An Analysis of User Behaviors for Objectively Evaluating Spoken Dialogue Systems

Koji Inoue, Divesh Lala, Keiko Ochi, Tatsuya Kawahara and Gabriel Skantze

Abstract Establishing evaluation schemes for spoken dialogue systems is important, but it can also be challenging. While subjective evaluations are commonly used in user experiments, objective evaluations are necessary for research comparison and reproducibility. To address this issue, we propose a framework for indirectly but objectively evaluating systems based on users' behaviors. In this paper, to this end, we investigate the relationship between user behaviors and subjective evaluation scores in social dialogue tasks: attentive listening, job interview, and first-meeting conversation. The results reveal that in dialogue tasks where user utterances are primary, such as attentive listening and job interview, indicators like the number of utterances and words play a significant role in evaluation. Observing disfluency also can indicate the effectiveness of formal tasks, such as job interview. On the other hand, in dialogue tasks with high interactivity, such as first-meeting conversation, behaviors related to turn-taking, like average switch pause length, become more important. These findings suggest that selecting appropriate user behaviors can provide valuable insights for objective evaluation in each social dialogue task.

Koji Inoue
Kyoto University, Japan, e-mail: inoue.koji.3x@kyoto-u.ac.jp

Divesh Lala
Kyoto University, Japan, e-mail: lala@sap.ist.i.kyoto-u.ac.jp

Keiko Ochi
Kyoto University, Japan, e-mail: ochi.keiko.5f@kyoto-u.ac.jp

Tatsuya Kawahara
Kyoto University, Japan, e-mail: kawahara@i.kyoto-u.ac.jp

Gabriel Skantze
KTH Royal Institute of Technology, Sweden, e-mail: skantze@kth.se

1 Introduction

In the research and development of spoken dialogue systems (SDSs), establishing evaluation methods is a significant challenge [1, 2, 3, 4, 5, 6]. The performance of dialogue understanding and response generation has seen remarkable progress in recent years, thanks to the development of large language models (LLMs). The advancement of LLM research and development has been supported by commonly used evaluation methods in the field, along with extensive text dialogue datasets. Both objective evaluation methods (automatic evaluation) and subjective ones (human evaluation) have been used complementarily. Objective evaluation allows for efficient scalability of evaluation data by enabling automated evaluation measures such as BLEU [7] and Distinct [8]. Additionally, using the same evaluation data as other studies ensures comparability and research reproducibility. On the other hand, subjective evaluation enables a more detailed assessment of each generated system response, capturing aspects that cannot be measured objectively. For instance, evaluating the empathy of responses currently relies on subjective evaluation by humans. Therefore, in SDSs, it is ideal to enhance the efficiency and reproducibility of research in the field by appropriately utilizing both objective and subjective evaluation methods.

For future research and development of speech dialogue systems, it is important to establish objective evaluation criteria. For typical task-oriented dialogues, such as restaurant searches, where the dialogue goal is clear, objective evaluation criteria like the accuracy of slot-filling tasks and the success rate of the dialogue have been utilized [9, 10]. However, the dialogue tasks for SDSs do not always have clearly defined goals. With the advent of conversational AI, SDSs are becoming more realistic in their ability to handle everyday social conversations. This includes brief exchanges like reception and information guidance [11, 12], as well as more extended conversations like counseling [13, 14, 15, 16] and interviews [17, 18, 19, 20]. In these dialogues, while the purpose of the dialogue can be described clearly, the goal itself is not always explicit and gradually becomes clear through the dialogue, fostering mutual understanding and relationships. Consequently, subjective evaluations have been more commonly employed than objective evaluations in the past.

In this study, our objective is to establish an objective evaluation method for SDSs in social scenarios where the goals are not clearly defined. Instead of analyzing system utterances or relying on subjective evaluations from users, we aim to indirectly evaluate the system based on users' spoken-dialogue behaviors during the dialogue, as depicted in Fig. 1. Users' behaviors in this context, such as the rate of speaking time or the number of spoken words, are objectively observable. However, it is not clear which behaviors in specific dialogue tasks can be used as clues for evaluation. To address this, we analyze the relationship between users' behaviors during the dialogue and their subjective evaluations using dialogue data from several social dialogue tasks, including attentive listening, job interview, and first-meeting conversation. The goal is to identify the behaviors that serve as clues for objective evaluation in different dialogue tasks, which will enable the appropriate selection of users' behaviors for objective evaluation in future research and development. By measuring

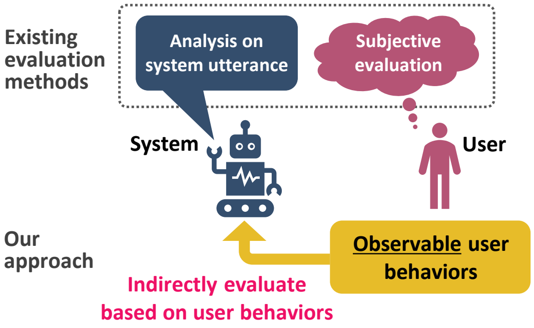

Existing evaluation methods
Analysis on system utterance
Subjective evaluation
System
User
Our approach
Observable user behaviors
Indirectly evaluate based on user behaviors

Fig. 1 Overview of proposed evaluation scheme

and comparing these behaviors, we can achieve the objective evaluation of SDSs across multiple studies, contributing to the overall expansion of the field, which is the ultimate goal of this study. For example, when comparing two systems, it is ideal to have a situation where not only traditional subjective evaluations are conducted, but also numerical behavior data related to the task is reported.

This paper positions itself as an initial analysis of the relationship between user behavior and subjective evaluation toward the aforementioned goal. The contributions of this paper are twofold:

- Propose an objective evaluation scheme for spoken dialogue systems based on users' objective behaviors
- Clarify the users' behaviors related to subjective evaluation in social dialogue tasks such as attentive listening, job interview, and first-meeting conversations

The rest of this paper is organized as follows. The proposed evaluation scheme is introduced in Section 2. The dialogue data used is explained in Section 3. Then, the relationship between users' behaviors and subjective evaluation is analyzed in Section 4. Finally, this paper concludes in Section 5.

2 Proposed Evaluation Scheme

The evaluation method proposed in this study is designed to indirectly assess SDSs by analyzing users' behaviors during the dialogue. The focus is on specific behaviors related to spoken dialogue, including speech, language, and dialogue features. It is important to note that future research will explore additional modalities, such as eye-gaze behaviors analyzed through image processing.

User behaviors analyzed in this study are listed below.

- Utterance time / min.
- Number of utterances (IPU segments) / min.
- Number of utterance words / min.

- Number of unique utterance words / min.
- Number of utterance content words / min.
- Number of unique utterance content words / min.
- Number of backchannels / min.
- Number of fillers / min.
- Number of laughs / min.
- Number of disfluencies / min.
- Average switching pause length

The criterion for dividing an IPU (inter pausal unit) is set as a silent interval of 200 milliseconds or more. Since the dialogue data used in this study is in the Japanese language, word segmentation is performed using MeCab[1]. Content words are defined as nouns, verbs, adjectives, adverbs, and conjunctions. Backchannels are defined as responsive interjections such as "yes" or "uh-huh", and emotional interjections such as "hmm" or "oh". Fillers are expressions used to bridge gaps in conversation, such as "um" or "well", while speech disfluencies are expressions used for self-correction, such as "spe, specifically". The linguistic behaviors described above were calculated based on manually transcribed data in this study. Switching pause length refers to the duration of the silent interval when the speaking floor transitions from the system to the user.

Intuitively, the above-mentioned behaviors vary depending on how natural the interaction with the system is. This can be understood by comparing human-system dialogues and human-human dialogues. For example, in human-system dialogues, users typically speak clearly and with a limited vocabulary. On the other hand, in human-human dialogues, it is natural to speak fluently and with a diverse vocabulary. Regarding backchannels, users in human-system dialogues rarely use them, while in human-human ones, backchannels are used frequently. Studies have shown that the average switching pause length in human-human dialogues is almost zero seconds [21, 22, 23], while in human-system dialogues, it often takes around 1 to 3 seconds. Based on these characteristics, evaluating a spoken dialogue system using the above-mentioned behaviors can be seen as evaluating its naturalness and similarity to human behavior. In this study, we aim to identify specific behaviors related to user subjective evaluation in three different social dialogue tasks.

3 Dialogue Data

The dialogue data used in this study is explained below. These data were recorded using android ERICA [24], as shown in Figure 2. It is important to note that in order to introduce variation in the quality of the dialogue, a mixture of dialogues between human-system and human-human (referred to as WOZ: Wizard-of-OZ) was used, depending on the dialogue task. Table 1 summarizes the number of dialogues used in this study based on the type of setting. In the case of WOZ, there was an operator in

[1] https://taku910.github.io/mecab

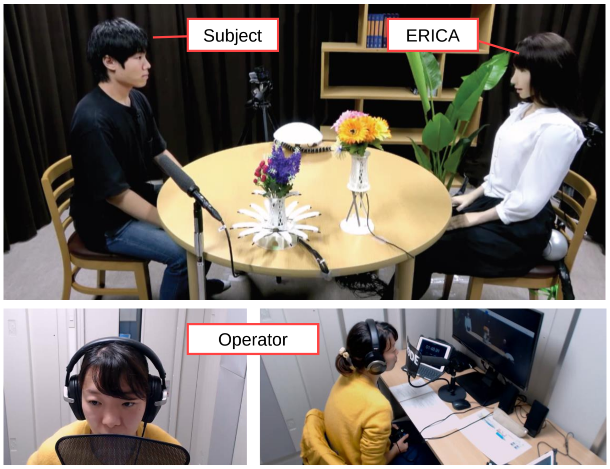


Fig. 2 Scene of dialogue data collection (Bottom: an operator in WOZ)

Table 1 Number of dialogue in different settings (The names of dialogue tasks are represented as AL, JI, and FMC for Attentive Listening, Job Interview, and First-meeting Conversation, respectively.)

Dialogue task	AL	JI	FMC
# Autonomous (human-system)	19	86	0
# WOZ (human-human)	50	0	50
Total	69	86	50

Table 2 Characteristics of dialogue tasks targeted in this study

Dialogue task	Attentive Listening	Job Interview	First-meeting Conversation
System role	Listening	Asking	All
Initiative	User	System	Mixed
Majority of utterances	User	User	Both
Majority of backchannels	System	System	Both
Turn-taking	Few	Explicit	Complicated

a separate room, and the operator's voice was played through the android's speaker. Non-verbal expressions, such as the android's gaze and gestures, were controlled by the operator using a handheld controller. All the dialogues in this study were conducted in Japanese.

In this study, we utilize dialogue data for three different social dialogue tasks, each with its own unique characteristics. Table 2 provides a summary of these characteristics, specifically in terms of implementing a spoken dialogue system for each task. Given the variations in the initiative of dialogue, as well as the frequency and clarity of turn-taking, it is anticipated that users' behaviors related to subjective evaluation will also differ.

3.1 Attentive Listening

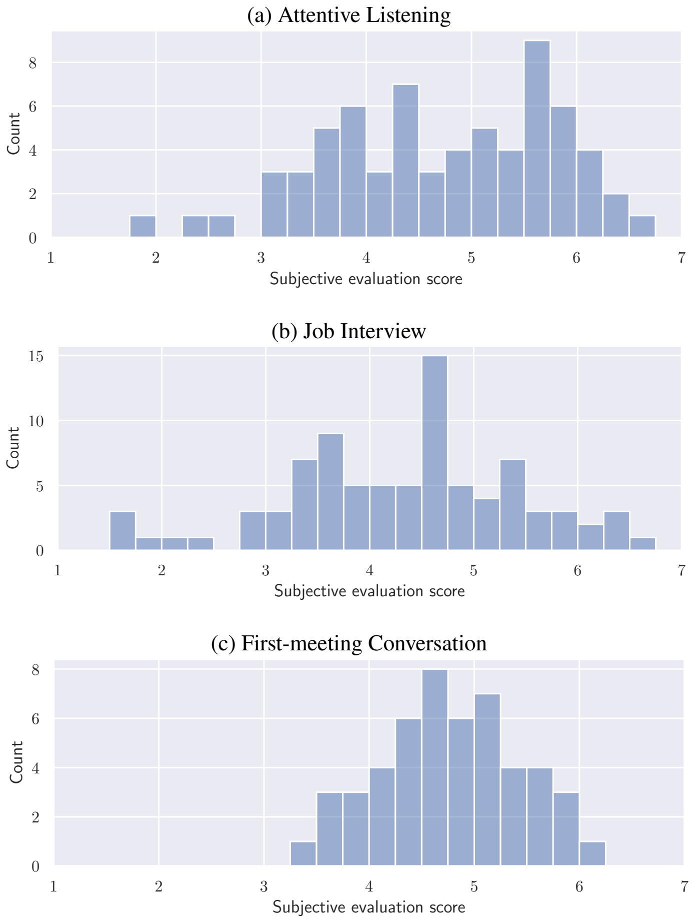
Attentive Listening involves the task of the listener (system) actively listening to the user's talk. The listener responds through various types of utterances such as backchanneling and elaborating questions. The authors have developed a real-time spoken dialogue system capable of generating listener responses [25]. This system was used to record the dialogue data for this study. We recruited 69 university students as users. They engaged in an 8-minute dialogue with the attentive listening dialogue system, focusing on the topic of "difficulties during the COVID-19 pandemic." The dialogue task also included human-human dialogue (WOZ). In total, there were 20 interactions with the autonomous system and 50 WOZ dialogues, resulting in a total of 70 dialogues.

After each dialogue, a subjective evaluation was conducted on the 19 items created in our previous study [25]. These items include statements such as "The words uttered by the robot were natural," "The robot understood the talk," and "The robot showed empathy towards me." The participants were asked to rate each item on a 7-point scale ranging from 1 to 7. For this study, we used 18 items, excluding one item that showed variation in interpretation. We calculated the average value for each dialogue (per participant), which serves as the dependent variable. In other words, the goal is to predict this average rating value based on the user's behaviors mentioned in the previous section. Fig. 3 (a) shows the distribution of the evaluation scores. Overall, the scores are somewhat high, but it can be observed that there are also a certain number of participants who gave low scores.

3.2 Job Interview

Job interview is a dialogue between an interviewer (system) and an applicant (user), where the applicant answers questions posed by the interviewer. For this study, we utilized a job interview dialogue system developed by the authors [20], and all interactions were conducted using this autonomous system. In this task, we treated it as a practice job interview and recruited 43 university students. Each participant engaged in two times of dialogue with two different systems, which differed in the presence of follow-up questions from the system. Follow-up questions are inquiries that delve deeper into the applicant's previous response, such as extracting keywords and asking, "Could you please provide more details about (keyword)?" Before starting the dialogue, participants were asked to select their desired industry and company, and also prepare answers to several expected questions.

After each dialogue, subjective evaluations were conducted on 19 items created in our previous study [20]. These items include statements such as "I was nervous during the interview," "Thanks to the interview, I was able to notice my weak points," and "The interviewer understood my answers." Participants were asked to rate each item on a 7-point scale ranging from 1 to 7. The average value was calculated for each dialogue using the 18 items, excluding the one item that seemed to have variation in


(a) Attentive Listening

(b) Job Interview

(c) First-meeting Conversation

Fig. 3 Distribution of subjective evaluation scores in each dialogue task

interpretation. This average value was then used as the dependent variable. Fig. 3 (b) shows the distribution of the evaluation scores. It can be seen that there is slightly more variation compared to those of Attentive Listening.

3.3 First-meeting Conversation

First-meeting conversation is a dialogue that allows both participants to get to know each other and establish a relationship. For our study, we recruited 50 university and graduate students as users to engage in conversations with a robot, simulating a first-meeting scenario. Note that the system was completely operated and controlled by the WOZ setup. Furthermore, the participants were given a list of commonly discussed topics in first-meeting conversations prior to the interaction.

After each dialogue, a subjective evaluation was conducted on 18 items. These items include "I was able to get to know the other person well," "The atmosphere of the conversation was pleasant," "I had a favorable impression of the other person." Participants were asked to rate each item on a 7-point scale, ranging from 1 to 7. The average value of these 18 items was calculated for each dialogue (participant) and used as the dependent variable. Fig. 3 (c) shows the distribution of the evaluation scores. It can be observed that the scores are not as distributed as the other two tasks. The reasons for this could be that all dialogues were performed by the WOZ operator, resulting in overall high quality of the conversations. Additionally, subjective evaluations of this task are often ambiguous, making it difficult to determine superiority or inferiority.

4 Analysis

The relationship between the subjective evaluation scores mentioned in the previous section and the user behavior mentioned in Section 2 was investigated. To analyze this relationship, we utilized SHAP (SHapley Additive exPlanations) [26][2]. SHAP analysis calculates the contribution level (SHAP value) of each feature in the output value of a trained model. Specifically, the SHAP value ϕ_f of a feature (behavior) $f \in \boldsymbol{x}$ is calculated as follows:

$$\phi_f = \sum_{A \subseteq \boldsymbol{x} \setminus \{f\}} \frac{|A|!(|\mathbf{x}| - |A| - 1)!}{|\boldsymbol{x}|!} (y(A \cup \{f\}) - y(A))$$

Note that A represents a subset that excludes the feature f, and $y(\cdot)$ represents the output of the model when using the given set of features. In other words, it is the average difference between the output values when using the feature f and when not using it, for all subsets of the features. The larger the absolute value of this SHAP value, the greater the interpretation that the corresponding feature has a larger impact on the trained model. One advantage of this analysis method is that it calculates the influence of each feature, taking into account the interaction among the features. The behaviors described so far are not independent and co-occur during the dialogue, so it is reasonable to analyze them considering the interaction among them, just like SHAP does.

The procedure for applying SHAP is as follows: For each dialogue task, we trained a regression model using the user's behavioral features described in Section 2 as explanatory variables and the subjective evaluation scores described in Section 3 as the target variable. XGBoost was used as the regression model in this case. Then,

[2] `https://pypi.org/project/shap/`

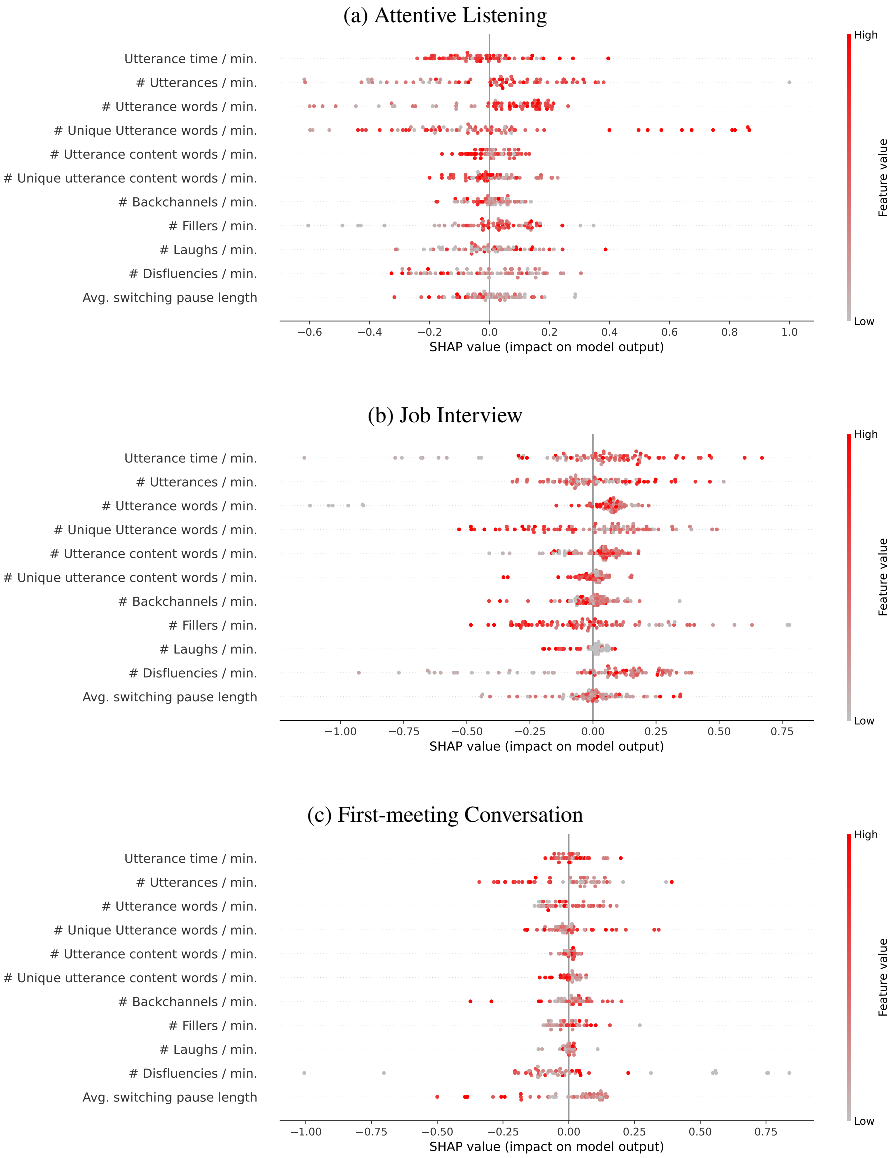
User behavior	Dialogue task		
	AL	JI	FMC
Utterance time / min.	0.098 (7.69)	**0.212** (15.07)	0.043 (5.42)
# utterances (IPU segments) / min.	**0.199** (15.63)	**0.124** (8.80)	**0.126** (15.87)
# utterance words / min.	**0.162** (12.72)	**0.145** (10.31)	0.070 (8.84)
# unique utterance words / min.	**0.242** (18.97)	**0.183** (12.98)	0.068 (8.58)
# utterance content words / min.	0.048 (3.74)	0.087 (6.15)	0.016 (1.98)
# unique utterance content words / min.	0.066 (5.16)	0.042 (2.96)	0.025 (3.21)
# backchannels / min.	0.050 (3.91)	0.067 (4.73)	0.062 (7.83)
# fillers / min.	**0.111** (8.66)	**0.187** (13.26)	0.047 (5.99)
# laughs / min.	0.093 (7.28)	0.042 (2.98)	0.016 (2.08)
# disfluencies / min.	**0.129** (10.13)	**0.224** (15.88)	**0.192** (24.23)
Average switching pause length	0.078 (6.10)	0.097 (6.87)	**0.126** (15.95)

Table 3 Absolute mean of SHAP value in each behavior and dialogue task (The numbers in parentheses represent the percentages in each dialogue task. The bold numbers represent the absolute mean SHAP values exceeding 0.100. The names of dialogue tasks are represented as AL, JI, and FMC for Attentive Listening, Job Interview, and First-meeting Conversation, respectively.)

we computed the SHAP value for each feature using the aforementioned method for the trained model.

Table 3 presents the average absolute SHAP value for each behavior in the dialogue tasks. Across all tasks, the number of utterances (IPU segments) and the number of disfluencies remained consistently high. When comparing Attentive Listening and Job Interview, similar trends were observed. Both tasks emphasize the user's speaking ability, leading to the importance of metrics like the number of utterances, words, and those unique numbers. However, the number of disfluencies was given greater significance in the job interview task. This is because effective questioning is crucial in Job Interview, so the behavior of disfluencies reflects the evaluation of the system. For First-meeting Conversation, in addition to the number of utterances and disfluencies, a higher SHAP value was observed for the average switch pause length. This type of dialogue involves mixed initiative and frequent exchanges, making smooth turn-taking vital for system evaluation. It should also be noted that First-meeting Conversation, being conducted in the WOZ setting, tends to be more dynamic compared to the autonomous setting so that these interactive factors would be more indicative. In summary, in dialogue tasks where user utterances are primary, such as Attentive Listening and Job Interview, indicators like the number of utterances and words play a significant role in evaluation. Additionally, in formal situations like Job Interview, disfluencies are also effective. On the other hand, in dialogue tasks with high interactivity, such as First-meeting Conversation, the importance of behaviors related to turn-taking, like average switch pause length, increases.

To analyze the specific tendencies of SHAP values, we present the distribution of SHAP values per behavior in each dialogue task in Fig. 4. Each point represents a dialogue, and the shading indicates the magnitude of the behavior feature value. The correlation between the number of utterances in Attentive Listening (Fig. 4 (a)) and the number of disfluencies in Job Interview (Fig. 4 (b)) with the magnitude of SHAP


(a) Attentive Listening

(b) Job Interview

(c) First-meeting Conversation

Fig. 4 Distributions of SHAP value for each behavior and dialogue task

values suggests a positive correlation between subjective evaluation scores and these behaviors. In contrast, the number of unique utterance words in Job Interview (Fig. 4 (b)), and the number of disfluencies and the average switching pause length in First-meeting Conversation (Fig. 4 (c)) exhibit an inverse relationship with SHAP values. Therefore, it can be concluded that there is a negative correlation between subjective evaluation scores and these behaviors.

Finally, we investigated the predictive performance of the regression model trained using leave-one-out cross-validation on subjective evaluation scores. The input consisted of all the behavioral features (11 dimensions) mentioned earlier, and the output was the subjective evaluation score of the system shown in Figure 3. XGBoost was used for the model. When calculating the mean absolute error for the test data, we obtained the following results: 0.970 for Attentive Listening, 0.953 for Job Interview, and 0.683 for First-meeting Conversation. In other words, for all dialogue tasks, the errors were smaller than the granularity of the evaluation scores (1.000). This confirms the feasibility of the proposed evaluation framework and the validity of users' behaviors selected for this study.

5 Conclusion

In this paper, we proposed an objective evaluation method for spoken dialogue systems used in social dialogue tasks. We examined the relationship between users' behaviors and subjective evaluation scores for three different dialogue tasks: attentive listening, job interview, and first-meeting conversation. Our findings revealed that the behaviors associated with the subjective evaluations vary depending on the characteristics of the dialogue task. Additionally, we evaluated the proposed evaluation method framework through cross-validation and found that it can accurately predict subjective evaluation scores from the users' behaviors, with an absolute error smaller than the evaluation score's granularity. Moving forward, our future work aims to expand the analysis by considering more target behaviors and dialogue tasks. We also plan to further develop the proposed evaluation method framework for other research and development purposes.

Acknowledgement

This work was supported by JST ACT-X (JPMJAX2103), JST Moonshot R&D (JPMJPS2011), and JSPS KAKENHI (JP19H05691 and JP23K16901). The authors also express appreciation to Dr. Masato Komuro for his insightful comments on this research.

References

1. Marilyn Walker, Diane Litman, Candace A Kamm, and Alicia Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *Annual Meeting of the Association for Computational Linguistics (ACL) and Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 271–280, 1997.

2. Doreen Ying Ying Sim and Chu Kiong Loo. Extensive assessment and evaluation methodologies on assistive social robots for modelling human–robot interaction – a review. *Information Sciences*, 301:305–344, 2015.
3. Alaa Abd-Alrazaq, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, and Kerstin Denecke. Technical metrics used to evaluate health care chatbots: Scoping review. *Journal of medical Internet research*, 22(6), 2020.
4. Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810, 2021.
5. Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. Automatic evaluation and moderation of open-domain dialogue systems. *arXiv preprint, 2111.02110*, 2021.
6. Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56:3055–3155, 2023.
7. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, page 311–318, 2002.
8. Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 110–119, 2016.
9. Matthew Henderson, Blaise Thomson, and Jason D. Williams. The second dialog state tracking challenge. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 263–272, 2014.
10. Pawe l Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. MultiWOZ - A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5016–5026, 2018.
11. William Swartout, David Traum, Ron Artstein, Dan Noren, Paul Debevec, Kerry Bronnenkant, Josh Williams, Anton Leuski, Shrikanth Narayanan, Diane Piepol, Chad Lane, Jacquelyn Moriel, Priti Aggarwal, Matt Liewer, Jen-Yuan Chiang, Jillian Gerten, Selina Chu, and Kyle White. Ada and grace: Toward realistic and engaging virtual museum guides. In *The annual conference on Intelligent Virtual Agents (IVA)*, pages 286–300, 2010.
12. Takamasa Iio, Satoru Satake, Takayuki Kanda, Kotaro Hayashi, Florent Ferreri, and Norihiro Hagita. Human-like guide robot that proactively explains exhibits. *International Journal of Social Robotics*, 12:549–566, 2020.
13. David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis P. Morency. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1061–1068, 2014.
14. Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, and Enrico Coiera. Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258, 2018.
15. Arielle AJ Scoglio, Erin D Reilly, Jay A Gorman, and Charles E Drebing. Use of social robots in mental health and well-being research: Systematic review. *Journal of medical Internet research*, 21(7), 2019.
16. Samira Rasouli, Garima Gupta, Elizabeth Nilsen, and Kerstin Dautenhahn. Potential applications of social robots in robot-assisted interventions for social anxiety. *International Journal of Social Robotics*, 14:1–32, 2022.
17. Sangdo Han, Kyusong Lee, Donghyeon Lee, and Gary Geunbae Lee. Counseling dialog system with 5W1H extraction. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, 2013.

18. Michael Johnston, Patrick Ehlen, Frederick G. Conrad, Michael F. Schober, Christopher Antoun, Stefanie Fail, Andrew Hupp, Lucas Vickers, Huiying Yan, and Chan Zhang. Spoken dialog systems for automated survey interviewing. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 329–333, 2013.
19. Zhou Yu, Vikram Ramanarayanan, Patrick Lange, and David Suendermann-Oeft. An open-source dialog system with real-time engagement tracking for job interview training applications. In *International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, 2017.
20. Koji Inoue, Kohei Hara, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. Job interviewer android with elaborate follow-up question generation. In *International Conference on Multimodal Interaction (ICMI)*, pages 324–332, 2020.
21. Stephen C. Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, 6(731):1–17, 2015.
22. Gabriel Skantze. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language*, 67:1–26, 2021.
23. Mark Dingemanse and Andreas Liesenfeld. From text to talk: Harnessing conversational corpora for humane and diversity-aware language technology. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5614–5633, 2022.
24. Koji Inoue, Pierrick Milhorat, Divesh Lala, Tianyu Zhao, and Tatsuya Kawahara. Talking with ERICA, an autonomous android. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 212–215, 2016.
25. Koji Inoue, Divesh Lala, Kenta Yamamoto, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. An attentive listening system with android ERICA: Comparison of autonomous and WOZ interactions. In *Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 118–127, 2020.
26. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Neural Information Processing Systems (NeurIPS)*, 2017.