Evaluating Dialogue Systems from the System Owners' Perspectives

Mikio Nakano, Hisahiro Mukai, Yoichi Matsuyama, and Kazunori Komatani

Abstract This position paper proposes to evaluate dialogue systems from the perspectives of the system owners, who decide funding the system development and operation, unlike evaluating based only on the users' satisfaction and experiences as usually done in academic research. We suspect this difference causes a gap between conducting academic research and providing services using dialogue systems. This paper presents an initial list of evaluation criteria from the system owners' perspectives. They consist of three groups, namely the system owners' benefits, costs, and risks. This paper also indicates some of them have been overlooked in academic research, and that considering them will lead to novel research topics.

1 Introduction: A Gap between Academic Research and Practical System Development

So far, numerous innovative techniques have been proposed in academia to improve dialogue systems, but not all of them have been used in practical applications. This gap can be viewed as a "Valley of Death". We conjecture this is because the evaluation criteria of academic research are different from those of the system owners, who fund the development and operation of dialogue systems.

Mikio Nakano
C4A Research Institute, Inc., Tokyo, Japan, e-mail: mikio.nakano@c4a.jp

Hisahiro Mukai
Nextremer Co., Ltd., Kochi, Japan, e-mail: anil.mukai@nextremer.com

Yoichi Matsuyama
Equmenopolis, Inc., Tokyo, Japan, e-mail: yoichim@equ.ai

Kazunori Komatani
SANKEN, Osaka University, Osaka, Japan, e-mail: komatani@sanken.osaka-u.ac.jp

Table 1 Initial List of Evaluation Criteria

Category	Subcategory	Evaluation criterion
Benefits		(1) Revenue from user fees (2) Reduction in labor costs (3) Income from advertisements linked to dialogue content (4) Information collection from users (user demands and dissatisfaction, user situation, etc.) (5) Collection of interaction data for future system development (6) Increase in the sales of products having the system as a component (7) Resident services (for municipalities) (8) Social contribution
Costs	Development costs	(1) Initial system development (2) Data collection and annotation (3) Model training and knowledge-base construction (4) System testing
	Operation costs	(5) Server costs (6) External API service usage fee (7) Improvement in the models, knowledge bases, and system (8) Incident response (9) System advertisements
Risks	Legal risks	(1) Personal information leakage (2) Copyright violation
	Ethical risks	(3) Generating inappropriate system behaviors including slander, defamation, and annoying advertisements (4) Presenting wrong information
	Social risks	(5) Being used for criminals (6) Damaging reputation due to poor performance
	Service discontinuation risk	(7) Termination of external API services

Most of the academic research evaluated the dialogue systems in terms of user satisfaction (e.g., improvement in task success rates and reduction in dialogue cost) [9, 10] and user experiences [1, 3, 6], However, these alone do not provide the system owners with a comprehensive basis for investment decisions. Therefore, for a variety of dialogue systems to be developed and used, we need to assess dialogue systems and their elemental technologies from the system owners' perspectives.

2 Evaluation Criteria from the System Owners' Perspectives

As a starting point of discussion, we enumerated the evaluation criteria that should be considered in developing and operating dialogue systems and grouped them into

three, namely, benefits, costs, and risks (Table 1). Which evaluation criteria need to be considered and which criteria are important more than the others depend on the type of the system and how the system is used. Generally speaking, avoiding risks, especially legal and ethical risks, is the most crucial.

As an example, let us consider a text-based customer service chatbot of a consumer electronics company and assume that it uses retrieval-based response generation [8] based on Sentence-BERT [7]. Its benefits include Benefits (2) and (4) in Table 1, as it can reduce the labor costs for the customer support center and can collect users' demands and dissatisfaction information from the dialogue logs. Its costs are all costs listed in Table 1 except (6) and (9). As for its risks, it is possible to eliminate the Risks (2), (3), and (4) if the response candidates are carefully written by the company. These risks would remain if the system uses large language model-based response generation instead. It is possible to ignore Risks (5) and (7) for this type of system. Risks (1) and (6) exist and must be addressed.

Some of the evaluation criteria in Table 1 correlate with the users' satisfaction and experiences, which have been used as the evaluation criteria in academic research. For example, if a user is satisfied with a dialogue with a customer service chatbot, she/he is more likely to use it again instead of making a phone call to a customer support center, resulting in a reduction in labor cost (Benefit (2)). As another example, generating safer responses [5] (mitigating Risk (3)) is one of the ways to improve user experiences.

However, not all criteria in Table 1 necessarily correlate with the users' satisfaction and experiences. Taking such criteria into account is crucial in designing practical systems. For example, server costs (Cost (5)) are expensive for systems using large machine learning models, and personal information leakage risk (Risk (1)) exists in chat-oriented dialogue systems that ask the user for his/her personal information. Such evaluation criteria tend to have been overlooked in previous research, although there are several exceptions.[1] We therefore think considering them would lead to unexplored and valuable research topics, which we hope bridge the gap between academia and industry.

3 Concluding Remarks

This paper presented an initial list of evaluation criteria from the system owners' perspectives. Some of them have been overlooked in academic research, and we hope this list will lead to novel research topics.

We plan to create a more comprehensive list of evaluation criteria by soliciting opinions from practical system owners and developers. Analyzing which evaluation criteria are important for different types of dialogue systems is also among our future work.

[1] For example, López-Cózar et al. [4] tried to reduce Cost (4) and Fazzinga et al. [2] addressed Risk (1).

References

1. Clark, L., Pantidi, N., Cooney, O., Doyle, P.R., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., Cowan, B.R.: What makes a good conversation?: Challenges in designing truly conversational agents. In: Proceedings of CHI, 475, pp. 1–12 (2019)
2. Fazzinga, B., Galassi, A., Torroni, P.: A privacy-preserving dialogue system based on argumentation. Intelligent Systems with Applications **16**(200113) (2022)
3. Johnston, M., Flagg, C., Gottardi, A., Sahai, S., Lu, Y., Sagi, S., Dai, L., Goyal, P., Hedayatnia, B., Hu, L., Jin, D., Lange, P., Liu, S., Liu, S., Pressel, D., Shi, H., Yang, Z., Zhang, C., Zhang, D., Ball, L., Bland, K., Hu, S., Ipek, O., Jeun, J., Rocker, H., Vaz, L., Iyengar, A., Liu, Y., Mandal, A., Hakkani-Tür, D., Ghanadan, R.: Advancing open domain dialog: The fifth alexa prize socialbot grand challenge. In: Alexa Prize SocialBot Grand Challenge 5 Proceedings (2023)
4. López-Cózar, R., De la Torre, A., Segura, J., Rubio, A., Sánchez, V.: Testing dialogue systems by means of automatic generation of conversations. Interacting with Computers **14**(5), 521–546 (2002)
5. Meade, N., Gella, S., Hazarika, D., Gupta, P., Jin, D., Reddy, S., Liu, Y., Hakkani-Tür, D.: Using in-context learning to improve dialogue safety. In: Proceedings of EMNLP (Findings), pp. 11,882–11,910 (2023)
6. Minato, T., Higashinaka, R., Sakai, K., Funayama, T., Nishizaki, H., Nagai, T.: Design of a competition specifically for spoken dialogue with a humanoid robot. Advanced Robotics **37**(21), 1349–1363 (2023)
7. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the EMNLP-IJCNLP, pp. 3982–3992 (2019)
8. Tao, C., Feng, J., Yan, R., Wu, W., Jiang, D.: A survey on response selection for retrieval-based dialogues. In: Proceedings of IJCAI, pp. 4619–4626 (2021)
9. Ultes, S., Maier, W.: User satisfaction reward estimation across domains: Domain-independent dialogue policy learning. Dialogue and Discourse **12**(2), 81–114 (2021)
10. Walker, M.A., Litman, D.J., Kamm, C.A., Abella, A.: PARADISE: A framework for evaluating spoken dialogue agents. In: Proceedings of ACL, pp. 271–280 (1997)