An Assessment Framework for DialPort

*Kyusong Lee, *Tiancheng Zhao, Stefan Ultes, Lina Rojas-Barahona, Eli Pincus, David Traum and Maxine Eskenazi

Abstract Collecting a large amount of real human-computer interaction data in various domains is a cornerstone in the development of better data-driven spoken dialog systems. The DialPort project is creating a portal to collect a constant stream of real user conversational data on a variety of topics. In order to keep real users attracted to DialPort, it is crucial to develop a robust evaluation framework to monitor and maintain high performance. Different from earlier spoken dialog systems, DialPort has a heterogeneous set of spoken dialog systems gathered under one outward-looking agent. In order to access this new structure, we have identified some unique challenges that DialPort will encounter so that it can appeal to real users and have created a novel evaluation scheme that quantitatively assesses their performance in these situations. We look at assessment from the point of view of the system developer as well as that of the end user.

*Kyusong Lee
Carnegie Mellon University, PA, USA e-mail: kyusongl@andrew.cmu.edu

*Tiancheng Zhao
Carnegie Mellon University, PA, USA e-mail: tianchez@andrew.cmu.edu

Stefan Ultes
Cambridge University, Cambridge, UK e-mail: stefan.ultes@eng.cam.ac.uk

Lina Rojas-Barahona
Cambridge University, Cambridge, UK e-mail: lmr46@cam.ac.uk

Eli Pincus
University of Southern California, CA, USA e-mail: pincus@ict.usc.edu

David Traum
University of Southern California, CA, USA e-mail: traum@ict.usc.edu

Maxine Eskenazi
Carnegie Mellon Unversity, PA, USA e-mail: max@cs.cmu.edu

*Both authors contributed equally to this paper

1 Introduction

Data-driven methods have become increasingly popular in developing better spoken dialog systems (SDS) due to their superior performance and scalability compared to manual handcrafting [7, 5]. DialPort [10, 9, 8] is providing a new solution for rapidly collecting conversational data. The goal of DialPort is to combine a large number of dialog systems that have diverse functionality in order to attract a group of stable real users and to maintain those users' interest. In order to ensure that the real users are attracted to DialPort over the long term, we need a principled framework for monitoring and improving its performance is required. In this manner, at any time we can at any time have a snapshot of system performance and quickly make changes so that we do not lose our users. Past SDS assessment paradigms [6, 1] may not be directly applicable to DialPort because it groups multiple remote agents that are heterogeneous in nature. Thus this paper proposes a novel assessment scheme based on the PARADISE framework [6] which was designed to measure SDSs' user satisfaction. Specifically, our assessment scheme assesses DialPort according to its ultimate goals: collecting large amounts of data, and satisfying the real users' needs. For the latter goal, the assessment must reflect *how* the portal achieves the former goal: the smooth character of the conversation, response delay and performance.

In order to verify the effectiveness of the proposed assessment paradigm, we conducted a real-user study using the DialPort portal agent, which transfers control of the dialog, according to user needs, to five remote agents: 1) a weather information system from CMU (using NOAA[1]), 2) a restaurant information system from CMU (using YELP[2]), 3) another restaurant system from Cambridge University [2], 4) a word guessing game agent from USC [3] and 5) a chat-bot from CMU and POSTECH. The CMU systems and the portal are on-site and the others are connected from off-site locations. We quantitatively assessed the performance of the DialPort portal in both task success rate, dialog management efficiency and speed of response. Finally, we show that our evaluation framework provides robust measures for DialPort.

2 Evaluation Framework of DialPort

PARADISE [6] measures the user satisfaction (US) in terms of the *success* and *cost*, where success measures *what* a system is supposed to accomplish and cost measures *how* a system achieves its goal. Also, in order to achieve the best US, a system should aim to maximize success and minimize cost. Based on this formulation, we lay out the structure of the measure of US for DialPort in Figure 1. The following sections explain the elements included in our measure of US in details.

[1] http://www.noaa.gov/

[2] https://www.yelp.com/developers/documentation/v2/overview

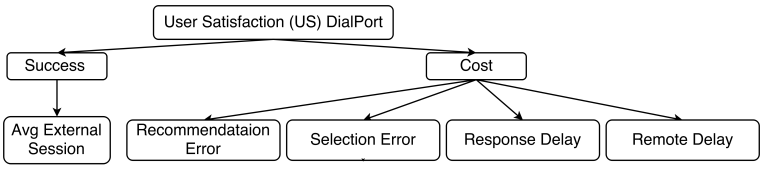

Fig. 1 The decompositions of user satisfaction for DialPort.

2.1 Success of DialPort

The goal of DialPort is to collect large amounts of spoken dialog system data from real users, including both successful and failed dialogs. For clarity, we refer to a dialog with a remote agent as *external sessions* and the whole conversation from the moment the user enters the Portal to the end when they leave as a *portal session* (a portal session is composed of multiple external sessions). Therefore, the success of DialPort could be measured in terms of the average number of external sessions per portal session (avgExtSess). However in order to keep real users attracted to DialPort and in turn create more data, it is essential to maintain good conversational flow and successful performance. Therefore, we also must minimize the cost, which is defined below.

2.2 Cost

The cost, in our context, can be defined as:

- Response Delay of DialPort (D): responding to users with a minimal delay is crucial for an SDS [4] in order to maintain the pace of the interaction and to avoid barge-ins and system interruptions of the user. We measure the average delay in milliseconds from the point at which the user finishes speaking to the beginning of the next system response.
- Selection Error (S): refers to the situation when the Portal agent, which must decide which agent to connect to the user, switches to a suboptimal external agent. We measure this in terms of selection error rate.
- Recommendation Error (R): refers to the case where users do not agree with the Portal agent's recommendation. We measure this in terms of recommendation error rate.
- Remote Delay (RD): Although the success of DialPort could depend on the number of dialogs it gathers, the quality of the external agents greatly impact user satisfaction. Therefore, we include the average response delay with remote agents as a part of the cost, in order to avoid disruptions such as barge-ins and system interruptions.

2.3 Overall Performance

The original PARADISE framework learns a linear regression [6] to weight the importance of each cost and success in order to predict the users' subjective scores. In this work, we assume each element is uniformly weighted. Thus we first normalize each input and compute the overall US of DialPort by:

$$\mathrm{US}_{meta} = \mathscr{N}(\mathrm{avgExtSess}) - (\mathscr{N}(S) + \mathscr{N}(R) + \mathscr{N}(D) + \mathscr{N}(\mathrm{RD})) \quad (1)$$

where $\mathscr{N}$ stands for z normalization to standardize the input into zero mean and unit variance, to ensure measurements with different scales can be combined together.

3 Evaluation

We assessed DialPort from the points of view of the assessment of the Portal agent giving the impression of seamless domain changes. We analyzed log data of the current version of DialPort. We gathered 119 dialogs from a group of 10 LTI students. The maximum number of dialogs per any one user was 11 and the minimum number of dialogs was 1. On average a dialog with DialPort lasted 10.97 turns. Two experts manually tagged the selection error and the recommendation error for the portal. Other measures were automatically obtained from the log data such as the number of sessions and response delay.

Evaluation of External Remote Agents

	Portal Agent	**All External Agents**	**Two On-site (w/o chatbot)**	**Three On-site (plus chatbot)**	**Two Off-site**
# of External Sessions	-	614	135	552	62
Avg # of Turn/External Session	-	2.35	2.54	1.37	11.01
Avg Resp Delay (ms)	457.80	683.54	518.97	683.39	683.89
Std Resp Delay (ms)	456.86	529.74	333.72	488.56	599.19

Table 1 Performance of external agents.

First, we show the statistics such as the number of external sessions and response delay of groups of external agents (Table 1). We looked at: all external agents combined, all three on-site agents, both off-site agents, and two CMU agents excluding the chatbot. We observed that many user requests are directed to the chat bot. Moreover, 28.8% of chatbot utterances were non-understanding recovery turns, such as "can you please rephrase that?", "sorry I didn't catch that". The remaining 71.2% of the chatbot turns were question-answering and chit-chat. Based on the log data collected, we observed that users tend to talk to the chatbot after finishing a conversation with one of the task-oriented external agents. After a few turns of interactions

with the chatbot, users usually initiate new external sessions with other task-oriented external agents. This shows the importance of handling out-of-domain utterances gracefully to maintain the flow of the dialog.

We also note that the external sessions with off-site agents are longer than those with the on-site ones. While the chatbot is responsible for a lot of this difference, it is good to note that users can carry on longer conversations with the external agents since our goal is to create a data flow for these systems.

There are two main causes of response delay: a network delay and a computation delay. The servers for the Portal agent and the on-site remote agents are located in Pittsburgh, so the network delay between them and the other systems was small. Off-site systems are have a longer network delay. Table 1 shows that the total response delay of on-site systems is longer than the delay of off-site systems. This is because the chatbot accesses a very large database, which introduces a significant amount of computation time. By excluding the chatbot in our calculations, we see that the response delay of on-site systems is $24.1\%(\text{P-value} < 0.001)$ faster than the off-site systems.

Evaluation of the Portal Agent

We then separately assessed the Portal agent, that is, the agent that the user interacts with that decides which SDS to connect to the user. We assess the Portal agent for 1) recommendation error 2) selection error. Recommendation errors were labeled for each system utterance with the recommendation intent with binary label $\{0, 1\}$, where 1 means correct and 0 means incorrect.

	avgExtSess	**select error**	**recommend error**
Master Agent	4.32	22.6%	37.15%

Table 2 Performance of the Portal agent. Since there is only one version of the Portal agent in this study, we cannot z-normalize the scores. AvgExtSess is number of external systems accessed in one portal session

A recommendation score gets label 1 if the users agree with the recommendation, otherwise it is 0. For example, when a system said ”Would you like to know about next weekends weather?”, a positive reward would be received if the next user utterance is ”yes”, ”okay” or ”what is the weather in Pittsburgh?”. Otherwise, the system would get the label 0. The result shows that 62.85% of the requests were successfully recommended. Selection errors were labeled for each system utterance that switches to a new agent. If the transition is correct, it gets label 1, otherwise 0. For example, if a user said Who is the president of South Korea? the Portal agent should select the chatbot. Any other selection will result in label 0. 78.40% were successfully sent to the appropriate agents. Figure 2 shows that the most frequent error was due to the Portal agent selecting a weather or restaurant agent when a location entity was detected in chit-chat utterances.

	System A	System B	System C	System D	System E
System A	0.958	0.000	0.000	0.000	0.042
System B	0.071	0.857	0.000	0.000	0.071
System C	0.000	0.000	0.900	0.000	0.100
System D	0.000	0.059	0.000	0.824	0.118
System E	0.140	0.053	0.009	0.000	0.798

Fig. 2 The external system confusion matrix for the Portal.

4 Conclusion

We have proposed a framework for the assessment of DialPort and evaluated the Portal agent and its access to the external system based on our collected annotated conversational data. We have identified unique challenges that DialPort faces and have provided a comprehensive evaluation framework that covers the performance of response delay, agent transition, and recommendation strategy for the portal. In future work, we plan to develop data-driven models to automatically predict the success and cost of the Portal and remote agents of DialPort, which suggests a promising research direction.

Acknowledgements We would like to thank Youngjun He for his help in the annotation of the data used in this paper. This work is funded by National Science Foundation grant CNS-1512973. The opinions expressed in this paper do not necessarily reflect those of the National Science Foundation.

References

1. Liu, C.W., Lowe, R., Serban, I.V., Noseworthy, M., Charlin, L., Pineau, J.: How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. arXiv preprint arXiv:1603.08023 (2016)
2. Mrkšić, N., Séaghdha, D.O., Thomson, B., Gašić, M., Su, P.H., Vandyke, D., Wen, T.H., Young, S.: Multi-domain dialog state tracking using recurrent neural networks. arXiv preprint arXiv:1506.07190 (2015)
3. Pincus, E., Traum, D.: Towards Automatic Identification of Effective Clues for Team Word-Guessing Games. In: Proceedings of the Language Resources and Evaluation Conference (LREC), pp. 2741–2747. European Language Resources Association, Portoro, Slovenia (2016)
4. Raux, A., Eskenazi, M.: A finite-state turn-taking model for spoken dialog systems. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 629–637. Association for Computational Linguistics (2009)
5. Vinyals, O., Le, Q.: A neural conversational model. arXiv preprint arXiv:1506.05869 (2015)
6. Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A.: Paradise: A framework for evaluating spoken dialogue agents. In: Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics, pp. 271–280. Association for Computational Linguistics (1997)
7. Williams, J.D., Young, S.: Partially observable markov decision processes for spoken dialog systems. Computer Speech & Language **21**(2), 393–422 (2007)

8. Zhao, T., Eskenazi, M., Lee, K.: Dialport: A general framework for aggregating dialog systems. EMNLP 2016 p. 32 (2016)
9. Zhao, T., Lee, K., Eskenazi, M.: Dialport: Connecting the spoken dialog research community to real user data. In: 2016 IEEE Workshop on Spoken Language Technology (2016)
10. Zhao, T., Lee, K., Eskenazi, M.: The dialport portal: Grouping diverse types of spoken dialog systems. In: Workshop on Chatbots and Conversational Agents (2016)