent times [36]. Sentiment analysis can be performed on large quantities of tweets and posts from different platforms to assess general opinion about a specific product or topic.

Different applications use a range of machine learning classification algorithms to categorise sentiment scores [37, 38], some use just two classes: positive and negative, while others use an n-point scale, e.g., very good, good, satisfactory, bad, very bad [39]. A review and a comparative study of existing techniques for opinion mining like machine learning and lexicon-based approaches is provided in [40].

Table 2. Main user sentiment studies in dialogue systems reviewed in the literature

Domain	Author	Year	Proposal / Findings
SDS	Schuller [41], Nwe[42]	2003	Emotion recognition in spoken dialogue using phonic features.
SDS and TOSS	Devillers [43]	2003/05	Automatic and ’robust’ cues for emotion detection using extra linguistic features, lexical and discourse context.
SDS	TH Bui [44]	2006	’Affective’ dialogue model: inferring user’s emotional state for an adaptive system’s response. Earlier work applied to spoken dialogue systems in.
TODS and SDS	Ferreira [45], Ultes [46]	2013/17	Proposed an expert-based reward shaping approach in dialogue management, and a live user satisfaction estimation model based on ’Interaction Quality’, a ”less subjective variant of user satisfaction”.

DS	Shin [47]	2018	Detecting user sentiment from multimodal channels (acoustic, dialogic and textual) and incorporating the detected sentiment as feedback into adaptive end-to-end DS
DS	Jaques [48]	2019	Deep reinforcement learning model (off-policy batch RL algorithm).
DS	Shin [49]	2019	Happybot: on-policy learning in conjunction with a user-sentiment approximator to improve a seq2seq dialogue model.
DS	Sasha [50]	2020	Applying Reinforced Learning to manage multi-intent conversations with sentiment based immediate rewards

DS: Dialogue Systems, **SDS**: Spoken Dialogue Systems, **TODS**: Task-oriented Dialogue Systems, **TOSS**: Task-oriented Spoken Systems

Early studies on sentiment analysis in the context of dialogue systems explored the inclusion of user sentiment in rule-based systems, towards adaptive spoken dialogue systems [51, 52]. Most of these studies investigated modular-based dialogue systems (conventionally referred to as pipeline models), with predefined rules for systems to adapt to variability in user sentiment. In recent studies, however, much focus has been placed onto sentiment-adaptive end-to-end dialogue systems, particularly due to their adaptability in comparison with modular-based ones [53], which are known to be harder to train, and adapt to new contexts [54].

Studies exploring the conjunction of dialogue systems with sentiment analysis are often motivated by the notion of system *adaptability*, assuming a correlation between adaptability of the systems to user sentiment and their satisfaction. Some recent work emphasises the importance for conversational agents to adapt to different user (personality) types [55, 56]. Attention is paid to studying user sentiment as a variable to guide the design of sentiment-adaptive dialogue systems [57, 58]. A comprehensive list of development milestones on sentiment analysis application to the analysis and evaluation of dialogue systems, as well as on sentiment-adaptive systems is provided in Table 2.

It should be noted, nonetheless, that sentiment analysis methods have not been extensively applied to conversational agents and dialogue systems. One reason for this is the fact sentiment analysis performs more effectively when pre-trained on a domain specific dataset, and would not often generalise to open domains of discourse inherent in many dialogue systems. One example is the well-known shortcomings when generalising sentiment classification of models trained on the IMDB movie database to classify sentiment about movies [59, 60].

However, as data becomes more accessible and the sentiment analysis techniques become more sophisticated, the performance and scalability of many sentiment analysis tools are constantly improving. This in fact can allow for further advances in the development of *sentiment-aware* dialogue systems, such that dialogue systems can adapt to the dynamics of user sentiment throughout the course of interaction. Depending on the objective function used to optimise, there can be multiple approaches to extract and use the variability in user-sentiment, which can be categorised into two groups:

- **Individual user utterance**: which looks at the sentiment score of individual user utterance, which can offer insight into the specific semantics and vectors of that single interaction such as that found in [58, 59]. This compartmentalised approach allows a deeper evaluation of the content of that one message, whether this is a product, experience or other entity.

- **Contextual user utterance**: examines the thread as a whole can be explored from a temporal perspective, the evolution of the thread, rather than just individual messages [60, 61]. This can give insight as to why the sentiment of the user is going up or down and allows evaluation as to why this is happening. When compared with other threads, trends can be found as to what is causing the fluctuation of sentiment. The difference of sentiment score between the first and last message, which can be referred to as the 'sentiment swing' can also be very useful, as this is an example of how the situation has progressed from the perspective of the user.

An example to illustrate user-sentiment swing during dialogue is provided in Tables 3, 4 and 5 which shows three resolved task-oriented interactions. The sentiment score corresponding to the user utterance at each turn is recorded. For simplicity, the variability in user-sentiment at each turn is smoothed in Figure 1. Conversation 1 remains fairly neutral throughout the interaction, ending with a slightly more positive sentiment than at the beginning of the exchange. Conversation 2 shows a positive uptick in sentiment as the relatively simple issue is solved. However, the sentiment of conversation 3 represents the frustration of the user, showing a severe drop in sentiment as they encounter issues with their query. However, as the issue is resolved in the end, the sentiment recovers accordingly to conclude with a positive sentiment score.

Table 3. Conversation One: A thread from the ITS helpdesk dataset.

Source	Utterance	Score	Swing
Conversation One			
User	Hi there, I am unable to copy and paste HTML text - or any text - into Cereus. We have been told by our web editor to paste the text from Word into an online HTML editor and then copy and paste the HTML into Cereus. Unfortunately it doesn't work, even when I right-click to paste, or use control C and V. Thanks,	0.128	
Helpdesk	Are you still having issues with copying and pasting into Cereus via HTML web editor ? What is the name of the Web Editor that gave you this advice ? Regards		
User	Yes I still am having the issues. We use https://html-online.com/editor/ Thank you.	0.27	↑

Helpdesk	I think this might be one of two possible issues. As a first step would you mind using IE11 to access the application via Cereus please? I know sometimes the text editing box can be a bit flaky on newer browsers. Kind regards,		
User	Thank you, but I don't have IE 11. Do you have a safe link you can send me as not sure which source to trust to download. I need IE 11 for Mac…	0.13	↓
Helpdesk	Agh! Sorry —*SR*— I didn't realise you were on a Mac. I don't quite know what to suggest in this case.I haven't heard of anyone else having issues on a Mac but that might be because no one else uses one when trying to use the News app. Cereus is a bit of an old dinosaur and due to be decommissioned soon I'm afraid. I don't suppose you have access to a PC do you? If not I think I will have to put you back to the Help Desk and get them to assign the job to someone who supports Macs. Sorry about this		
User	Hi —*SR*—, I am due to pick up my PC laptop from UWE, but not heard back as to when that could be yet. Plus I need to put out a press release tomorrow morning... Yes, please do put me in touch with one of your Mac guys. Huge thanks for your help though!, Regards	0.24	↑

Table 4. Conversation Two: A thread from the ITS helpdesk dataset.

Conversation Two			
User	Hi BB, Where has the guidance about sign up sheets been moved to?	0	
Helpdesk	Hi, —*SR*— (Request for Information) has been assigned to Learning and Research at the status of 'In Progress'. Open the ticket Thank you		
User	Any news?	-0.12	↓
Helpdesk	Good afternoon *—Person—*, We have had a big clear out of the web site, and are pointing people to the the main support pages for blackboard, if you need further assistance please feel free to contact our help desk. i have found this guidance on sign up sheets here: *—Misc—* thanks		
User	Hey ITS, I might be going blind but where does it mention sign up sheets? I thought multi-sign up sheets were something UWE built? Appreciate your time! *—Person—*	0.36	↑
Helpdesk	Hi *—Person—*, If you are referring to sign up sheets as related to the creation of groups please see the following link: —*Misc*— otherwise if you are referring to the third party 'SignUp Lists' function then the link for that can be found on the above staff guides link page. Regards		
User	That's absolutely grand, thanks —*Person*—!	0.71	↑

Table 5. Conversation One: A thread from the ITS helpdesk dataset.

Conversation Three			
User	Hi Folks, I need to arrange to have 3 laptops (mac or pc) to use for the *—Module—* week at the —*Location*—. How do i go about this please. Groups of students will be planning/editing mixing desk set-ups and making spreadsheets. The desk editing software is a free download Midas M32-Edit software available here —*Misc*— The masterclass runs —*Misc*— to —*Misc*— Cheers, much appreciated!	0.84	
Helpdesk	Hi —*SR*— by *—Person—* for 3 PC or mac laptops for *—Date—* to *—Date—* has been assigned to Client Services Regional - Assignment Details: 3 PC or mac laptops for *—Date—* to *—Date—* Open the ticket Thank you		
User	Hi Folks, I will need to pick up these computers tomorrow for the early start on Monday morning at —*Location*—. Can you tell me where I can collect them from and if the software isn't on them already how we can install it. Many thanks for your help.	0.47	↓
Helpdesk	Hello *—Person—* Sorry but IT Services do not have a stock of loanable laptops. I would suggest trying the FET Project room. If students are using *—Misc—* built laptops off site they will have to log in to them on site beforehand to create their user profile. If software needs installing ask the Project room to liaise with the *—Room—* ITS helpdesk who will assist with this. Regards		
User	Hi *—Person—*, i realise this probably isn't your fault . . . BUT To wait for 7 days to tell me this is a little bit off. Can you understand why I might think that this falls short of reasonable service? I'm not very happy to find out at the last moment something which you might have told me at the start of this week when I would have had time to do something about it!	-0.76	↓
Helpdesk	*—Person—* has turned up trumps with 2 machines .. they will need to be set up with logins that 45 students can use at the *—Misc—* and with this software . . . I will bring them to *—Room—* in a short while for you to action this. The desk editing software is a free download Midas M32-Edit software available here —*Misc*— The masterclass runs 0900 *—Date—* to 1900 *—Date—* Regards		
User	Hi Folks, All sorted now, crisis averted, many thanks! Cheers	0.36	↑

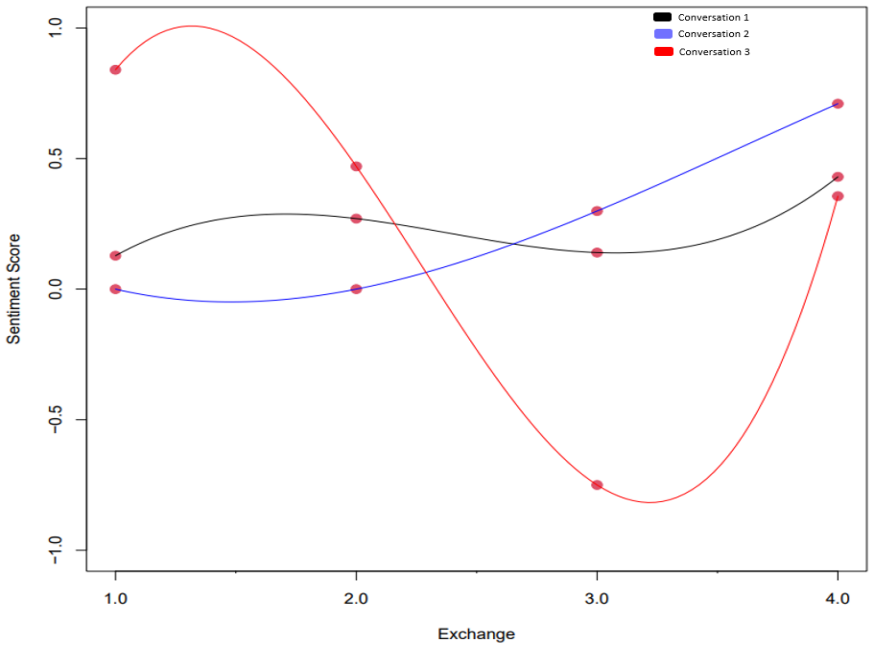

Figure 1. An illustration of the user sentiment score of conversation 1, 2 and 3 from Tables 3, 4 and 5.

Despite the scores fluctuating throughout the interaction, all threads end with a neutral to positive conclusion, indicating that the user was satisfied or happy with the outcome. Whilst this is insightful in itself, the highs and lows provide a chance to understand why these values were exhibited at that point, which could allow for the examination of the objective attributes or semantics used. The values could also just simply be the result of a contextual issue, such as, in this case, a restaurant being fully booked.

However, regardless of the domain in which sentiment analysis is utilised, a cautious apprehension should be taken in interpreting the obtained scores. Modern sentiment analysis tools are advancing, but they are still not mature enough to accurately recognise sarcasm, jokes and nuances of language. There is also the limitation of a lack of distinctive sentiment annotations amongst an already limited amount of datasets readily available, as observed in [60] which subsequently makes it harder to perform accurate analysis on dialogues of a more extensive lexicon.

What's more, sentiment analysis is sensitive to social conventions which are prevalent in human communication. Many interactions through email, for example, will exhibit some form of generic greeting such as 'Good Morning' as well as a sign off (sometimes inserted automatically through a template) such as 'Best Wishes'. These terms are often used by individuals, regardless of the context of their email, which can therefore skew the sentiment score to be higher than the actual substance that their email might elicit.

Therefore, it could be argued that the current state of sentiment analysis makes it a useful tool to gain analytical insight from a corpus of text, but to utilise them as the sole driver for action could potentially lead to erroneous decision making. The context of its usage is important.

2.4. Dialogue Cost

The term ‘dialogue cost’ appears frequently throughout dialogue system literature [62, 63, 64] and typically refers to multiple aspects of resource retrieval and utilisation ranging from the data itself, to the computational power required by the model being utilised. Some literature even refers to the explicit monetary cost of the dialogue system based on the manual labour required to label the data, often using the method of crowdsourcing [65, 66].

Relevant and feature rich data is the foundation for a performing dialogue system, and no matter how good a model is, it cannot compensate for a small or poor quality dataset. Therefore, such resources can be expensive to acquire, whether in terms of time or money [67]. In more domain specific dialogue, the data native to these sometimes unfamiliar domains, plays an even more important role as it highlights semantic and pragmatic phenomena that is unique to that domain.

Alongside task-resolution, dialogue cost is often considered to infer ‘dialogue strategies’ [68] which specify at each stage what the next action to be taken by the system. A dialogue strategy can have the objective of converging towards the goal state in the most efficient way possible through a series of interactions with the user. ‘Efficiency’ can for example, mean access to external resources, the dialogue duration, internal computation time, or resource use. The goal is to reduce these ‘costs’ to allow the system to achieve higher performance.

However, the ambiguity of the term ‘dialogue cost’ can make it a difficult area to assess. The PARADISE framework describes efficiency measures such as the number of turns or elapsed time to complete a task [68, 80, 6], as well as qualitative measures such as inappropriate or repair utterances [70, 71] as potential dialogue costs. Whereas, some researchers explore the term from a reinforcement learning perspective, in which the dialogue cost is a penalisation assigned for taking the wrong action predicated on a pre-defined function. Therefore, it can be a difficult to quantify cost in relation to a dialogue. Even when considering what is typically agreed on, regardless of the context, that dialogue ‘cost’ should be minimised, i.e., to maximise system efficiency, there isn't such established foundation to suggest that, for instance, a shorter —hence more ‘efficient’– dialogue is directly correlated to a better user experience. In fact, it can simply be the opposite.

2.5. Dialogue Cost

The retention rate of a TODS is often referred to as a measure of the number of users that return to use the system within a given time frame. This is another important, yet accessible metric for quantifying dialogue systems' performance. If a company’s chatbot aims to replace other communication channels (e.g., lowering call volume), the goal is to obtain significantly higher retention, which can be indicative of higher consumer satisfaction [72]. However, there are plenty of other automated options that allow users to manage accounts easily without speaking to a human. Thus, if a chatbot is focused on customer support, a high retention rate does not necessarily have to be the measure of success [73].

The context and domain in which the TODS is deployed is an important factor to consider when looking at the retention rate of a given dialogue system. If the dialogue system in question is a health-based chatbot for a one-off issue, then the user is unlikely to have to reuse the chatbot, and therefore the metric is less valuable. However, if the chatbot is being deployed as a customer service replacement, then a high retention rate can be interpreted as a positive performance indicator, as it shows the user has enough confidence in the system to reuse it.

Related metrics are those of *dropout* rate and *bounce* rate. The dropout rate refers to the number of users who quit the session with the dialogue system before an outcome had been reached. A high dropout rate for a dialogue system can be a substantial indication of poor performance. The bounce rate is the volume of users that do not utilise the dialogue system for its intended use. A high retention rate with low dropout and bounce rates would suggest a high level of performance.

However, only so much can be derived from the metric of retention rate without some form of user feedback, as the metric is sensitive to anomalies. A dialogue system could perform perfectly, yet a user might not return for other, unknown reasons. This should not be indicative of the performance of the system, yet the metric might suggest this to be the case. Therefore, the larger the set of interactions retention is analysed on, the more insightful the findings will potentially be. Because of this, it could be argued that the rate of retention offers a good overview perspective of system performance, but such considerations should prevent retention rate from being a primary form of performance insight. It is also important to note that the ability to extract the retention rate is not always feasible, as is the case with the ITS helpdesk dataset.

2.6. Response Time Cost

The literature exploring dialogue response time is typically concerned with reducing the time it takes a conversational agent to respond to the user. The consensus is that a user wants responses as quickly as possible, and for the interaction to be as efficient as possible in terms of session time. The focus is often on the mechanics of the model in question, rather than the effect that response time could have on user satisfaction [74].

Alternative studies on response time shift the focus from the desire for instant responses to adding more human-like delays. In their study of using dynamic response delays for machine generated messages, Gnewuch et al [75] prioritise the 'feel' of the conversation over speed of response, opting to 'calculate a timing mechanism based on the complexity of the response and complexity of the previous message as a technique to increase the naturalness of the interaction'. As a result of these dynamic delays, they showed an increase in both the perception of humanness and social presence, as well as a greater satisfaction with the overall dialogue interaction; a faster response time is not necessarily better.

However, as with the majority of the quality attributes, the context and domain are very important to consider. 'Replika' [76] is an anthropomorphised chatbot designed as a companion to help battle loneliness. It utilises a slight delay to make the interaction feel more genuine and human-like, as instant replies would make the interaction feel too machine-like and break the social illusion. Conversely, '911bot' [77] is a chatbot that allows a user to describe an emergency situation, and because of this context, any artificial delays would not be appropriate. This highlights the importance of context when considering such conversational attributes to evaluate TODS performance.

Computationally, response time has become much less of a pressing issue in recent times for smaller to medium scale dialogue systems, as abundant computational resources, and innovation in machine learning \ NLP approaches, make instantaneous responses entirely feasible, and as a result, expected. Therefore, it could be argued that whilst a dialogue system might not get praised on its performance for optimal response times, whether instant or timed, it will be negatively graded for sub-optimal response times.

2.6. Conversation Length

The literature exploring the explicit length of conversation is limited. This is due to the fact that the developers predominantly focus on the substance of a message first, with the subsequent message length being as long or short as it needs to be. However, the length of an agent's responses can significantly alter the dynamic of an interaction, as it determines how much information can be conveyed in a single turn. Depending on the topic at hand, if the messages are too short, there is a risk the user will grow frustrated with the lack of detail in the answer, but if the messages are too long, the user's attention may wander.

In their guide to developing "better" chatbots for mental health, Dosovitsky et al [78] argue that "developers should strive to find a module length that enhances intervention fidelity without compromising engagement" and "should focus on creating a few engaging and effective modules at the beginning rather than developing a large variety of untested modules". Simply put, system utterance length should be adaptive, changing relative to the stage of the conversation.

Other work examined the effect of message length relative to the dialogue domain, e.g., [79], emphasising that one of the most important chatbot performance metrics is conversation length and structure. Industry trends suggest aiming for shorter conversations with simple structure, in line with the notion of efficient service. For example, banking chatbots are assumed to provide quick solutions such as sending and receiving money, or checking a balance. When the social aspect of the conversation is more important, fast and concise responses may turn counter-productive.

However, just looking at conversation length from an objective perspective can be misleading. If an analysis is performed in which it is deemed shorter messages are preferred for a given domain, and are subsequently rewarded, then this may undermine the very relevant factor of context. Dialogue systems often have the objective of being as efficient as possible, which would encourage the idea of concise discourse, which may not be a problem. However, some issues and topics simply do not lend themselves to this approach and require further development in the conversation. Therefore, it would be detrimental to the system to simply penalise longer message without any thought to the semantics and context involved. This is not to say conversation length is not a useful quality attribute, as the literature suggests, it is, yet the optimisation of this parameter needs more than just the configuration of a value for utterance length or number of turns.

3. DISCUSSION AND CONCLUSIONS

It is clear that there is no shortage of studies exploring the field of TODS and their performance [80]. However, research into TODS in conjunction with conversational quality attributes, beyond that of task-resolution, are less abundant. One potential reason for this is because these attributes, such as conversation length, response time and user-sentiment are often referred to more as bi-products of the dialogue systems performance in meeting user information requirements.

Although many studies on optimising TODS performance examined metrics for performance evaluation beyond that of task-resolution, thus far, however, the modelling of TODS performance as a multivariate function of multiple conversational quality attributes remains an open question.

Additionally, TODS are still difficult to evaluate. Although there are established methods and frameworks which are frequently referred to in literature, with PARADISE arguably the most applied, yet there is still no standard in place for a novel TODS to be measured against. This is

undoubtedly a hindrance to the field, as it gives a lack of consistency when designing a system and subsequently comparing it with others in the industry. Also, with the growing complexity of modern virtual assistants such as Siri, Bixby and Alexa to name a few, where each could be described as a *sophisticated* TODS, the task of objectively evaluating such systems is only going to become a more complex process.

Therefore, although significant progress has been made in the field of TODS over a relatively short period, there are still various challenges to be overcome. Arguably the most pressing issue is the lack of a standardised protocol for human evaluation, which makes it challenging to compare different approaches to one another [95]. On the other hand, automatic evaluation metrics have proven their utility with their efficiency and undemanding approach to dialogue assessment but are still considered less reliable in comparison to human judgement [132]. A shortage of task-oriented open-source datasets also acts as a bottleneck in the progression of the field, especially when approaching multiple domains. All of which is compounded by a growing expectation of the average user, as TODS are generally becoming more and more innovative on a global scale.

REFERENCES

[1] P.-H. Su, M. Gasic, N. Mrksic, L. Rojas-Barahona, S. Ultes, D. Vandyke, T.-H. Wen, S. Young, On-line active reward learning for policy optimisation in spoken dialogue systems, arXiv preprint arXiv:1605.07669 (2016).

[2] O. Vinyals, Q. Le, A neural conversational model, arXiv preprint arXiv:1506.05869 (2015).

[3] M. Henderson, B. Thomson, J. D. Williams, The second dialog state tracking challenge, in: Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL), 2014, pp. 263–272.

[4] Chatbot market size to reach $1.25 billion by 2025 — cagr: 24.3%: Grand view research, inc, shorturl.at/gjqwT, accessed: 2021-07-14.

[5] W. Wang, K. Siau, Trust in health chatbots, Thirty ninth International Conference on Information Systems, San Francisco 2018 (2018).

[6] M. A. Walker, D. J. Litman, C. A. Kamm, A. Abella, Paradise: A framework for evaluating spoken dialogue agents, arXiv preprint cmp-lg/9704004 (1997).

[7] S. Moller, R. Englert, K.-P. Engelbrecht, V. Hafner, A. Jameson, A. Oulasvirta, A. Raake, N. Reithinger, Memo: towards automatic usability evaluation of spoken dialogue services by user error simulations., Ninth International Conference on Spoken Language Processing (01 2006).

[8] J. D. Williams, K. Asadi, G. Zweig, Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 665–677.

[9] C. Kamm, User interfaces for voice applications, Proceedings of the National Academy of Sciences 92 (22) (1995) 10031–10037. arXiv:https://www.pnas.org/content/92/22/10031.full.pdf.

[10] M. Walker, J. C. Fromer, S. Narayanan, Learning optimal dialogue strategies: A case study of a spoken dialogue agent for email, in: COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics, 1998.

[11] N. Fraser, D. Gibbon, R. Moore, R. Winski, Assessment of interactive systems., Mouton de Gruyter, 1998, pp. 564–615.

[12] J. M. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, M. Cieliebak, Survey on evaluation methods for dialogue systems, Artificial Intelligence Review 54 (1) (2020) 755–810

[13] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, S. Young, Agenda-based user simulation for bootstrapping a pomdp dialogue system, in: Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics;Companion Volume, Short Papers, 2007, pp. 149–152.

[14] H. H. Clark, S. E. Brennan, Grounding in communication., Perspectives on socially shared cognition (1991)

[15] H. H. Clark, Using language, Cambridge university press, 1996.
[16] P.-H. Su, D. Vandyke, M. Ga˘si´c, D. Kim, N. Mrk˘si´c, T. H. Wen, S. Young, Learning from real users: Rating dialogue success with neural networks for reinforcement learning in spoken dialogue systems, arXiv preprint arXiv:1508.03386 (09 2015)
[17] B. Thomson, S. Young, Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems, Computer Speech Language 24 (4) (2010) 562–588
[18] J. Planells, L. Hurtado Oliver, E. Segarra, E. Sanchis, A multi-domain dialog system to integrate heterogeneous spoken dialog systems,
[19] M. Noseworthy, J. C. K. Cheung, J. Pineau, Predicting success in goal-driven human-human dialogues, in: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Association for Computational Linguistics, Saarbr ¨ucken, Germany, 2017, pp. 253–262.
[20] C.-S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, P. Fung, Transferable multi-domain state generator for task-oriented dialogue systems, in: ACL, 2019.
[21] Y. Huang, J. Feng, M. Hu, X. Wu, X. Du, S. Ma, Meta-reinforced multi-domain state generator for dialogue systems, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp.7109–7118.
[22] N. Mrksic, D. O Seaghdha, T.-H. Wen, B. Thomson, S. Young, Neural belief tracker: Data-driven dialogue state tracking, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1777–1788.
[23] P. Xu, Q. Hu, An end-to-end approach for handling unknown slot values in dialogue state tracking, 2018, pp. 1448–1457.
[24] C.-S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, P. Fung, Transferable multi-domain state generator for task-oriented dialogue systems, in: ACL, 2019.
[25] J. D. Williams, S. Young, Partially observable markov decision processes for spoken dialog systems, Computer Speech & Language 21 (2) (2007) 393–422.
[26] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang, X. Zhu, Recent advances and challenges in task-oriented dialog systems, Science China Technological Sciences (2020) 1–17
[27] M. A. Walker, D. J. Litman, C. A. Kamm, A. Abella, Paradise: A framework for evaluating spoken dialogue agents, arXiv preprint cmp-lg/9704004 (1997).
[28] J. F. Allen, B. W. Miller, E. K. Ringger, T. Sikorski, Robust understanding in a dialogue system, in: Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Vol. 62, 1996, p. 70
[29] R. Meena, G. Skantze, J. Gustafson, Data-driven models for timing feedback responses in a map task dialogue system, Computer Speech & Language 28 (4) (2014) 903–922
[30] Z. Callejas, R. Lopez-Cozar, Relations between de-facto criteria in the evaluation of a spoken dialogue system, Speech Communication 50 (8-9) (2008) 646–665.
[31] S. M. Robinson, A. Roque, A. Vaswani, D. Traum, C. Hernandez, B. Millspaugh, Evaluation of a spoken dialogue system for virtual reality call for fire training, Tech. rep., University of Southern California Marina Del Rey Ca Inst for Creative Technologies (2007).
[32] L. Lamel, S. Rosset, J.-L. Gauvain, Considerations in the design and evaluation of spoken language dialog systems (03 2001).
[33] A. Kamm, M. Walker, D. Litman, Evaluating spoken language systems (06 1999).
[34] A. Malchanau, V. Petukhova, H. Bunt, Multimodal dialogue system evaluation: a case study applying usability standards, in: 9th International Workshop on Spoken Dialogue System Technology, Springer, 2019, pp. 145–159.
[35] E. Raita, A. Oulasvirta, Too good to be bad: Favorable product expectations boost subjective usability ratings, Interacting with Computers 23 (4) (2011) 363–371
[36] V. M., J. Vala, P. Balani, A survey on sentiment analysis algorithms for opinion mining, International Journal of Computer Applications 133 (2016) 7–11

[37] W. Medhat, A. Hassan, H. Korashy, Sentiment analysis algorithms and applications: A survey, Ain Shams Engineering Journal 5 (4) (2014) 1093–1113.
[38] H. Sinha, A. Kaur, A detailed survey and comparative study of sentiment analysis algorithms, in: 2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS), 2016, pp. 94–98
[39] R. Prabowo, M. Thelwall, Sentiment analysis: A combined approach, Journal of Informetrics 3 (2) (2009) 143–157.
[40] V. Kharde, S. Sonawane, Sentiment analysis of twitter data: A survey of techniques, International Journal of Computer Applications 139 (2016) 5–15
[41] B. Schuller, G. Rigoll, M. Lang, Hidden markov model-based speech emotion recognition, in: 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698), Vol. 1, 2003, pp. I–401.
[42] T. L. Nwe, S. W. Foo, L. C. De Silva, Speech emotion recognition using hidden markov models, Speech Communication 41 (4) (2003) 603–623
[43] L. Devillers, L. Lamel, I. Vasilescu, Emotion detection in task-oriented spoken dialogues, in: 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698), Vol. 3, 2003, pp. III–549.
[44] - T. Bui, J. Zwiers, M. Poel, A. Nijholt, Toward affective dialogue modeling using partially observable markov decision processes, in: 1st workshop on Emotion and Computing – Current Research and Future Impact, 2006, pp. 47–50.
[45] E. Ferreira, F. Lef`evre, Expert-based reward shaping and exploration scheme for boosting policy learning of dialogue management, in: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, 2013, pp. 108–113.
[46] S. Ultes, P. Budzianowski, I. Casanueva, N. Mrksic, L. M. Rojas-Barahona, P. hao Su, T.-H. Wen, M. Gai´c, S. J. Young, Domain-independent user satisfaction reward estimation for dialogue policy learning, in: INTERSPEECH, 2017.
[47] J. Shin, P. Xu, A. Madotto, P. Fung, Happybot: Generating empathetic dialogue responses by improving user experience look-ahead (2019). arXiv:1906.08487
[48] N. Jaques, A. Ghandeharioun, J. H. Shen, C. Ferguson, A. Lapedriza, N. Jones, S. Gu, R. Picard, Way off-policy batch deep reinforcement learning of human preferences in dialog (2020).
[49] T. Saha, S. Saha, P. Bhattacharyya, Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning, PLOS ONE 15 (7) (2020) 1–28.
[50] J. Acosta, Using emotion to gain rapport in a spoken dialog system., in: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium, 2009, pp. 49–54
[51] J. Pittermann, A. Pittermann, W. Minker, Emotion recognition and adaptation in spoken dialogue systems, International Journal of Speech Technology 13 (2010) 49–60.
[52] B. Liu, I. Lane, End-to-end learning of task-oriented dialogs, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, New Orleans, Louisiana, USA, 2018, pp. 67–73
[53] N. Braunschweiler, A. Papangelis, Comparison of an End-to-end Trainable Dialogue System with a Modular Statistical Dialogue System, in: Proc. Interspeech 2018, 2018, pp. 576–580.
[54] Q. V. Liao, W. Geyer, M. Muller, Y. Khazaen, Conversational Interfaces for Information Search, Springer International Publishing, Cham, 2020, pp. 267–287.
[55] E. Ruane, S. Farrell, A. Ventresque, User perception of text-based chatbot personality, in: A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, E. Luger, M. Goodwin, P. B. Brandtzaeg (Eds.), Chatbot Research and Design, Springer International Publishing, Cham, 2021, pp. 32–47
[56] W. Shi, Z. Yu, Sentiment adaptive end-to-end dialog systems, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018, pp. 1509–1519.
[57] T. Saha, S. Saha, P. Bhattacharyya, Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning, PLOS ONE 15 (2020)

[58] H. Kumar, B. Harish, H. Darshan, Sentiment analysis on imdb movie reviews using hybrid feature extraction method., International Journal of Interactive Multimedia & Artificial Intelligence 5 (5) (2019).
[59] A. Yenter, A. Verma, Deep cnn-lstm with combined kernels from multiple branches for imdb review sentiment analysis, in: 2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON), IEEE, 2017, pp. 540–546
[60] H. Saif, Y. He, H. Alani, Semantic sentiment analysis of twitter, in: International semantic web conference, Springer, 2012, pp. 508–524.
[61] K. Scheffler, S. Young, Automatic learning of dialogue strategy using dialogue simulation and reinforcement learning, in: Proceedings of HLT, Vol. 2, 2002.
[62] M. A. Walker, D. J. Litman, C. A. Kamm, A. Abella, Paradise: A framework for evaluating spoken dialogue agents, arXiv preprint cmplg/9704004 (1997).
[63] J. Rela ̃no-Gil, D. Tapias, M. C. Gancedo, M. Charfuel ́an, L. Hern ́andez, Robust and flexible mixed-initiative dialogue for telephone services, in: Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999, pp. 287–290.
[64] M. Mitchell, D. Bohus, E. Kamar, Crowdsourcing language generation templates for dialogue systems, in: Proceedings of the INLG and SIGDIAL 2014 Joint Session, 2014, pp. 172–180.
[65] P. Shah, D. Hakkani-T ̈ur, B. Liu, G. T ̈ur, 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), Association for Computational Linguistics, New Orleans - Louisiana, 2018, pp. 41–51
[66] R. Manuvinakurike, M. Paetzel, D. DeVault, Reducing the cost of dialogue system training and evaluation with online, crowd-sourced dialogue data collection, Proceedings of SEMDIAL (2015) 113–121
[67] E. Levin, R. Pieraccini, W. Eckert, Learning dialogue strategies within the Markov decision process framework, in: 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings, IEEE, 1997, pp. 72–79.
[68] A. Abella, M. K. Brown, B. Buntschuh, Development principles for dialog-based interfaces, in: Workshop on Dialogue Processing in Spoken Language Systems, Springer, 1996, pp. 141–155
[69] L. Hirschman, C. Pao, The cost of errors in a spoken language system, in: Third European Conference on Speech Communication and Technology, 1993
[70] M. Danieli, E. Gerbino, Metrics for evaluating dialogue strategies in a spoken language system, in: Proceedings of the 1995 AAAI spring symposium on Empirical Methods in Discourse Interpretation and Generation, Vol. 16, 1995, pp. 34–39.
[71] A. Simpson, N. M. Eraser, Black box and glass box evaluation of the sundial system, in: Third European Conference on Speech Communication and Technology, 1993.
[72] M. Dhyani, R. Kumar, An intelligent chatbot using deep learning with bidirectional rnn and attention model, Materials Today: Proceedings 34 (2019) 817–824
[73] A. Nursetyo, E. R. Subhiyakto, et al., Smart chatbot system for e-commerce assitance based on aiml, in: 2018 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), IEEE, 2018, pp. 641–645.
[74] M. Kowsher, A. Tahabilder, M. Z. I. Sanjid, N. J. Prottasha, M. M. H. Sarker, Knowledge-base optimization to reduce the response time of bangla chatbot, 2020 Joint 9th International Conference on Informatics, Electronics and Vision and 2020 4th International Conference on Imaging, Vision and Pattern Recognition, ICIEV and icIVPR 2020 (8 2020).
[75] U. Gnewuch, S. Morana, M. T. Adam, A. Maedche, Faster is not always better: understanding the effect of dynamic response delays in human-chatbot interaction, in: 26th European Conference on Information Systems: Beyond Digitization-Facets of Socio-Technical Change, ECIS 2018, Portsmouth, UK, June 23-28, 2018. Ed.: U. Frank, 2018, p. 143975.
[76] Replika. URL https://replika.ai/
[77] J. Martin, 911bot, https://github.com/surgeforward/911bot (2016).
[78] G. Dosovitsky, B. S. Pineda, N. C. Jacobson, C. Chang, M. Escoredo, E. L. Bunge, Artificial intelligence chatbot for depression: Descriptive study of usage, JMIR Formative Research 4 (11 2020).
[79] A. Przegalinska, L. Ciechanowski, A. Stroz, P. Gloor, G. Mazurek, In bot we trust: A new methodology of chatbot performance measures, Business Horizons 62 (6) (2019) 785–797.

[80] Fellows, R., Ihshaish, H., Battle, S., Haines, C., Mayhew, P. and Deza, J.I., 2021. Task-oriented Dialogue Systems: performance vs. quality-optima, a review. *arXiv preprint arXiv:2112.11176*.