Understanding User Satisfaction with Task-oriented Dialogue Systems

Clemencia Siro
University of Amsterdam
Amsterdam, The Netherlands
c.n.siro@uva.nl

Mohammad Aliannejadi
University of Amsterdam
Amsterdam, The Netherlands
m.aliannejadi@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

Dialogue systems (DSs) are evaluated depending on their type and purpose. Two categories are often distinguished: (1) task-oriented dialogue systems (TDSs), which are typically evaluated on utility, i.e., their ability to complete a specified task, and (2) open-domain chat-bots, which are evaluated on the user experience, i.e., based on their ability to engage a person. What is the influence of *user experience* on the user satisfaction rating of TDSs as opposed to, or in addition to, *utility*? We collect data by providing an additional annotation layer for dialogues sampled from the ReDial dataset, a widely used conversational recommendation dataset. Unlike prior work, we annotate the sampled dialogues at both the turn and dialogue level on six dialogue aspects: *relevance*, *interestingness*, *understanding*, *task completion*, *efficiency*, and *interest arousal*. The annotations allow us to study how different dialogue aspects influence user satisfaction. We introduce a comprehensive set of user experience aspects derived from the annotators' open comments that can influence users' overall impression. We find that the concept of satisfaction varies across annotators and dialogues, and show that a relevant turn is significant for some annotators, while for others, an interesting turn is all they need. Our analysis indicates that the proposed user experience aspects provide a fine-grained analysis of user satisfaction that is not captured by a monolithic overall human rating.

CCS CONCEPTS

• **Computing methodologies** → Discourse, dialogue and pragmatics; • **Information systems** → Users and interactive retrieval; Evaluation of retrieval results.

KEYWORDS

Fine-grained user satisfaction, task-oriented dialogues, user experience

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531798

ACM Reference Format:
Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2022. Understanding User Satisfaction with Task-oriented Dialogue Systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531798

1 INTRODUCTION

Recent research into the evaluation of conversational systems such as dialogue systems and conversational recommender systems has proposed automatic metrics that are meant to correlate well with human judgements [3]. Many of these standard evaluation metrics have been shown to be ineffective in dialogue evaluation [3, 16]. As a consequence, a significant amount of dialogue research relies on human evaluation to measure a system's effectiveness. Recently, estimating a user's overall satisfaction with system interaction has gained momentum as the core evaluation metric for task-oriented dialogue system (TDS) [8, 13]. Though useful and effective, overall user satisfaction does not necessarily give insights on what aspect or dimensions the TDS is performing well. Knowing why a user is satisfied or dissatisfied helps the conversational system recover from an error and optimise toward an individual aspect to avoid total dissatisfaction during an interaction session.

Understanding *user satisfaction* with a TDS at a fine-grained level is vital at both the design and evaluation stages. Metrics such as engagement, relevance, and interestingness have been investigated to understand fine-grained user satisfaction and how they determine a user's overall satisfaction in different scenarios and applications [7, 20, 24]. For TDS, user satisfaction is modelled as an evaluation metric for measuring a system's ability to achieve a functional goal with high accuracy (i,e task success rate and dialogue cost) [19]. Unlike TDS, the main focus in chat-bot evaluation is on the user experience during interaction (i.e how engaging, interesting etc. the system is) [14].

The metrics proposed in [7, 24] provide a granular analysis on how they influence user satisfaction for chat-bots – but it is not known how these aspects influence user satisfaction of TDSs [see, e.g., 12, 26]. In this study, we focus on understanding the significance of several dialogue aspects on overall impression rating of a TDS. We investigate some of the metrics from [7] originally introduced for chat-bots (viz. *interestingness*, *relevance*, and *understanding*) and how they determine a user's overall impression of a TDS. We also propose a new aspect, *interest arousal*, as a metric in measuring TDS effectiveness. We find that this newly proposed metric achieves the highest correlation with overall *user satisfaction* with a TDS, compared to other metrics that focus on the user experience, with a Spearman's ρ of 0.7903.

To understand the influence of the dialogue aspects, we collect human quality annotations for the recommendation dialogues (ReDial) dataset [15]. The dialogues are annotated at both the turn and dialogue levels on several aspects stated above and extensive analysis is conducted on the annotated dataset. With these we sought to answer the following questions: What dialogue aspects influence overall impression of TDSs? and; What role does the utility and user experience dimensions play when rating the overall impression of a task oriented dialogue?

The contributions we make in this paper are: (i) We add an extra annotation layer for the ReDial dataset. A human quality annotation effort is set up on Amazon Mechanical Turk (AMT),[1] for the annotation of 40 sampled dialogues at the turn and dialogue level on six dialogue aspects: *relevance*, *interestingness*, *understanding*, *task completion*, *efficiency*, and *interest arousal*. (ii) We analyse the annotated dataset to identify dialogue aspects that influence the overall impression. (iii) We propose additional dialogue aspects with significant contributions to the overall impression of a TDS. We classify the annotators' open comments left in the justification box into different categories. Apart from the six dialogue aspects investigated in this study, *natural conversation*, *success in the last interaction* and *repetition* are among other aspects stated by the annotators that influenced their overall impression.

2 RELATED WORK

User satisfaction. Kelly [10] defines *user satisfaction* as the fulfilment of a user's specified desire or goal. User satisfaction has gained popularity as an evaluation metric of information retrieval (IR) systems based on implicit signals [8, 9, 11–13]. Due to the reliance on the user's intelligence and emotions to measure user satisfaction, user satisfaction in information systems depends on the user's interaction experience and the fulfilment of their specified desires, and goals [10]. Factors such as system effectiveness, user effort, user characteristics and expectations influence a user's satisfaction rating for IR systems [1]. In TDS, user satisfaction is measured by rating a dialogue at both the turn and dialogue level on overall impression [2, 23]. We rate both the turn and dialogue level with fine-grained aspects of user satisfaction in our work.

Dialogue qualities. Dialogue systems are often evaluated on their overall impression [3]. Recently, research into fine-grained user satisfaction has increased. Walker et al. [25] propose a framework for evaluating dialogues in a multi-faceted way. It measures several dialogue qualities and combines them to estimate user satisfaction. Mehri and Eskenazi [18] develop an automatic evaluation metric for evaluating dialogue systems at a fine-grained level, including interestingness, engagingness, diversity, understanding, specificity, and inquisitiveness. Several other publications have investigated human evaluation of dialogue systems on different dialogue qualities [see, e.g., 4, 20]. Our work rates user satisfaction at both turn and dialogue level on six fine-grained user satisfaction aspects, unlike previous research rating both levels on overall impression. We also propose a new aspect, *interest arousal*, which strongly correlates with the overall impression.

[1] https://mturk.com

3 METHODOLOGY

To establish dialogue aspects that lead to overall user satisfaction with a TDS, we create an additional annotation layer for the ReDial dataset. We set up an annotation experiment on Amazon Mechanical Turk (AMT) using so-called master workers. The AMT master workers annotate a total of 40 conversations on six dialogue aspects. Following Mehri and Eskenazi [18] work, we hand-selected three system responses from each conversation for turn-level annotation. Each response has two previous turns as context plus the following user utterance. Unlike [23], we ask the annotators to decide on the label by considering the user's next action. We display all three turns on a single page and instruct the annotators to answer questions for each turn. After completing the turn-level annotation, the same annotators are taken to a new page where they provide dialogue-level annotations on the same dialogue. The annotators cannot return to the turn-level annotation page. This restriction is based on two considerations: (i) to avoid bias of annotators on the turn-level labels when making decisions on the dialogue-level annotations; and (ii) to prevent annotators from going back to change their turn-level ratings. With this we aim to capture how well an annotator's turn ratings correlate with their dialogue-level ratings and overall ratings.

We crowd-source the annotations to enable scalable and efficient annotation labels while capturing different points of view. We refined the instructions in a series of pilot studies and internal reviews to ensure the workers understood the task. Moreover, we clearly define each evaluation aspect, backed by real examples from the dataset. In the instructions, we stress the fact that the workers need to base their judgements on evidence present in the dialogue (e.g., "I really liked your suggestion.") to show relevance, not their personal taste or guess. We annotate each dialogue with 5 workers. The annotators answered two questions for each turn and five at the dialogue level. In total, the annotated dataset includes 1,200 turn-level and 1000 dialogue-level data points. At the end of each dialogue, we ask each annotator to leave an open comment to justify their overall impression rating of the TDS. We obtain 200 open comments from 32 workers, which we use for analysis to propose additional aspects to be studied with respect user satisfaction as shown in Table 3.

Recommendation dialogue dataset. The ReDial dataset [15] is a large dialogue-based human-human movie recommendation corpus. It consists of 11,348 dialogues. One person is the movie seeker, and the other is the recommender. The movie seeker should explain the kind of movie they like and ask for suggestions. The recommender tries to understand the seeker's movie tastes and recommends movies. This dataset is categorised as both chit-chat and task-oriented since the recommender needs to discover the seeker's movie preference before recommending.

Turn-level annotation. At the turn level, given the previous context, and the next user utterance, we instruct the workers to assess the system's response according to two fine-grained aspects on a scale of 1 to 3:

Relevance (1–3): The system's response is appropriate to the previous turn and fulfils the user's interest [7, 18, 20, 21].

Interestingness (1–3): The system makes chit-chat while presenting facts [7, 18, 20, 22].

Table 1: Correlation of overall impression with turn-level and dialogue-level annotations. All correlations in this table are statistically significant ($p < 0.01$).

Level	Aspect	Spearman's ρ	Pearson's r
Turn	Relevance	**0.5199**	**0.5622**
	Interestingness	0.3374	0.3603
Dialogue	Understanding	0.7589	0.7928
	Task completion	0.7895	0.8280
	Interest arousal	**0.7903**	**0.8341**
	Efficiency	0.5946	0.5697

The options for each aspect were: *No, Somewhat, Yes*. For *Relevance*, we also provided a *Not applicable* option in case an annotator believes there is not enough evidence to determine whether the response is relevant (e.g., if no movie is recommended).

Dialogue-level annotation. The workers rate the system's quality considering the entire dialogue for the dialogue-level annotation. This level is evaluated based on four aspects. For *understanding, task completion, and interest arousal* we provide three options: *No, Somewhat, Yes*. For *efficiency* raters are given a binary choice [5, 13]. We also ask workers to rate the system on overall impression. The *overall impression* is rated on a Likert scale of 1 to 5 (with 1 being very dissatisfied and 5 very satisfied). Below is a summary of the definitions we provide our workers for the dialogue aspects:

Understanding (1–3): The system understands the user's request and fulfils [5, 18, 24].

Task completion (1–3): The system makes suggestions that the user finally accepts [13].

Efficiency (0–1): The system can make suggestions that meet the user's interest within the first three interactions [5, 13].

Interest arousal (1–3): The system attempts to intrigue the user's interest into accepting a suggestion they are not familiar with.

Overall impression (1–5): The worker's overall impression of the system's performance, given the dialogue context [7, 13, 18, 20, 22].

Finally, we ask the workers to justify their rating on *overall impression*. We use the justifications to contextualise the given ratings and analyse and discover additional aspects that affect the quality of a dialogue.

Participants. A total of 32 AMT workers took part in the human annotation effort, 18 female and 14 male. Their ages range from 18 to 49.

4 RESULTS AND ANALYSIS

This section presents the results from our annotation effort and an analysis of the annotators' comments on their overall impression labels. We intend to answer the following questions: (RQ1) To what extent do turn-level aspects correlate with the overall user impression in TDS? And (RQ2) What dialogue-level aspects have a more significant influence on the overall user impression?

As explained above, apart from rating dialogues on six dialogue aspects, annotators also rated the system on overall impression. We use these ratings to classify the dialogues into several categories. First, a dialogue is *satisfactory* if it has a majority rating of 3 or more; it is *unsatisfactory* if it has a majority rating of less than 3. Second, we label a dialogue as being *subjective* if: (i) two or more annotators selected labels that indicated both satisfactory and unsatisfactory, and (ii) only two annotators agreed on a label whereas the other three selected different labels from each other. That is, we have four different labels selected. There are 26 *satisfactory* and 6 *unsatisfactory* dialogues; 8 dialogues are categorised as *subjective*. Inter annotator agreement for the overall impression ratings was fair, with a Fleiss Kappa score of 0.412.

Table 2: Determinant coefficients computed with regression showing the effect size of all aspects to overall impression.

	Aspect	Utility	User experience	R^2
Turn (T)	Relevance (R)	+		0.568
	Interestingness (I)		+	0.258
	R + I	+	+	**0.583**
Dialogue (D)	Understanding (U)		+	0.629
	Task completion (TC)	+		0.686
	Interest arousal (IA)		+	0.696
	Efficiency (E)		+	0.325
	IA + TC + U + E	+	+	**0.825**
D + T	R + TC	+		0.761
	IA + U + I + E		+	0.803
	IA + TC + U + I + E + R	+	+	**0.844**

4.1 Turn-level aspects influencing overall impression

We compute Spearman's ρ and Pearson's r correlation coefficients with the overall impression for each turn and average across the three turns; see Table 1 (top). The *relevance* aspect exhibits the highest correlation at the turn-level. Out of the dialogues classified as satisfactory, 46% of the turns were rated relevant (= 3) compared to 31% rated interesting (= 3). Hence, more system responses are found to be relevant but not interesting as a TDS is traditionally expected to optimise towards task success and not engagement.

When a turn is relevant, the dialogue's overall impression is more likely to be satisfactory (96% of the turns). The same does not hold for a nonrelevant turn (43% of the turns led to a satisfactory dialogue), suggesting that in this case, the user's overall impression depends not only on *relevance* but on other dialogue aspects too. A system's success in making a successful suggestion[2] in the final turn has more weight on the overall impression than the preceding turns. This conforms to the findings of [13, 17], which shows that the latest interactions with a system more influence the overall satisfaction of users. Although relevance is essential in determining the overall impression, it is not the only influencing aspect.

4.2 Dialogue-level aspects with significant influence on overall impression

In Figure 1 we plot the distribution of the ratings for the dialogue-level aspects against overall impression. We see a clear dependency of the *overall impression* on the *interest arousal* aspect; out of the

[2]A successful suggestion is a movie suggestion that the user accepts.

Table 3: Additional aspects captured from the open comments. The % show how often the aspect was stated.

Aspect	Definition	Annotator comment
Opinion (2.4%)	System expresses general opinions on a generic topic or expressing strong personal opinion	"I don't think that the system should be providing its own opinions on the movies"
Naturalness (5.42%)	The flow of the conversation is good and fluent	"The conversation flow naturally from one exchange to the next"
Success on the last interaction (10.8%)	System gets better as time goes by	"The system finally recommends a good movie at the very end"
Repetition (1.8%)	The system repeating itself or suggestions	"The system has good suggestions, but it repeats itself over and over which is strange."
User (4.21%)	User's actions influencing the overall impression	"The system was being helpful but the user was difficult in answering preference questions"

dialogues classified as satisfactory, 73% were rated high in terms of interest arousal (see Figure 1a). We also notice that all dialogues rated low (= 1) are unsatisfactory overall. Thus the ability of a TDS to intrigue the user's interest in watching a novel suggestion can be the determinant of the overall impression.

Table 1 (bottom) reports Spearman's ρ and Pearson's r correlation coefficients of the dialogue-level aspects with overall impression rating. *Efficiency* is the least correlating aspect for both scores. In our study, this aspect captures the system's ability to make relevant recommendations meeting the user's need within the first three exchanges. Unlike chatbots, which are meant to engage with a user for a long period, TDS dialogues should be concise [6].

We see in Figure 1b that more dialogues are rated inefficient than efficient (53.5% vs. 46.5%), suggesting that efficient suggestions of movies contribute to a dialogue being satisfactory. Our analysis, however, indicates that the opposite cannot be said for inefficient dialogues: some of them were rated satisfactory (64.44%). We note from the annotators' open comments that though a system took extra turns to make a relevant suggestion, as long as the user got a suggestion, they rate the system as satisfactory. This indicates that a system that fails to satisfy the user's need in the first three interactions is less likely to do so in further interactions.

To understand the significance of the investigated dialogue aspects to the overall impression, we train various regression models

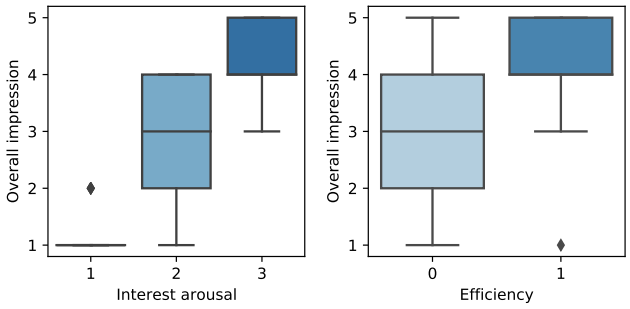

Figure 1: Box plots showing distribution of the (a) *interest arousal* and (b) *efficiency* aspects ratings against overall impression ratings.

considering different aspect combinations (both single and multiple aspects); see Table 2 for the results. At the turn-level, an approach that combines both aspects outperforms the best turn-level single aspect (*relevance*). As for the dialogue-level aspects, *interest arousal* exhibits the highest significance among all other aspects, taken individually. The combination of dialogue-level aspects clearly shows a stronger relationship to the overall rating model than individual aspects. Unsurprisingly, combining all aspects performs better than that of individual aspects or different levels.

Tables 1 and 2 show that dialogue-level aspects have a bigger influence on the overall impression than turn-level aspects. This suggests that turn-level aspects cannot be used solely to estimate the user's overall satisfaction effectively. This is attributed to cases where a system's response at a turn is sub-optimal, thus not representing the entire dialogue impression. The turn and dialogue aspects concern two evaluation dimensions: utility and user experience. *Relevance* and *task completion* measure the utility of a TDS, i.e., its ability to accomplish a task by making relevant suggestions. The user experience dimensions (*understanding*, *interest arousal*, *efficiency* and *interestingness*) focus on the user's interaction experience. Combining dialogue aspects from both dimensions has a strong relationship to the overall impression, unlike the individual aspects. In Table 2 the columns Utility and User experience show the two dimensions: combining both dimensions (the last row in each section in Table 2) leads to the best performance. The combination of turn and dialogue level aspects (D+T, third group) achieves the highest R^2. In summary, leveraging aspects from both dimensions (utility and user experience) is essential when designing a TDS that is meant to achieve a high overall impression.

4.3 Analysis of the justifications

We report on a manual inspection of the workers' open comments. We went through the comments and assigned them to evaluation aspects based on the worker's perspective. E.g., a comment that mentions "the system kept recommending the same movie" signals the existence of a novel aspect that concerns repeated recommendations in a dialogue. Table 3 lists the (dominant) novel categories discovered from the comments, together with a gloss and example. Several interesting aspects are observed by the annotators. For example, most annotators disliked the fact that the system expressed its opinion on a genre or movie. In cases where the system

is repetitive (in terms of language use or recommended items), the annotators' assessments were negatively impacted. Some annotators noted the positive impact of a dialogue being natural and human-like or that the system made a good recommendation after several failed suggestions (i.e., success on the last interaction). There were some examples where all annotators agreed that the suggestions were good, but the user did not react rationally.

5 DISCUSSION AND CONCLUSIONS

In this paper, we focus on providing a fine-grained understanding of what the overall user impression means in TDSs. We asked annotators to follow a dialogue and assess both at the turn and dialogue level on multiple aspects. While related work highlights the significance of these aspects [20, 22, 27], not much work has been done on the impact of these aspects on TDSs.

Providing relevant recommendations throughout a dialogue is crucial for user satisfaction, but it does not tell the whole story. Both from the annotations and open-ended comments, we find that engaging with users in the form of chit-chat can have two effects. If a user is already happy with a provided recommendation, more engagement can lead to further *interest arousal*, and hence more satisfaction; but if the system fails to meet the user's expectations, it can have a negative effect. This is in line with [22], who stress the importance of finding the right amount of chit-chat in a dialogue.

Our analysis of open-ended comments and justifications revealed new aspects that can affect users' satisfaction. In line with our quantitative analysis and related work [13, 17], many annotators mentioned the importance of user experience in the final turns or at least one successful interaction in the dialogue. Other aspects such as repeated utterances and recommendations negatively impacted the user experience. This indicates the need for jointly optimising turn- and dialogue-level metrics and for a fine-grained model of user satisfaction that incorporates multiple aspects.

One limitation of our work is that the annotators assess user satisfaction based on the user's utterances and reactions to the system's responses. While we observed a high level of agreement for most dialogues, we noticed a disagreement between annotators on some dialogues. We plan to collect a set of fine-grained annotation labels directly from users. Also, we plan to extend this study to learn to predict the overall impression of the users in a TDS.

ACKNOWLEDGMENTS

We thank our reviewers for valuable feedback. This research was supported by Huawei Finland and by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Azzah Al-Maskari and Mark Sanderson. 2010. A review of factors influencing user satisfaction in information retrieval. *Journal of the American Society for Information Science and Technology* 61, 5 (2010), 859–868.

[2] Wanling Cai and Li Chen. 2020. *Predicting User Intents and Satisfaction with Dialogue-Based Conversational Recommendations.* Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/3340631.3394856

[3] Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2020. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review* 54 (2020), 755 – 810.

[4] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating Coherence in Dialogue Systems using Entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 3806–3812. https://doi.org/10.18653/v1/N19-1381

[5] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication* 50, 8 (2008), 630–645. https://doi.org/10.1016/j.specom.2008.04.002 Evaluating new methods and models for advanced speech-based interactive systems.

[6] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and Challenges in Conversational Recommender Systems: A Survey. *AI Open* 2 (July 2021), 100–126.

[7] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019.* 1891–1895. https://doi.org/10.21437/Interspeech.2019-3079

[8] Seyyed Hadi Hashemi, Kyle Williams, Ahmed El Kholy, Imed Zitouni, and Paul A. Crook. 2018. Measuring User Satisfaction on Smart Speaker Intelligent Assistants Using Intent Sensitive Query Embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18).* Association for Computing Machinery, New York, NY, USA, 1183–1192. https://doi.org/10.1145/3269206.3271802

[9] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W. White. 2015. Understanding and Predicting Graded Search Satisfaction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (Shanghai, China) *(WSDM '15).* Association for Computing Machinery, New York, NY, USA, 57–66. https://doi.org/10.1145/2684822.2685319

[10] Diane Kelly. 2009. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends Inf. Retr.* 3, 1—2 (jan 2009), 1–224. https://doi.org/10.1561/1500000012

[11] Youngho Kim, Ahmed Hassan, Ryen W. White, and Imed Zitouni. 2014. Modeling Dwell Time to Predict Click-Level Satisfaction. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (New York, New York, USA) *(WSDM '14).* Association for Computing Machinery, New York, NY, USA, 193–202. https://doi.org/10.1145/2556195.2556220

[12] Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting User Satisfaction with Intelligent Assistants. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16).* Association for Computing Machinery, New York, NY, USA, 45–54. https://doi.org/10.1145/2911451.2911521

[13] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (Carrboro, North Carolina, USA) *(CHIIR '16).* Association for Computing Machinery, New York, NY, USA, 121–130. https://doi.org/10.1145/2854946.2854961

[14] Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons. *ArXiv* abs/1909.03087 (2019).

[15] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018).*

[16] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Austin, Texas, 2122–2132. https://doi.org/10.18653/v1/D16-1230

[17] Jiqun Liu and Fangyuan Han. 2020. Investigating Reference Dependence Effects on User Search Interaction and Satisfaction: A Behavioral Economics Perspective. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR.* ACM, 1141–1150.

[18] Shikib Mehri and Maxine Eskenazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Online, 681–707. https://doi.org/10.18653/v1/2020.acl-main.64

[19] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 8689–8696. https://doi.org/10.1609/aaai.v34i05.6394

[20] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 1702–1723. https://doi.org/10.18653/v1/N19-1170

[21] Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2020. Generating empathetic responses by looking ahead the user's sentiment. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7989–7993.

[22] Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding Chit-Chat to Enhance Task-Oriented Dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 1570–1583. https://doi.org/10.18653/v1/2021.naacl-main.124

[23] Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. *Simulating User Satisfaction for the Evaluation of Task-Oriented Dialogue Systems*. Association for Computing Machinery, New York, NY, USA, 2499–2506. https://doi.org/10.1145/3404835.3463241

[24] Anu Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, Shaohua Yang, and Anirudh Raju. 2018. On Evaluating and Comparing Open Domain Dialog Systems. *arXiv: Computation and Language* (2018).

[25] Marilyn A. Walker, Diane J. Litman, Candace A. Kamm, and Alicia Abella. 1997. PARADISE: A Framework for Evaluating Spoken Dialogue Agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics* (Madrid, Spain) *(ACL '98/EACL '98)*. Association for Computational Linguistics, USA, 271–280. https://doi.org/10.3115/976909.979652

[26] Shuo Zhang and Krisztian Balog. 2020. *Evaluating Conversational Recommender Systems via User Simulation*. Association for Computing Machinery, New York, NY, USA, 1512–1520. https://doi.org/10.1145/3394486.3403202

[27] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2204–2213. https://doi.org/10.18653/v1/P18-1205