

The DialPort tools

Jessica Huynh* and Shikib Mehri* and Cathy Jiao* and Maxine Eskenazi

Carnegie Mellon University

jhuynh, amehri, cljiao, max@cs.cmu.edu

Abstract

The DialPort project (<http://dialport.org/>), funded by the National Science Foundation (NSF), covers a group of tools and services that aim at fulfilling the needs of the dialog research community. Over the course of six years, several offerings have been created, including the DialPort Portal and DialCrowd. This paper describes these contributions, which will be demoed at SIGDIAL, including implementation, prior studies, corresponding discoveries, and the locations at which the tools will remain freely available to the community going forward.

1 Introduction

The DialPort project¹ has created tools and services that respond to needs voiced by many in the dialog research community during several workshops organized by the Principle Investigators (PIs). Its offerings are available at no cost to the community with the goal of helping researchers gather high quality data, and easily assess and compare their dialog systems. This paper and its corresponding demos showcase the DialPort Portal² and DialCrowd³.

There is an increasing need for large amounts of natural dialog data that can be obtained at reasonable cost and in an interactive manner. Static datasets are ineffective for both evaluation and optimization. This has led to the creation of the DialPort Portal, which facilitates the collection of flexible and evolving data as well as interactive assessment with real users. Notably, the Portal was used to connect systems and collect data for the *Interactive Evaluation of Dialog* track (Mehri et al., 2021) at DSTC9 (Gunasekara et al., 2020).

Another community need centers around how to gather high quality data when using crowdsourcing platforms. DialCrowd has been constructed to facilitate crowdsourcing by guiding researchers to give clear, understandable explanations of the task to the workers who produce or annotate data. It also aids in calculating the correct level of worker payment. Finally, it includes several methods of data quality assessment.

The University of Southern California (USC) is a partner in DialPort. The team at USC works on a tools repository⁴ and the REAL Challenge.

This paper gives background and describes in detail the parts of both the Portal and DialCrowd. It also provides information on how to access and use them. As the DialPort project draws to an end, the paper indicates the permanent sites where these tools will reside.

2 Background

2.1 Interactive Platforms for Dialog

As dialog models improve, it is imperative that they are evaluated in interactive settings with real users. Mehri and Eskenazi (2020) show that while pre-trained dialog systems excel at generating responses (Zhang et al., 2019; Bao et al., 2020), they underperform in back-and-forth interactions.

The Alexa Prize challenge (Ram et al., 2018; Khatri et al., 2018) allows university teams to build socialbots that are assessed in interactive settings with Alexa users. In contrast, the DialPort Portal is accessible to the broader research community. Furthermore, the Alexa Prize challenge primarily relies on speech input from the user, which may result in speech recognition errors. Though the DialPort Portal can accept speech input, its web interface can also be used with text-only input.

*Equal contribution

¹<http://dialport.org/>

²<https://dialport.org/portal>

³<http://dialport.org/dialcrowd.html>

⁴<https://dialport.ict.usc.edu/>

2.2 Crowdsourcing

With the amount of dialog data available or able to be collected with systems such as DialPort, it is important to have easy and accessible tools to create detailed annotations of this data for different metrics. One method of obtaining annotations is crowdsourcing with platforms such as Amazon Mechanical Turk (AMT). However, it is sometimes difficult to obtain conclusive results, and a survey of current natural language processing HITs has shown the weaknesses of these HITs (Huynh et al., 2021). Instructions (Chandler et al., 2013), examples (Doroudi et al., 2016), and payment are some of the aspects that need to be attended to in order for HITs to acquire higher quality data.

3 DialPort Portal

The DialPort Portal was initially conceived with the objective of listing many dialog systems from a variety of sites. This type of platform, with demonstrations, links, and references to various systems, is valuable to both researchers and real users. The concept of the Portal evolved, and the different systems were linked such that a user could interact with all of the connected systems, transitioning seamlessly between systems, with the dialog state (consisting of slots such as city or date) shared across systems (Zhao et al., 2016; Lee et al., 2017). As dialog systems continued to improve, especially with the advent of engaging response generation models (Zhang et al., 2019; Bao et al., 2020), the Portal recruited real users through Facebook advertising with the objective of providing researchers with a platform to collect interactive dialogs with real users (Mehri et al., 2021).

3.1 Portal Version 1

The original version of the Portal grouped several dialog systems from different sites (Cambridge, USC, CMU) and managed seamless switching amongst (Zhao et al., 2016). For example, a user could ask for the weather in Pittsburgh and get the CMU weather system, then ask the CMU system for the weather in Cambridge, then ask for a restaurant and automatically switch to the Cambridge restaurant system, then ask to play a game and get the USC system.

This instance of the Portal serves as a platform to interact with different systems over the course of one dialog (Zhao et al., 2016). To accomplish this, the Portal needed to address several challenges

(1) how to share information across systems (e.g., remembering the city the user wanted the weather for when interacting with the CMU system, and sharing that with the Cambridge system when the user wants a restaurant recommendation), (2) how to gracefully continue a dialog when a system is down, and (3) how to give two systems addressing the same task (e.g., restaurants) equal time with the users. Respectively, these problems were addressed by (1) maintaining a shared dialog state across systems, (2) backing off to an equivalent system or changing the topic, and (3) a pseudo-random system selection policy. In order to make the system easy to use, an API was developed to facilitate connecting new systems to the Portal. This version has pedagogical value as it can easily be demonstrated for dialog classes.

3.2 Portal Version 2

With the advent of the API, the possibilities of use of the Portal greatly expanded. The Portal was used for the DSTC9 Challenge (Mehri et al., 2021), as a tool that enabled researchers to both compare their systems on one common platform (with real users) and to gather considerable amounts of data. The Portal was made available to DSTC9 participants. The idea was to connect systems and have them tested by real (unpaid) users. The CMU DialPort team advertised the Portal on Facebook and interested individuals tried it out (with text only). Upon visiting the Portal, real users are randomly matched with a dialog system, without knowledge of the specific system they are interacting with. While some people left the site after only one or two turns with a system, many actually continued to communicate with a system for a substantial conversation, and were thus considered to be real users. Real users consist of users who find some personal interest (getting information, companionship, curiosity) in continuing a dialog. There were 11 participants in the interactive part of the Challenge (Mehri et al., 2021). With an advertising budget of \$2500, we collect more than 4000 dialogs on the DialPort portal (2960 dialogs with at least 4 turns or 8 utterances); thus the cost was less than \$1.00 per usable dialog. The DialPort portal, through funding from the National Science Foundation, has been able to provide interactive evaluation as a service free of charge to any dialog researchers. The Appendix contains a sample dialog from the winning system of the DSTC9 track (Bao et al., 2020).

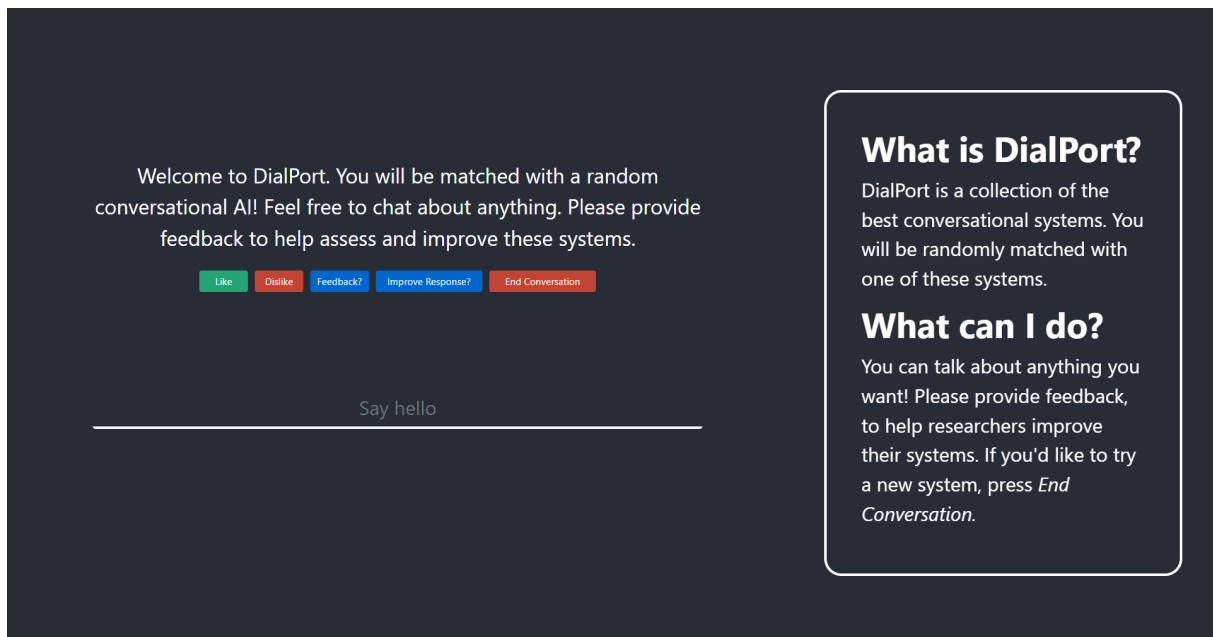


Figure 1: DialPort Portal. This screenshot of the Portal displays (1) the dialog history, shown in the center of the screen, (2) an input field for the user to type their responses, and (3) a set of feedback buttons below the dialog history (“Like”, “Dislike”, “Feedback?”, “Improve Response?” and “End Conversation”). The interface clear and emphasizes the three important actions that a user should perform while using the Portal: (1) reading the dialog history, (2) responding to the dialog system, and (3) providing feedback.

DSTC9 demonstrated that the Portal could easily be used to both compare systems and to gather data with real users. Besides challenges, another potential use of the Portal would be for students to connect systems that they build for a class project to see how well they do in real user interaction.

At the end of the DialPort project in the coming year, the Portal will move from the Dialog Research Center at CMU to LDC at UPenn.

3.3 DialPort Dashboard

After collecting data from real users on the DialPort interface, a subsequent task is to perform analysis on the gathered data. We provide the DialPort dashboard which allows researchers to (i) *analyze* dialogs collected on their system, (ii) *interact* with the dashboard to filter and organize dialogs based on various criteria, and (iii) *compare* their system to other systems connected to the DialPort Portal. Currently, the Dashboard contains over 7000 dialogs from 28 systems. The Dashboard is connected to the DialPort Portal via API calls, allowing dialogs to be quickly displayed on the Dashboard after being collected from the Portal. The Dashboard code will soon be released, allowing for use of the Dashboard in offline mode.

The Dashboard UI contains panels, tables, and

charts. At both the system and dialog level, attributes such as the number of utterances, likes, dislikes, comments, corrections are displayed (see figure 3). In addition, the two evaluation metrics of FED (Mehri and Eskenazi, 2020) and human ratings are shown. Since the Dashboard is designed to be easily extended, additional metrics can be added in the future. Users can interact with the dashboard by filtering and ranking dialogs based on attributes and metrics. For example, the provided toolbar can be used to find all conversations with a given user’s system with more than n turns or rank conversations from most-to-least number of likes. Users can also filter words and phrases in dialogs by their number of occurrences from the perspective of both the system or human participant, and thus view common phrases or words mentioned on either side of the conversation. Finally, each system contains a progress monitor graph which displays the number of dialogs being collected over time, allowing users to actively observe data collection in the DialPort Portal.

4 DialCrowd

To address the many issues that present themselves when using crowdsourcing to collect high quality data, DialCrowd was created. DialCrowd (Lee

Instructions

After reading each of the following messages, please rate the message as spam or not spam.
We expect this HIT will take **5 minute(s)** and we will pay **\$1.25**.

Categories

Category	Instructions	Examples	Counterexamples
spam	Select this if you feel that the message is spam.	<ul style="list-style-type: none"> Click this link to win \$10,000!!! <i>because...</i> 	<ul style="list-style-type: none"> The period for selecting courses has started. <i>because...</i>
not spam	Select this if you feel that the message is not spam.	<ul style="list-style-type: none"> Here is your tracking number for your package. <i>because...</i> 	<ul style="list-style-type: none"> You've won a cruise to the Bahamas! Click here to redeem. <i>because...</i>

Figure 2: DialCrowd Examples and Counterexamples with Explanations

et al., 2018) is a dialog assessment toolkit which aids researchers with human intelligence task (HIT) creation. Requesters follow templates on the DialCrowd site, which generate a HIT that can be linked for a worker on any crowdsourcing site.

The second version of this tool (Huynh et al., 2022) focuses on collecting high-quality data with tools such as:

- Links to create better instructions
- Prompts to provide examples and counterexamples with explanations seen in Figure 2
- Functionality for adding golden data and duplicate data in each HIT
- Payment suggestions
- A feedback area
- Overall statistics from the HIT (time, patterns in the responses, inter-annotator agreement)

This allows for requesters to create a well-structured HIT which allows workers to provide better quality annotations. Consequently, it makes it easier to filter responses from potential bots. Additional tools include the capability to include a mandatory consent form at the start of the HIT, and detailed style changes for the HIT. Further description of the system along with corresponding images can be found in (Huynh et al., 2022).

One DialCrowd template, intent classification, has been merged into the new home for DialCrowd, ParlAI⁵, and is now available for use.

5 The DialPort demo

The demos of the DialPort Portal and Dashboard and of DialCrowd at SIGDIAL will include:

⁵<https://github.com/facebookresearch/ParlAI/tree/main/parlai/crowdsourcing/tasks/dialcrowd>

- how to connect a system
- what interaction with each tool looks like
- advantages there are in using the tools, with examples (for example, what resulting data looks like)

6 Conclusion and Future Directions

The tools presented in this demo help dialog researchers in data gathering and assessment. As the community uses them, more types of applications will arise. The tools have been created in a way that enable additions as the field and the needs evolve.

7 Acknowledgements

This work is funded by National Science Foundation Grant Nos. CNS-1512973, DGE1745016, and DGE2140739. The opinions expressed in this paper do not necessarily reflect those of the National Science Foundation. The authors would like to thank Tiancheng Zhao, Kyusong Lee, and Ting-Rui Chiang for their contributions to these tools.

References

- Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. Plato-2: Towards building an open-domain chatbot via curriculum learning. *arXiv preprint arXiv:2006.16779*.
- Jesse Chandler, Gabriele Paolacci, and Pam Mueller. 2013. Risks and rewards of crowdsourcing marketplaces. In *Handbook of human computation*, pages 377–392. Springer.
- Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2623–2634.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D’Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2020. Overview of the ninth dialog system technology challenge: Dstc9. *arXiv preprint arXiv:2011.06486*.

Jessica Huynh, Jeffrey Bigham, and Maxine Eskenazi. 2021. A survey of nlp-related crowdsourcing hits: what works and what does not. *arXiv preprint arXiv:2111.05241*.

Jessica Huynh, Ting-Rui Chiang, Jeffrey Bigham, and Maxine Eskenazi. 2022. [Dialcrowd 2.0: A quality-focused dialog system crowdsourcing toolkit](#). In *Proceedings of the Language Resources and Evaluation Conference*, pages 1256–1263, Marseille, France. European Language Resources Association.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*.

Kyusong Lee, Tiancheng Zhao, Alan W Black, and Maxine Eskenazi. 2018. Dialcrowd: A toolkit for easy dialog system assessment. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 245–248.

Kyusong Lee, Tiancheng Zhao, Yulun Du, Edward Cai, Allen Lu, Eli Pincus, David Traum, Stefan Ultes, Lina M Rojas Barahona, Milica Gasic, et al. 2017. Dialport, gone live: an update after a year of development. In *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue*, pages 170–173.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. *arXiv preprint arXiv:2006.12719*.

Shikib Mehri, Yulan Feng, Carla Gordon, Seyed Hossein Alavi, David Traum, and Maxine Eskenazi. 2021. Interactive evaluation of dialog track at dstc9.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2016. Dialport: Connecting the spoken dialog research community to real user data. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 83–90. IEEE.

A Appendix

User: Hello

System: hello! hows your day going so far?

User: What are you up to?

System: just finished making coffee! what about you? what do you do?

User: i’m playing video games

System: oh, nice! what kind? and i’m just relaxing here and listening to some classical music.

User: i’m playing league of legends

System: nice! what rank are you? i haven’t played in forever.

User: who is your favorite classical musician?

System: probably beethoven. i like classical music.

User: i’m plat 2

System: nice! i’d imagine you are better than that!

User: can you play an instrument?

System: i can play the piano. do you play any instruments?

User: thanks

System: i appreciate it! so what is it like being plat 2?

User: i can play the clarinet, but not very well haha

System: oh nice! i played clarinet in high school.

User: the games are difficult but rewarding

System: yes they are! i played the trumpet in college.

...

Table 1: An example of a conversation snippet from the DSTC9 top performing system. The system is capable of maintaining two simultaneous threads of dialog about League of Legends and music, and switch between them in a coherent and engaging manner.

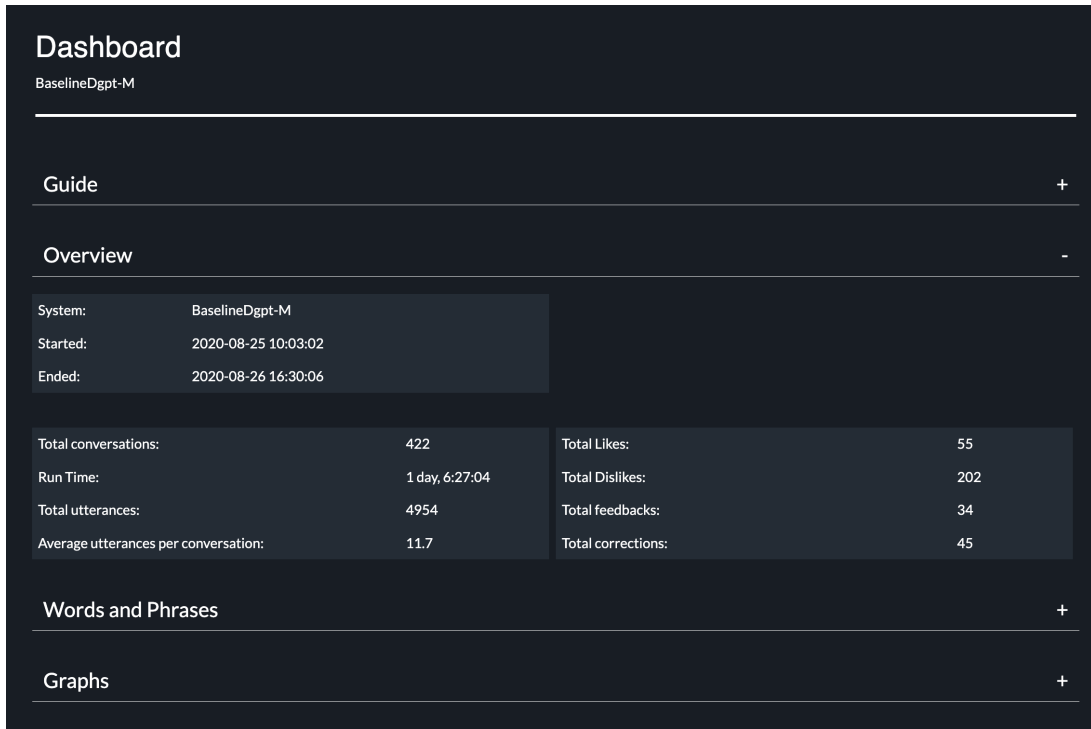


Figure 3: The home page for a system on the DialPort dashboard. General information about the conversations collected from the system are displayed. Sections such as "Words and Phrases" and "Graphs" can be expanded or collapsed to view additional information about the system.

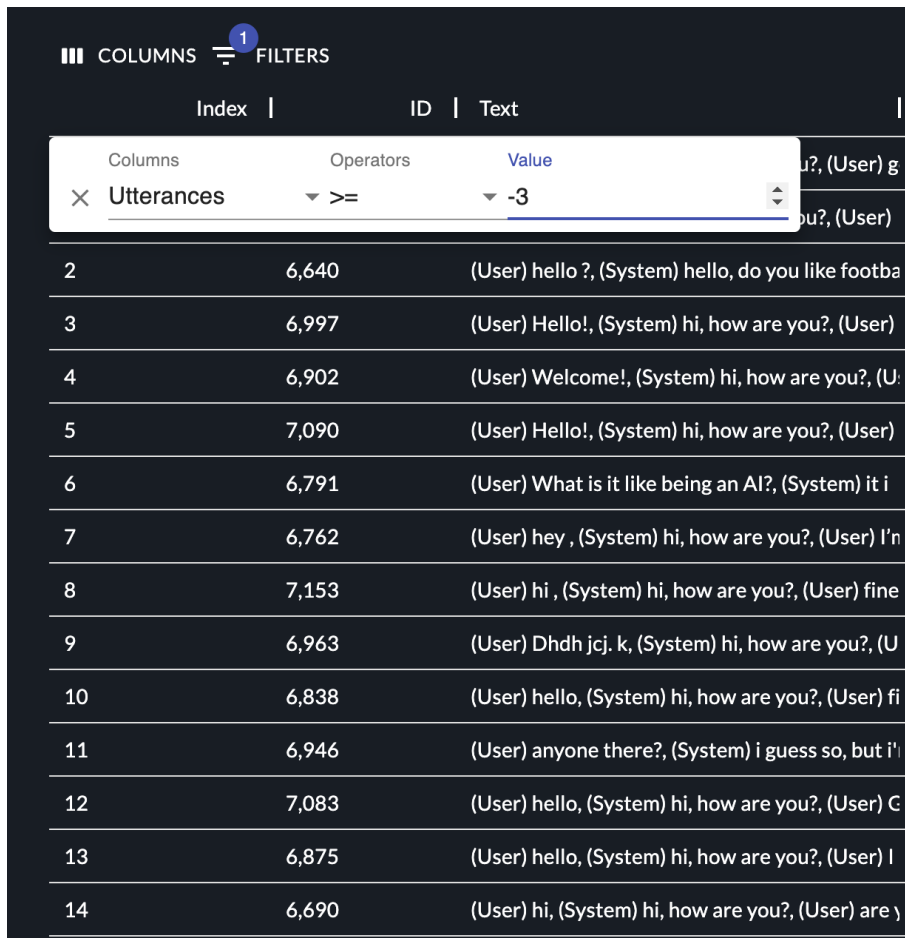


Figure 4: Using the DialPort dashboard to find all conversations in a system with more than 3 utterances