UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Understanding and Predicting User Satisfaction with Conversational Recommender Systems

Siro, C.; Aliannejadi, M.; de Rijke, M.

DOI
10.1145/3624989

Publication date
2024

Document Version
Final published version

Published in
ACM Transactions on Information Systems

License
CC BY-ND

Link to publication

Citation for published version (APA):
Siro, C., Aliannejadi, M., & de Rijke, M. (2024). Understanding and Predicting User Satisfaction with Conversational Recommender Systems. *ACM Transactions on Information Systems*, *42*(2), Article 55. https://doi.org/10.1145/3624989

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:27 May 2024

Understanding and Predicting User Satisfaction with Conversational Recommender Systems

CLEMENCIA SIRO, MOHAMMAD ALIANNEJADI, and MAARTEN DE RIJKE, University of Amsterdam, The Netherlands

User satisfaction depicts the effectiveness of a system from the user's perspective. Understanding and predicting user satisfaction is vital for the design of user-oriented evaluation methods for **conversational recommender systems (CRSs)**. Current approaches rely on turn-level satisfaction ratings to predict a user's overall satisfaction with CRS. These methods assume that all users perceive satisfaction similarly, failing to capture the broader dialogue aspects that influence overall user satisfaction.

We investigate the effect of several dialogue aspects on user satisfaction when interacting with a CRS. To this end, we annotate dialogues based on six aspects (i.e., *relevance*, *interestingness*, *understanding*, *task-completion*, *interest-arousal*, and *efficiency*) at the turn and dialogue levels. We find that the concept of satisfaction varies per user. At the turn level, a system's ability to make relevant recommendations is a significant factor in satisfaction. We adopt these aspects as features for predicting response quality and user satisfaction. We achieve an F1-score of 0.80 in classifying dissatisfactory dialogues, and a Pearson's r of 0.73 for turn-level response quality estimation, demonstrating the effectiveness of the proposed dialogue aspects in predicting user satisfaction and being able to identify dialogues where the system is failing.

With this article, we release our annotated data.[1]

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

Additional Key Words and Phrases: Conversational recommender system, understanding user satisfaction, predicting user satisfaction, user experience

ACM Reference format:
Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2023. Understanding and Predicting User Satisfaction with Conversational Recommender Systems. *ACM Trans. Inf. Syst.* 42, 2, Article 55 (November 2023), 37 pages.
https://doi.org/10.1145/3624989

1 INTRODUCTION

Evaluation is a major concern when developing **information retrieval (IR)** systems, and it can be conducted based on measures of result relevance or user experience, such as user satisfaction,

---

[1] https://github.com/Clemenciah/Understanding-User-Satisfaction-Data

Authors' addresses: C. Siro, M. Aliannejadi, and M. de Rijke, University of Amsterdam, Amsterdam, The Netherlands; e-mails: {c.n.siro, m.aliannejadi, m.derijke}@uva.nl.

This work is licensed under a Creative Commons Attribution-NoDerivs International 4.0 License.

© 2023 Copyright held by the owner/author(s).

1046-8188/2023/11-ART55
https://doi.org/10.1145/3624989

which focuses on the user's perspective. While relevance metrics such as nDCG or average precision [34] have been commonly used, are re-usable and allow for system comparison, they often demonstrate poor correlation with the user's actual interaction experience [2, 63]. As a result, in recent years, there has been a growing interest in user-oriented evaluation approaches that rely on various user interaction signals, in contrast to system-oriented evaluation methodologies, i.e., the Cranfield paradigm [13, 14].

In traditional **recommender systems (RSs)**, user-oriented evaluation strategies often rely on implicit user feedback, such as user clicks and mouse scroll events to assess whether a user finds the recommended item appealing or not. However, such interaction signals are not available for **conversational recommender systems (CRSs)** whose main interaction with users is in natural language either by text or speech [26]. In CRSs, users interact with the system through natural language with utterances such as "I like the movie, I will watch it" expressing their preference in more detail [54]. This distinction in user interaction poses unique challenges in evaluating CRSs, both in terms of design and deployment, to ensure that these systems effectively cater to the user's needs.

User satisfaction. CRSs are recommender systems designed to provide recommendations that address the specific needs of users. As such, they fall under the category of task-oriented dialogue systems **task-oriented dialogue systems (TDSs)**. Standard automatic evaluation metrics such as BLEU [52], ROUGE [45], and METEOR [16] have shown poor correlation with human judgment [46], thus making them unsuitable for the evaluation of TDSs. In recent years, the research community has shown significant interest in developing new automatic evaluation metrics tailored to dialogue systems. These metrics not only exhibit stronger correlation with human judgment, but also consider various aspects of dialogues, such as relevance, interestingness, and understanding, without relying solely on word overlap [27, 32, 51, 64, 70]. While these metrics perform well during system design, their efficacy during system deployment is still a subject of ongoing investigation.

As a consequence, a significant number of TDSs rely on human evaluation to measure the system's effectiveness [29, 42]. An emerging approach for evaluating TDSs is to estimate a user's overall satisfaction with the system from explicit and implicit user interaction signals [29, 42]. While this approach is valuable and effective, it does not provide insights into the specific aspects or dimensions in which the CRS is performing well. Understanding the reasons behind a user's satisfaction or dissatisfaction is crucial for the CRS to learn from errors and optimize its performance in individual aspects, thereby avoiding complete dissatisfaction during an interaction session.

Understanding user satisfaction in a task-oriented setting. Understanding user satisfaction with CRSs is crucial, mainly for two reasons. Firstly, it allows system designers to understand different user perceptions regarding satisfaction, which in turn leads to better user personalization. Secondly, it helps prevent total dialogue failure by enabling the deployment of adaptive conversational approaches, such as failure recovery or topic switching. By conducting fine-grained evaluations of CRSs, the system can learn an individual user's interaction preferences, leading to a more successful fulfillment of the user's goal.

Various metrics, including engagement, relevance, and interestingness, have been investigated to understand fine-grained user satisfaction and their correlation with overall user satisfaction in different scenarios and applications[28, 59, 64]. While recent research has seen a surge in fine-grained evaluation for dialogue systems, most of these studies have focused on open-domain dialogue systems that are non-task-oriented [22, 27, 51]. On the other hand, conventionally, TDSs such as CRSs are evaluated on the basis of **task success** (**TS**) and overall user satisfaction. In CRSs, user satisfaction is modeled as an evaluation metric for measuring the ability of the system to achieve a pre-defined goal with high accuracy, that is to make the most relevant recommendations [55]. In contrast, for non-task-based dialogue systems (i.e., chat-bots), the evaluation focus is primarily on the user experience during interaction (i.e., how engaging or interesting the system is) [43].

Evaluating user satisfaction. Recent studies have examined user satisfaction in dialogue systems, particularly in the context of CRSs. These studies typically estimate user satisfaction by collecting overall turn-level satisfaction ratings from users during system interactions or by leveraging external assessors through platforms like **Amazon mechanical turk (MTurk)**.[2] In these evaluations, users[3] are typically asked to provide ratings for each dialogue turn by answering questions such as, *Are you/Is the user satisfied with the system response?* While overall turn-level satisfaction ratings provide a measure of user satisfaction, they may not capture the broader aspects that contribute to a user's satisfaction [60]. When humans are asked to evaluate a dialogue system, they often consider multiple aspects of the system [22]. Therefore, the satisfaction label aims to summarize the user's opinion into one single measure. Venkatesh et al. [64] argue that user satisfaction is subjective due to its reliance on the user's emotional and intellectual state. They also demonstrate that different dialogue systems exhibit varying performance when evaluated across different dialogue aspects, indicating the absence of a one-size-fits-all metric.

Previous studies have proposed metrics that offer a granular analysis of how various aspects influence user satisfaction in chat-bot systems [28, 64]. However, it is unclear how these aspects specifically influence user satisfaction in the context of TDSs [see, e.g., 41, 71]. With most aspect-based evaluations focusing on chat-bot systems [50, 51], only a few studies have so far investigated the influence of dialogue aspects for TDSs [37, 60]. Jin et al. [37] present a model that explores the relationship between different conversational characteristics (e.g., adaptability and understanding) and the user experience in a CRS. Their findings demonstrate how conversational constructs interact with recommendation constructs to influence the overall user experience of a CRS. However, they do not specifically examine how individual aspects impact a user's satisfaction with the CRS. In our previous work [60], we proposed several dialogue aspects that could influence a user's satisfaction with TDSs. We found that, in terms of turn-level aspects, *relevance* strongly influenced a user's overall satisfaction rating(Spearman's ρ of 0.5199). Additionally, we introduced a newly defined aspect, *interest arousal* which exhibited a high correlation with overall user satisfaction(Spearman's ρ of 0.7903). However, we did not establish a direct relationship between turn-level aspects and turn-level user satisfaction in our previous study.

Research questions. In this study, we seek to extend the study we carried out in [60]. Our aim is to understand a user's satisfaction with CRSs by focusing on the dialogue aspects of both the response and the entire dialogue. We intend to establish the relationship between individual dialogue aspects and overall user satisfaction to understand how they relate with **satisfactory (Sat)** and **dissatisfactory (DSat)** dialogues.

In addition, we aim to evaluate how effective the proposed aspects are in estimating a user's satisfaction at the turn and dialogue levels. To this aim, we carry out a crowdsourcing study with workers from MTurk on recommendation dialogue data, viz. the ReDial dataset [44]. The ReDial dataset provides a high-quality resource to investigate how several dialogue aspects affect a user's satisfaction during interaction with a CRS. We ask workers to annotate 600 dialogue turns and 200 dialogues on six dialogue aspects following our previous work [60]: *relevance*, *interestingness*, *understanding*, *task completion*, *interest arousal*, and *efficiency*. The dialogue aspects are grouped into utility and **user experience (UX)** dimensions of a **TDS**. Different from [60], we also ask workers to give their turn-level overall satisfaction rating and use it to establish a relationship between turn-level aspects and turn-level user satisfaction.

---

[2] https://www.mturk.com

[3] Here, *users* represent both actual users and external assessors.

Our aim is to answer the following research questions:

- **RQ1** How do the proposed dialogue aspects influence overall user satisfaction with a CRS?
- **RQ2** Can we estimate user satisfaction at each turn from turn-level aspects?
- **RQ3** How effective are the dialogue-level aspects in estimating user satisfaction compared to turn-level satisfaction ratings on CRSs?

Main findings. To address our research questions, we perform an in-depth analysis of the annotated turns and dialogues in order to understand how the proposed dialogue aspects influence a user's overall satisfaction. We note that for most annotators, at the turn level, the ability of a CRS to make relevant recommendations has a high influence on their turn-level satisfaction rating with a Spearman's ρ of 0.6104. In contrast, at the dialogue level, arousing a user's interest in watching a novel recommendation along with completing a task are the most influential determinants for overall satisfaction ratings from annotators with a Spearman's ρ of 0.6219 and 0.5987, respectively.

To evaluate the effectiveness of the proposed dialogue aspects, we experiment with several machine learning models on user satisfaction estimation and compare their performance using the annotated data. At the turn-level user satisfaction estimation task, we achieve a Spearman's ρ of 0.7337 between a random forest regressor model's prediction and the ground truth ratings. We achieved a correlation score of 0.7956 for predicting user satisfaction at the dialogue level. These results show the efficacy of the proposed dialogue aspects in estimating user satisfaction. Additionally, these results also demonstrate the significance of assessing the performance of a CRS at the aspect level; they can help system designers to identify on what dialogue quality a CRS is not performing as expected and optimize it.

Contributions. Our contributions in this article can be summarized as follows.

1. In our previous work [60], we conducted a study on 40 dialogues and 120 responses. In order to gain more insights, we extend that study with an extra 160 dialogues and 480 responses. In total, we conducted our investigations on 200 dialogues and 600 responses.
2. We ask annotators to assess dialogues on six dialogue aspects and overall user satisfaction. In addition, they provide judgments on turn-level satisfaction. User satisfaction ratings at the turn level allow us to establish the relationship between turn-level aspects and not only overall dialogue satisfaction but also turn-level satisfaction, which we did not experiment with in our previous work.
3. We carry out an in-depth feature analysis on individual dialogue aspects and at the class level (i.e., Sat and DSat classes) so as to understand which dialogue aspects correlate highly with each of the classes.
4. Leveraging the annotated data, we experiment with several classical machine learning models and compare their performance in estimating user satisfaction at the turn and dialogue levels.
5. Our findings indicate that predictive models perform better at estimating user satisfaction based on the proposed dialogue aspects than based on turn-level satisfaction ratings.

To the best of our knowledge, our work is the first attempt to establish a relationship between the proposed dialogue aspects and user satisfaction at both the turn and dialogue levels and to evaluate their effectiveness in estimating user satisfaction with CRSs.

Organization of the paper. The rest of this article is organized as follows. In Section 2, we discuss related work. We describe the dialogue aspects investigated in this study in Section 3. In Section 4, we detail our annotation process and instructions given to the annotators. In Section 5, we analyse the annotated data to answer **RQ1**. Section 6 discusses our problem formulation and predictive models used to estimate turn- and dialogue-level user satisfaction, while Section 7 presents the results to our experiments and answers **RQ2** and **RQ3**. We discuss our results and limitations of this study in Section 8 and make our conclusions, implications, and future work in Section 9.

2 RELATED WORK

Our work is relevant to three main research areas: (i) conversational recommender systems, (ii) evaluation of dialogue systems, and (iii) user satisfaction in task-oriented dialogue systems because we provide a means to comprehend and measure overall user satisfaction with conversational recommender systems.

2.1 Conversational Recommender Systems

Research on recommendation via conversational interactions with information retrieval systems is increasingly receiving attention from both industry and academia. With multi-turn interactions, a CRS enables users to find their most relevant recommendations [24]. The CRS can interactively elicit users' current preferences from their feedback and build a more complete user model to make better recommendations. Conventional recommender systems, on the other hand, only support a single interaction mode, i.e., displaying a set of suggestions depending on users' historical activities [56]. Some older CRSs took advantage of user interface elements, such as critiquing-based systems [12], where users give input on suggestions by picking from a list of pre-defined criticisms [33]. Nonetheless, recent developments in natural language technology have led to more interest in developing a CRS based on conversational user interfaces (CUIs), where users can converse with the recommender system [38]. Several other approaches have been explored to enhance the effectiveness of recommendations, such as knowledge graph integration [74], prompt learning [66], and topic guidance [75].

The evaluation of CRSs is based on offline experiments that try to simulate a user's behavior relying on their past interaction data. One line of research evaluates the performance of a CRS based on how well it accomplishes the user's goal by making relevant recommendations using metrics such as TS and recommendation accuracy. Another line of work focuses on dialogue generation aspects, assessing the quality of the responses using word-overlap metrics such as the ROUGE score [17]. However, as argued by Deriu et al. [17], such individual measures do not reflect the overall quality of the system. Thus, current evaluation metrics that rely heavily on the system's utility do not provide us with information about the evaluation findings in practical settings. On the other hand, research shows that empirical studies conducted using user-centric approaches can accurately assess the system's performance in actual scenarios [4]. Ideally, a system should be assessed separately on each specific dialogue-level aspect to capture its performance on individual aspects [60].

So far, little work has been done to establish the relationship between dialogue aspects and overall response and system quality [60].

2.2 User Satisfaction

Kelly [39] defines *user satisfaction* as the fulfillment of a user's specified desire or goal. User satisfaction has gained popularity as an evaluation metric of IR systems based on implicit signals [29, 35, 40–42]. In IR, user satisfaction is usually estimated based on the user's interaction experience and goal fulfillment [39]. Factors such as system effectiveness, user effort, characteristics, and expectations influence a user's satisfaction rating in IR systems [1]. Dialogue systems are often evaluated on their overall satisfaction [17], where users give their satisfaction rating at the turn and dialogue levels [10, 62]. Though subjective, user satisfaction provides valuable insights into users' perceptions, preferences, and overall evaluation of a system's performance. Additionally, it is a widely used and accepted metric in user experience research [see, e.g., 7, 29, 42].

However, for task-oriented dialogue systems such as CRS, which should optimize toward recommendation and user experience, overall satisfaction does not capture the broad and diverse aspects influencing a user's satisfaction [60]. Thus, in this research, we seek to investigate this concept.

2.3 Fine-Grained Evaluation

Due to poor correlation between automatic metrics such as BLEU and human judgment, accurate evaluation of dialogue systems rely on human evaluation [46]. Non-task-oriented dialogue systems are evaluated on specific aspects such as relevance and engagingness [51, 64]. However, task-oriented dialogue systems are often limited to estimating the user's overall satisfaction [41, 62]. Recent research suggests that user satisfaction is multifaceted and subjective, thus, should not be reduced to a single label [64].

Several recent studies have proposed to evaluate dialogue systems at an aspect level. For example, one would measure the performance of a system in making *relevant* or *understandable* responses, instead of the overall quality of the response. PARADISE [65] is one of the first popular evaluation frameworks that decoupled a dialogue system's task requirements from its behavior. With predictive factors such as TS, dialogue efficiency, and dialogue quality, a system's effectiveness can be measured without having to collect user satisfaction ratings. Walker et al. [65] propose a framework for evaluating dialogues in a multi-faceted manner. They measure several dialogue aspects and combine them to estimate user satisfaction [65]. Mehri and Eskenazi [51] develop an automatic metric for evaluating dialogue systems at a fine-grained level, including interestingness, engagingness, diversity, understanding, specificity, and inquisitiveness. In their study, Venkatesh et al. [64] investigate the performance of multiple dialogue systems involved in the Alexa competition on several dialogue aspects and show that different systems perform well in specific dialogue aspects. Moreover, they show that no single measurement can be used to evaluate the overall performance of a system accurately. Several other studies have been carried out on human evaluation of multiple dialogue aspects [see, e.g., 19, 50, 59, 69, 72].

2.4 Predicting User Satisfaction

Predicting user satisfaction is critical in capturing whether a user's goal has been fulfilled or not. In web search, user satisfaction is viewed as a subjective measure of a user's experience during search [39]. Different from traditional IR relevance measures, such as precision and recall, user satisfaction takes into account both TS and user interaction experience [29, 30, 41]. For search systems, rich user interaction signals such as clicks, dwell time, and mouse scroll events are used to predict a user's satisfaction [35, 40]. Such interaction signals cannot be collected from dialogue-based systems whose main interaction is through natural language, either in text or spoken. Research on spoken dialogue systems, such as intelligent assistants, has addressed this challenge by suggesting the use of features such as spoken implicit features, intent-sensitive query embeddings, and touch-related features, showing their effectiveness in predicting user satisfaction [29, 41]. Several other features have been suggested in line with text-based dialogue systems including implicit dialogue features, user intent, utterance length, and user-system actions, and proven to be effective [10, 62]. Bodigutla et al. [6] demonstrate the effectiveness of traditional machine learning models in predicting user satisfaction. Using predicted turn-level ratings with implicit dialogue features, models such as gradient boosting classifiers demonstrate competitive performance [6]. In task-oriented systems, several publications predict user satisfaction from turn-level overall quality user judgment ratings [62], user intents [10, 53], and implicit features such as utterance length and sentiment analysis.

Despite the success of related work in predicting user satisfaction with task-oriented systems, there has been less focus on trying to understand which dialogue aspects have an effect on a user's satisfaction with these systems. Recent work by Siro et al. [60], tries to understand the relationship between several dialogue aspects and overall user satisfaction in a TDS, especially at the dialogue level. However, compared to related work, our work in this article is different in a number of ways: (i) unlike Siro et al. [60], who focus on dialogue-level user satisfaction, in this work, we establish the relationship between turn- and dialogue-level user satisfaction; (ii) we show the effectiveness of the dialogue aspects in estimating user satisfaction by experimenting with several classical machine learning models; and (iii) we increase the data sample size by re-annotating data from [60] with one more aspect (turn-level satisfaction) and annotating an additional 160 dialogues and 480 turns. Thus, in total, we have 200 dialogues and 600 turns annotated.

3 ASPECTS INFLUENCING USER SATISFACTION

In this section, we discuss the dialogue aspects we use in our crowdsourcing study. We map the qualities from prior work [37, 51, 60, 64], highlighting their definitions in different settings and defining them in our work. These qualities are derived from two TDS dimensions defined in our previous work [60]; the *utility* and *user experience* dimensions.

3.1 Utility

The *utility* dimension focuses on the objective nature of a CRS, that is to make relevant recommendations and accomplish a user's goal. In this dimension, we investigate two qualities, namely, *relevance* measured at the turn level and *task completion* measured at the dialogue level.

3.1.1 Relevance. *Relevance* is a central concept in the field of IR and plays an important role in the evaluation of conversational systems [57]. In essence, *relevance* is logically defined in the relationship between the information at hand and the user's information need [15]. In the field of conversational agents, it is used as a criterion for assessing the effectiveness of a dialogue system to potentially convey a piece of information that meets the user's needs. Ideally, relevance judgment labels should be collected from actual users to reflect their opinions (i.e., whether the suggested responses meet their information needs or not). However, it is hard to collect relevance judgments from actual users during an interaction, especially for conversational systems. This approach can be intrusive and may negatively impact the user's overall interaction experience with the system. In recent work, crowdsourcing has emerged as a reliable platform for collecting relevance labels for web search and conversational systems [3].

In our work, we employ crowdsourcing to collect relevance labels for dialogue responses. To assess the relevance of a response, we instruct annotators to rely solely on the user's explicit feedback provided in the current user's utterance. For instance, expressions such as "I do not think that is a horror movie," "I like it," "I have seen that one," "Could you recommend more like that one?" following a system's recommendation indicate whether the items suggested are relevant to the user's needs. In contrast to web search, where assessors judge the relevance of a query-document pair, relevance assessment for dialogue systems focuses on the appropriateness of the response [50]. In this study, we primarily evaluate the relevance of recommended movies rather than the appropriateness of the dialogue response itself. Therefore, we first ask annotators to determine if a movie is recommended in the response or not. If a response does not include a movie recommendation, we skip the relevance assessment. However, if a movie is recommended, we ask the annotators to determine a three-level relevance label (see Section 4 for more details). We adopted this definition because of the nature of our study, which is task-oriented, where our focus is on the utility of the system. Hence, relevance indicates how well the recommendations provided by the system align with the user's needs and preferences in the given conversational context. Assessing relevance at the turn level allows us to evaluate the immediate impact of a recommendation on the ongoing conversation and its ability to address the user's current needs.

3.1.2 *Task Completion.*

Task completion is a crucial aspect of task-oriented **conversational recommendation systems** (**CRS**), as they are designed with a predefined goal in mind. Traditionally, the main evaluation metric for task-oriented systems has been TS, which measures the system's ability to fulfill a user's goal [67]. However, in the case of interactive CRS, TS alone may not capture the overall satisfaction of the user with the dialogue. This is due to the interactive nature of the system and the fact that TS can vary depending on individual users and task complexity [48]. Simply relying on system logs to infer user search success is inadequate because task complexity and individual user needs cannot be accurately depicted in the logs.

To address this limitation, recent research has proposed using additional interaction cues, such as self-reported user TS or expert-annotator labels on TS [17]. In our work, we investigate how the system's ability to accomplish a user's goal influences the overall impression of the dialogue for the user with a CRS. We assess the system's capability to understand the user's intent and provide recommendations that satisfy their needs.

To measure the quality of task completion, we rely on the user's acknowledgment within the conversation. Utterances such as "I like it, I will watch it tonight" and "I think I will add that to my watching list" serve as signals indicating the successful accomplishment of the task from the user's perspective. By considering these explicit expressions of satisfaction or intent to engage with the recommended items, we can assess how effectively the CRS understands and addresses the user's needs.

By incorporating task completion as an evaluation metric, we aim to capture the system's ability to achieve the user's desired outcome and provide recommendations that align with their preferences. This approach allows us to evaluate the CRS beyond the traditional notion of TS and consider the overall dialogue satisfaction from the user's point of view.

3.2 User Experience

In the *UX* dimension, we assess how different dialogue aspects of a CRS during interaction could affect a user's satisfaction. The ideal requirement would be a system that interacts in a natural way with the user, making the interaction experience pleasing. Thus, inspired by related work [51, 60, 64, 72], we investigate the *interestingness*, *understanding*, *interest arousal*, and *efficiency* aspects.

3.2.1 *Interestingness.*

Due to recent advances in machine learning and natural language understanding, conversational agents such as Alexa and Siri have become increasingly common. While these agents are classified as task-oriented, there is an emerging interest in building dialogue systems that can socially engage with users while accomplishing a task [60, 61]. This quality has been used as a metric for evaluating non-goal-oriented dialogue systems in recent work [51, 64, 72].

Several proxies have been suggested for measuring interestingness such as the number of dialogue turns and total duration of a conversation [36, 64]. Though useful, these proxies assume the dialogue is non-goal-oriented. For goal-oriented systems, a dialogue is often supposed to be as short as possible, so that the user's needs can be met quickly. Therefore, conversation length is not an accurate proxy for measuring interestingness in task-oriented systems. In our work, *interestingness* is the ability of the system to chit-chat while making relevant recommendations, that is, a system making a recommendation in a natural manner as found in casual human conversations. It reflects the system's ability to suggest items that pique the user's curiosity or meet their personal interests in a natural manner, thus enhancing their overall conversational experience. By annotating interestingness at the turn level, we aim to assess the immediate impact of a recommendation on the user's level of interest and engagement.

3.2.2 Understanding. The aspect of "understanding" has been investigated at both the system response and dialogue level. A system's response is said to be understandable if it makes sense in the provided context history [51]. For instance, a system is not supposed to make an utterance about racing car movies when the context is on religion (such a response will be rated as not understandable).

At the dialogue level, a system is said to be understanding if it is able to track the user's preference and intent along the whole dialogue [60]. An understanding system is expected to conform its dialogue style to the user's preference in order to make sensible utterances. We show that in order for a dialogue system to meet a user's needs, it should be able to understand the user's preference and intent of interaction, thus, this quality is crucial in a CRS's ability to accomplish a user's task.

3.2.3 Interest Arousal. We introduced interest arousal in our previous work [60], as an aspect highly correlating with overall user impression at the dialogue level. The ability of a task-oriented dialogue system to arouse a user's interest is significant enough to determine satisfactory dialogues [60]. This quality can be seen as a merge of two qualities: *novelty* and *explainability*. To measure the two together, we define interest arousal as "the ability of the system to suggest novel items to the user and give a brief explanation in the form of synopsis or main actors in order to attract the user's interest to accept the item."

We rely mostly on the user's immediate utterance to capture this quality. User utterances such as, "I do not know that movie" or "Who's the main actor?" indicate that the suggested movie is not known by the user and the CRS's next action should be to give a brief explanation. Note that we do not measure this quality at the response level because annotators require at least two turns to determine user interest arousal, as it is measured after a novel suggestion has been made. In this work, we are interested in quantifying the relationship between interest arousal and user satisfaction.

3.2.4 Efficiency. Task-oriented systems are expected to be efficient, i.e., accomplish a specified task within a minimal number of turns of interactions. In web search, a system's efficiency is measured by considering how many comparisons a user has to make before getting the needed results (number of documents examined before getting the relevant one).

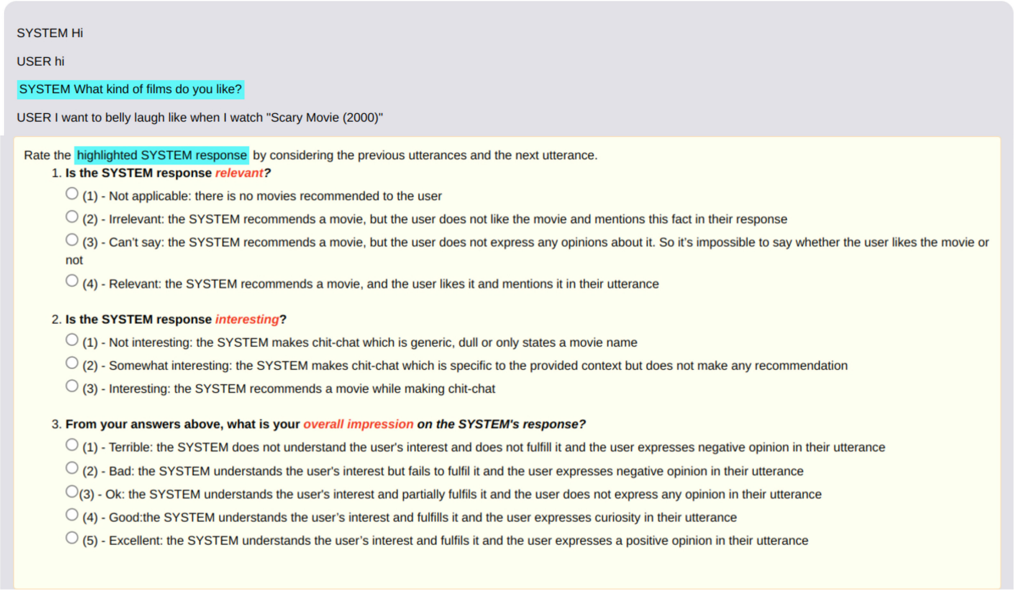
Various interaction signals are used to measure this aspect including conversation length, conversation duration for spoken dialogue systems, and search session length in web search systems. Since ReDial is a text-based dataset, we use conversation length to measure a system's efficiency, that is, the ability of the system to make suggestions that meet the user's needs within minimal turns. From our analysis, we note that in most conversations, a user acknowledges a recommendation within the first three turns, and thus, we conform to our previously proposed definition [60].

4 DATA ANNOTATION

To establish how the dialogue aspects in Section 3 exert a user's overall satisfaction, we create an additional annotation layer for the ReDial [44] dataset. We set up an annotation experiment on **Amazon mechanical turk (MTurk)** using the so-called master workers to assess:

(1) Three randomly selected responses from each dialogue on two aspects, namely, *relevance* and *interestingness*;
(2) The quality of the system at the dialogue level on the following aspects: *understanding*, *task completion*, *interest arousal*, and *efficiency*; and
(3) The overall satisfaction of the system response and the entire dialogue.

The complete instructions and definitions given to the assessors are provided in Tables 9, 10, and 11 (see the appendix). We display all three turns on a single page and instruct the annotators to answer questions for each turn, as shown in Figure 1. After completing the turn-level annotation,



SYSTEM Hi

USER hi

SYSTEM What kind of films do you like?

USER I want to belly laugh like when I watch "Scary Movie (2000)"

Rate the highlighted SYSTEM response by considering the previous utterances and the next utterance.

1. **Is the SYSTEM response *relevant*?**
   - ○ (1) - Not applicable: there is no movies recommended to the user
   - ○ (2) - Irrelevant: the SYSTEM recommends a movie, but the user does not like the movie and mentions this fact in their response
   - ○ (3) - Can't say: the SYSTEM recommends a movie, but the user does not express any opinions about it. So it's impossible to say whether the user likes the movie or not
   - ○ (4) - Relevant: the SYSTEM recommends a movie, and the user likes it and mentions it in their utterance
2. **Is the SYSTEM response *interesting*?**
   - ○ (1) - Not interesting: the SYSTEM makes chit-chat which is generic, dull or only states a movie name
   - ○ (2) - Somewhat interesting: the SYSTEM makes chit-chat which is specific to the provided context but does not make any recommendation
   - ○ (3) - Interesting: the SYSTEM recommends a movie while making chit-chat
3. **From your answers above, what is your *overall impression on the SYSTEM's response?***
   - ○ (1) - Terrible: the SYSTEM does not understand the user's interest and does not fulfill it and the user expresses negative opinion in their utterance
   - ○ (2) - Bad: the SYSTEM understands the user's interest but fails to fulfil it and the user expresses negative opinion in their utterance
   - ○(3) - Ok: the SYSTEM understands the user's interest and partially fulfils it and the user does not express any opinion in their utterance
   - ○ (4) - Good:the SYSTEM understands the user's interest and fulfills it and the user expresses curiosity in their utterance
   - ○ (5) - Excellent: the SYSTEM understands the user's interest and fulfils it and the user expresses a positive opinion in their utterance

Fig. 1. Turn-level annotation interface. A turn comprising of two user and system utterances with three follow up questions regarding the highlighted system utterance.

the same annotators are taken to a new page where they provide dialogue-level annotations on the same dialogue (see Figure 8 in the appendix). We do not allow the annotators to 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-level ratings correlate with their dialogue-level ratings and the overall satisfaction ratings.

4.1 Recommendation Dialogue Dataset

The ReDial dataset [44] is a conversational movie recommendation dataset. It consists of $11,348$ dialogues and the dataset are collected following the Wizard of Oz approach, i.e., one person acts as the movie seeker, while the other is the recommender. The dialogues are both system and user-initiated. The movie seeker should explain their movie preferences based on the genre, actor, and movie title and ask for suggestions. The recommender's role is to understand the seeker's movie taste and intent and make the right suggestions to the user. Due to this back-and-forth process of eliciting a user's preference, which mostly involves chit-chat, this dataset is categorized as both chit-chat and goal-oriented, thus allowing us to investigate dialogue aspects from both the utility and UX dimensions of a CRS.

4.2 Turn-Level Annotation

Unlike previous work [50, 51, 62], the annotators in our study have access to the user's current utterance. We treat the response quality annotation as a turn-level task. Considering the interactive nature of a CRS, a turn is defined as a single exchange between the user and the system [62]. Unlike previous work, a turn, in this case, consists of two exchanges between the user and the system. Therefore, we define a turn in this work as

$$T_i = S_{i-1}U_{i-1}, S_iU_i,$$

where U is the user utterance, S the system utterance and i is the current response position. In a recent study [62], turn-level annotation is conducted with workers having access to all previous system and user utterances up to the current system utterance as context and their role is to assess if the user would be satisfied with the current system response given the context without viewing the user utterance at position i. This approach requires annotators to understand the user's intent during the interaction and make judgments based on previous interactions. We argue that a user has a dynamic preference and intent during dialogue interactions and this can change from turn to turn, thus affecting their overall satisfaction of the system. In order to remedy this, we ask annotators to exclusively rely on the user's current utterance while making judgments on the dialogue aspects. That is, for each system response S_i to be annotated, the annotator has access to the previous user (U_{i-1}) and system (S_{i-1}) utterances as context and the current user utterance U_i from which they should make their judgment. In this way, we aim to limit annotators' bias, in that instead of annotators making judgments influenced solely by their own opinions, they reflect the opinions of the actual user as closely as possible.

Following Mehri and Eskenazi [51], we hand-selected three system responses from each conversation for turn-level annotation. In order to ensure three responses cover most of the dialogue, we only select dialogues with at most fifteen turns. We limit the context window to two such that each annotated response (S_i) has two previous utterances from the system (S_{i-1}) and the user (U_{i-1}) as context plus the current user utterance (U_i). This way, we ensure that an annotator does not have to keep track of a long conversation context when annotating a single response and each response has a reasonably long context during annotation.

For each response, we ask the annotators to assess them on relevance and interestingness and based on their ratings for the two aspects give their turn-level overall impression (satisfaction) rating as shown in Figure 1. The questions the annotators answered in this subtask are:

- *Is the system response relevant?*
- *Is the system response interesting?*
- *Based on your ratings above, what is your overall impression of the system response?*

As our annotators are not actual system users, we ask them to base their judgments solely on the next user's utterance to make the label judgment. For example, if the user states, "I do not like that movie," an annotator should be able to judge the system's response and recommendation as irrelevant since the suggested movie does not meet the user's needs. For "I have seen that and like it" the response should be rated as relevant. For the overall impression rating, we ask the annotators to base their judgment on the relevance and interestingness aspects. Each aspect comes with three options, namely, *No*, *Somewhat*, and *Yes*. For *relevance*, we also provide a *Not applicable* option when a system response does not contain a movie suggestion (e.g., if the system chit-chats or tries to elicit a user's preference). Due to limited annotation resources, we chose to focus on *relevance* and *interestingness* as the primary aspects for turn-level annotation, as they provide a strong foundation for evaluating the quality of recommendations in CRSs.

4.3 Dialogue-Level Annotation

At the dialogue level, we ask the annotators to assess the quality of the entire dialogue based on four aspects: understanding, task completion, efficiency, and interest arousal. We instructed the annotators to answer the following questions:

- *Is the system understanding the user's request?*
- *Did the system manage to complete the task?*
- *Is the system efficient?*
- *Does the system arouse the user's interest?*

Understanding and task completion are rated on a scale of 1–3 with the options of *No*, *Somewhat*, and *Yes.* Interest arousal is judged on a 4-point scale with a *Not Applicable* option for when no novel movie is recommended to the user or a novel movie is recommended but the user does not follow up about it. Lastly, efficiency is assessed on a binary scale [20, 42] where the system has either made a recommendation meeting a user's request within the first three turns or not. Following [51, 60], we also ask annotators to rate the entire dialogue on *overall impression* using a 5-point Likert scale based on their turn and dialogue levels aspects' ratings. Finally, we ask the workers to justify their rating on *overall impression* in a few words. We use the justifications to contextualize the given ratings and analyze and discover additional aspects that affect the quality of dialogue, as shown in Table 4.

4.4 Quality Control and Filtering

Here, we describe the demographics of our participants, followed by more details on the collected data and the measures we took to ensure the high quality of the data.

4.4.1 Participants. A total of 70 unique workers participate in the annotation. 56% male and 44% female, their age ranges from 18–40, with the majority aged between 24–35. A large number of the workers report not having experience with dialogue systems—78% have no experience vs. 22% who do have experience. To ensure quality annotations, we filter workers based on their MTurk approval rate. We recruit workers located in the United States to ensure they are all English-proficient, with an approval rate of 95% for more than 1000 hits.

4.4.2 Data. The number of turns in each dialogue used in this study ranges between 12 and 13. From the analysis we carried out on the dataset, we note that most of the long dialogues with more than 20 turns tend to deviate from the movie recommendation subject into other subjects such as politics. Each dialogue is initially annotated with at least three annotators. We always use an odd number of workers to allow for majority voting. If we lack a single agreed-upon label, an additional assessment is made with two more workers (mostly for the overall impression aspect). For the rest of the dialogue aspects, we use the labels as they are from the annotation scale to cater for the subjectivity of users in annotating the aspects. It is worth noting that we collected a set of additional annotation labels for a subset of 40 dialogues.

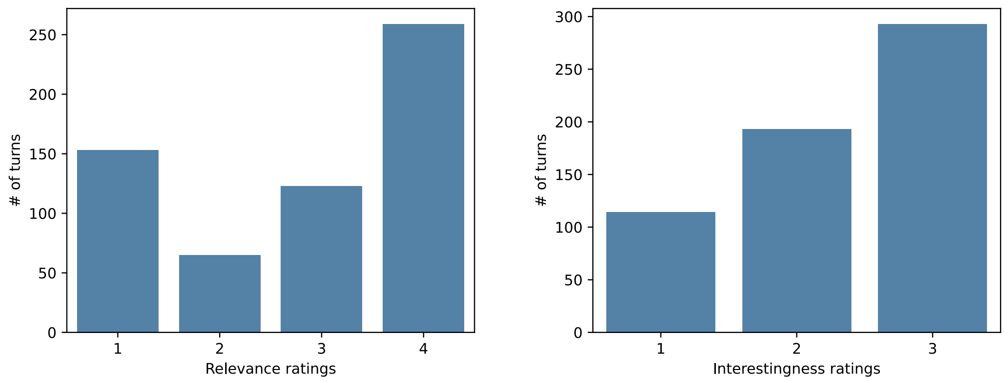
To get to a single label for each dialogue, we treat as outliers all labels different by more than 1.5 from the mean label. In case, we do not achieve a single majority label after the additional annotation, the authors re-annotate the dialogues themselves and come to an agreement for a single label.

5 DIALOGUE DATASET ANALYSIS

Using the annotated data, we first investigate **RQ1**: *How do the proposed dialogue aspects influence user satisfaction with a CRS?* To answer this question, we conducted several analyses to study the relationship between overall user satisfaction and both turn- and dialogue-level aspects. In addition, we identify essential aspects for the Sat and DSat classes.

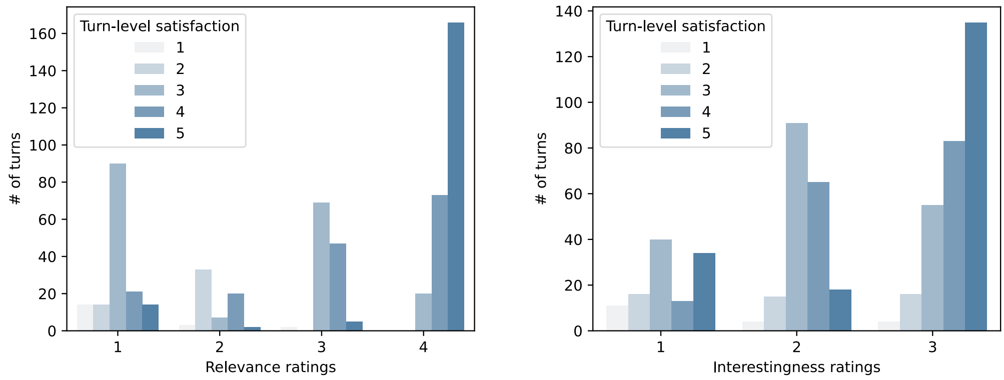
5.1 Turn-Level Analysis

At each turn, the aspects *relevance*, *interestingness*, and *overall turn quality* are rated. We show the distribution of the ratings for these aspects in Figures 2(a), 2(b), and 3 for *relevance*, *interestingness* and turn-level satisfaction, respectively. Note that the distributions in Figure 2 are computed over the three annotated turns in each dialogue. We can see that around 25% of the turns were annotated as not containing any movie recommendation ($R = 1$), while over 40% are annotated as very relevant. This result is not surprising because of the nature of the ReDial dataset, where a



(a) (b)

Fig. 2. Marginal distribution of (a) *relevance-R* annotations and (b) *interestingness-I* annotations. The values 1–3 mean not relevant/interesting, somewhat relevant/interesting, and very relevant/interesting, respectively, and with 1 for relevance meaning no movie is recommended.



(a) (b)

Fig. 3. Distribution of (a) relevance ratings, (b) interestingness ratings against turn level satisfaction, showing how assessors rated each response based on individual dialogue aspect.

recommender system needs to elicit a user's preference before making a suggestion, thus having multiple chit-chat turns. Meanwhile, turning to Figure 3, we observe that turns rated as very relevant and interesting at the same time overall led to a satisfactory turn, showing that CRS though goal-oriented should not only focus on making relevant recommendations but also in a natural and interesting manner.

Figure 4 shows Pearson's r between turn-level user satisfaction and (i) relevance (annotators assess if the recommended movie meets the user's preference), (ii) interestingness of system's response. Also, we report the correlation between relevance and interestingness in the figure. We note that the relevance and interestingness aspects have a moderate positive correlation with each other (~0.4). However, we see that relevance exhibits a higher correlation with overall turn impression than interestingness. Our analysis indicates that when a turn is rated as relevant, the turn's overall impression is more likely to be satisfactory (96% of the relevant turns).[4] On the other hand,

---

[4] We use "overall impression" and "overall user satisfaction" interchangeably; both refer to overall user satisfaction.

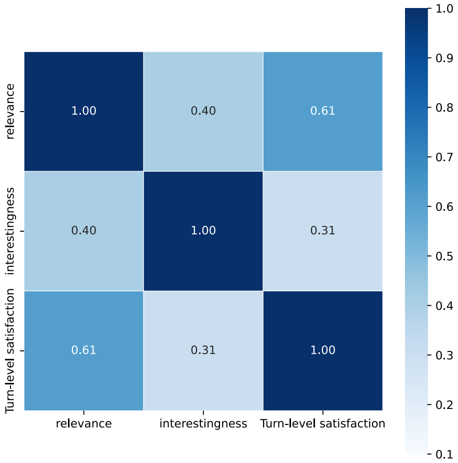


Fig. 4. Correlation of turn-level aspects to each other and to turn-level user satisfaction.

Table 1. Correlation of Dialogue-Level Overall Impression with Turn-Level and Dialogue-Level Aspects' Ratings

Level	Aspect	Spearman's ρ	Pearson's r
Turn	Relevance	0.3756	0.3935
	Interestingness	*0.1710*	0.2061
	Turn-level satisfaction (TSat)	**0.5397**	**0.5774**
Dialogue	Understanding	0.4929	0.5940
	Task completion	0.5987	**0.6429**
	Interest arousal	**0.6219**	0.6038
	Efficiency	0.3653	0.4004

All correlations in this table are statistically significant ($p < 0.01$).

the same does not hold for turns rated as irrelevant (43% of the irrelevant 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 such as response interestingness.

In summary, we note that at the turn level, the relevance and interestingness aspects are important in understanding a user's satisfaction. Specifically, we can rely on the relevance aspect to identify Sat responses, while interestingness can be used to identify DSat responses. Characterizing the relationship between these two classes could be useful in the automatic estimation of response quality.

5.2 Dialogue-Level Analysis

Table 1 reports Spearman's ρ and Pearson's r correlation coefficients of all six quality aspects, including *turn-level satisfaction* (TSat), with the overall dialogue satisfaction rating. Since three turns were annotated for each dialogue, we report the average results over all three turns for the three aspects. Note that both *relevance* and *turn-level satisfaction* have a moderate correlation (second row) with the overall dialogue satisfaction ratings. Compared to *interestingness*, *relevance* has a higher correlation, confirming our previous findings [60].

Notice that the *turn-level satisfaction* rating exhibits a high correlation with dialogue-level user satisfaction. This indicates that one can use a single overall turn-level quality metric to estimate a user's overall dialogue satisfaction, which has been used in previous studies [62]. We also do a

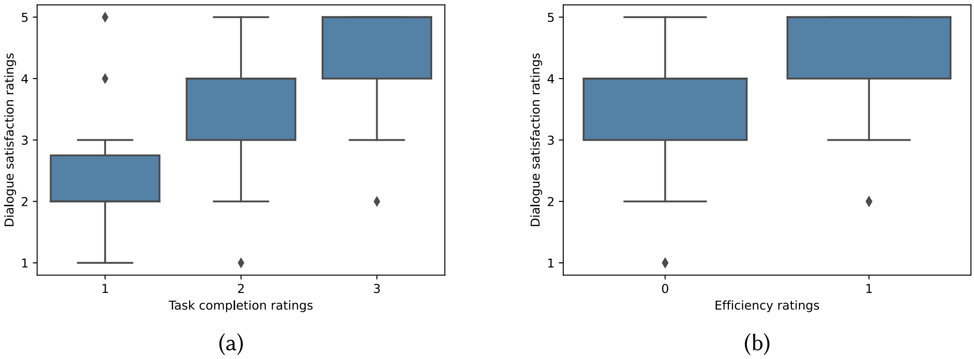


Fig. 5. Box plots showing distribution of the (a) *task completion* and (b) *efficiency* aspects ratings against overall impression ratings.

correlation analysis on each turn separately and note that both *relevance* and *turn-level satisfaction* achieve a high correlation in their third and last interaction turn compared to the other two previous turns. This shows that a system's success in making a successful suggestion[5] in the final turn has more weight on the overall impression than the preceding turns. This conforms to the findings of [42, 47, 60], showing that the latest interactions with a system have more influence on the overall satisfaction of users.

At the dialogue level, interest arousal achieves a high Spearman's ρ coefficient while task completion achieves a high Pearson's r coefficient, as shown in Table 1 (third row). *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 and efficient [24].

In Figure 5, we plot the distribution of the ratings for the dialogue-level aspects against the overall impression. We see a clear dependency of the *overall impression* on the *task completion* aspect; out of the dialogues classified as satisfactory, 68% were rated high in terms of task completion (see Figure 5(a)). We also notice that most dialogues rated low (= 1) in terms of task completion are unsatisfactory overall with a few outliers. Thus, we conclude that the ability of a CRS to complete a user's specified task can be the determinant of the overall impression.

We see in Figure 5(b) that more dialogues are rated efficient than inefficient (72.5% vs. 27.5%). We note that an efficient system, making suggestions meeting a user's need within three turns, leads to a satisfactory dialogue. Our analysis, however, indicates that the opposite cannot be said for inefficient dialogues: most of them were rated satisfactory (61.5%). We note from the annotators' open comments that even 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 considering different aspect combinations (both single and multiple aspects) and report their R^2; see Tables 2 and 3 for the results. R^2 represents the coefficient of determination for the regression model, which indicates the proportion of the variance in turn and dialogue level satisfaction that is explained by the independent and combined aspects [11]. At the

---

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

Table 2. Determinant Coefficients Computed with Regression Showing the Effect Size of Turn-Level Aspects to Turn-Level Satisfaction

Aspect	Utility	User experience	R^2
Relevance (R)	+		0.377
Interestingness (I)		+	*0.092*
R + I	+	+	**0.382**

Table 3. Determinant Coefficients Computed with Regression Showing the Effect Size of both Turn and Dialogue Levels Aspects to Overall Dialogue Satisfaction

	Aspect	Utility	User experience	R^2
Turn (T)	Relevance (R)	+		0.186
	Interestingness (I)		+	*0.036*
	Turn-level satisfaction (TSat)		+	0.290
	R + I + TSat	+	+	**0.310**
Dialogue (D)	Understanding (U)		+	0.353
	Task completion (TC)	+		0.413
	Interest arousal (IA)		+	0.365
	Efficiency (E)		+	0.160
	IA + TC + U + E	+	+	**0.559**
D + T	R + TC	+		0.452
	IA + U + I + E + TSat		+	0.572
	IA + TC + U + I + E + R + TSat	+	+	**0.607**

All results except the italicized results are significantly significant to ($p < 0.05$).

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 individual aspects or different levels.

Tables 1 and 3 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. The combination of dialogue aspects from both dimensions has a strong relationship with the *overall impression*, unlike the individual aspects. In Table 3, the columns Utility and User experience show the two dimensions: combining both dimensions (the last row in each section in Table 3) 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.

Analyzing annotators' open-comments. To identify additional dialogue aspects that influence a user's satisfaction with a CRS, we conduct a manual inspection of the worker's open comments. We only report aspects based on dialogue-level user satisfaction.

Table 4. Additional Aspects Captured from the Open Comments

Aspect	Definition	Annotator comment
Opinion (2.4%)	System expresses general opinions on a generic topic or expressing strong personal opinion	"I do not 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 repeats 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"

The % shows how often the aspect was stated.

We go through the comments and assign them to evaluation aspects based on the worker's perspective. For example, 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 4 lists the (dominant) novel categories discovered from the comments, together with a gloss and example. Several notable aspects are observed by the annotators. For instance, most annotators dislike the fact that the system expresses 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 are negatively impacted. This observation is in line with [25], where they show that overexposure of an item to a user in a short time period leads to a drop in user satisfaction. Some annotators note the positive impact of dialogue being natural and human-like or that the system makes a good recommendation after several failed suggestions (i.e., success on the last interaction). There are some examples where all annotators agree that the suggestions are good, but the user does not react rationally.

5.3 Summary

To summarize, in this section, we have first established the relationship between several dialogue aspects with user satisfaction. We have then conducted an analysis of the annotators' open comments to identify additional aspects that could influence a user's satisfaction. We conclude that at the turn-level relevance is the most important aspect whereas, at the dialogue level, the ability of the system to generate a user's interest and accomplish a task is significant in determining a user's overall satisfaction with a CRS. Therefore, we notice that the proposed dialogue aspects influence users' interaction with CRS differently. For some, a relevant recommendation has more effect on their overall rating whereas others consider the ability of the system to make relevant recommendations in a natural way as the most important factor influencing their overall satisfaction. Thus, user satisfaction is subjective to individual users, and the design and development of CRS should cater for personalization to individual users.

6 PREDICTING USER SATISFACTION

In this section, we present our approach to predicting user satisfaction in CRS. We discuss the problem formulation, models used, and the evaluation metrics for both turn and dialogue level user satisfaction.

6.1 Turn-Level Satisfaction Estimation

Task success [58] is a measure used in the evaluation of dialogue systems. This metric evaluates the quality of a dialogue with the assumption that users only care about their tasks being accomplished at the expense of **interaction quality (IQ)**. Since an annotator has to accurately determine a user's intended task, the metric is not accurate enough to estimate the quality of a dialogue response. Differently, in this work, we choose turn-level satisfaction (TSat) to determine the overall quality of a response in a dialogue. TSat estimation requires each turn to be annotated at a 5-point Likert scale. Unlike Sun et al. [62], who obtain the overall response quality at each turn, our annotation scheme requires annotators to rate three randomly sampled responses from each dialogue on two dialogue aspects, namely, *relevance and interestingness.* Then, we ask them to provide their overall quality rating. Response quality estimation could be used to identify the effect of a certain response on overall user satisfaction from a user's perspective.

6.1.1 Problem Definition. To answer **RQ2**: *Can we estimate user satisfaction at each turn from turn-level aspects?*, we formulate turn-level user satisfaction estimation as a regression problem. That is, given a randomly sampled turn T_i, with ratings for both the relevance (R_i) and interestingness (I_i) aspects, can we estimate a user's overall quality (O_i) rating for the given response? For example, a dialogue response rated 4 and 3 for the relevance and interestingness aspects, respectively, our task is to predict the user's overall response rating from these dialogue aspects. Using these turn-level aspects alleviates the need to manually craft features to predict turn-level satisfaction since our results show a comparative performance of simple machine learning models in estimating the quality at the response level.

6.1.2 Regression Methods. We consider various regression models similar to [6] for predicting overall response quality rating on a continuous scale of 1–5. We experiment with five popularly used models for regression, including **linear regression (LR)** [68], linear **support vector machine (SVM)** [18], **decision tree regressor (DTR)** [9], **random forest regressor (RFR)** [8], and **gradient boosting regressor (GBR)** [23], which ranks features by their importance.

6.1.3 Evaluation Criteria. For evaluation, we use popular evaluation metrics for regression tasks, namely, mean-squared error (MSE), root-mean-squared error (RMSE), and mean-absolute error (MAE). Following Bodigutla et al. [7], we also report the Pearson's r correlation coefficient for the performance of each model's 1–5 prediction, compared to the ground-truth human labels.

We implement the regression models (mentioned in Section 6.1.2) using scikit-learn.[6] For each model, we use five-fold cross-validation to tune the hyper-parameters and select the best values based on mean-squared error (MSE) on the validation set.

6.2 Dialogue-Level Satisfaction Estimation

We now investigate **RQ3** in this section: *How effective are dialogue aspects in estimating user satisfaction compared to turn-level satisfaction ratings?* In previous work, dialogue-level user satisfaction for task-oriented systems has been estimated leveraging rich signals such as user intents, dialogue acts, turn-level satisfaction ratings, and implicit turn and dialogue features [41, 62]. One major limitation of estimating overall user satisfaction using turn-level satisfaction ratings is the inability to capture specific aspects influencing a user's overall impression with a dialogue system. In this work, we propose to estimate overall user satisfaction from several dialogue aspects annotated in Section 4. We report on a performance comparison between the two approaches and show that estimating user satisfaction from dialogue aspects leads to a better-performing model.

---

[6] https://scikit-learn.org/

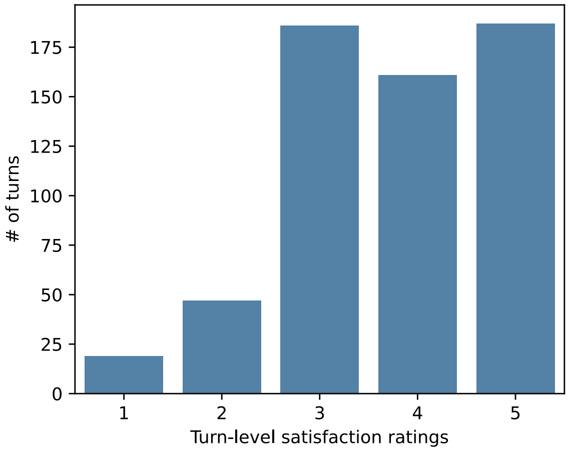


Fig. 6. Distribution of turn-level overall quality ratings.

6.2.1 Problem Definition. We formulate the overall user satisfaction estimation problem as a supervised binary classification task. Given the dialogue aspects' ratings, the goal is to classify the dialogue as either Sat or DSat. Due to label imbalance, we split the classes with dialogues (rating > 3) representing the satisfactory class and dissatisfactory class for dialogues (rating ≤ 3).

6.2.2 Classification Methods. To estimate the overall quality of a dialogue system, we consider several classification models: **logistic regression (Lr)**, a **support vector machine (SVM)** [18], a **decision tree classifier (DTC)** [9], a **random forest classifier (RFC)** [8], a **Gaussian naive Bayes (GNB)** [31], and a **gradient boosting classifier (GBC)** [23].

6.2.3 Evaluation Criteria. As evaluation metrics, we adopt four commonly used metrics for binary-classification task: **precision (Prec)** measures the proportion of correct predicted dialogue labels to the number of predicted dialogue labels, **recall (Rec)** refers to the percentage of correct predicted dialogue labels to the actual number of dialogue labels, and **F1-score (F1)** is the harmonic mean of precision and recall. Due to the high label imbalance for the **Sat** class (the **Sat** class is three times the size of the **DSat** class), we do not use the accuracy metric. To understand how each model is performing, we report results for each class separately.

As with the models in Section 6.1.2, we implement the classification models with scikit-learn. For each model, we use five-fold cross-validation. To search for optimal hyper-parameters, we use grid-search. The best values were selected based on F1-DSat. We train our predictors based on several aspects of combination variants.

7 RESULTS

In this section, we present our prediction results for both turn- and dialogue-level user satisfaction. **turn-level satisfaction (TSat)** is predicted with ratings from turn-level aspects (i.e., *relevance and interestingess*) whereas dialogue level user satisfaction is predicted from three types of ratings: First from the TSat ratings, second, dialogue-level aspects' ratings, and finally ratings combined from both the dialogue level aspects and TSat ratings.

7.1 Turn-Level Satisfaction

Figure 6 shows the distribution of human-annotated response quality ratings. We note that 62% of the turns are Sat (rating > 3) compared to DSat (38%) (rating ≤ 3). Turn-level satisfaction prediction

Table 5. Comparison of the Performance of Regression Models in Estimating Response Quality Measured Using Mean Squared Error (MSE), Root-Mean-Squared Error (RMSE), and Correlation between the Predicted and Ground Truth Labels

Model	MSE	RMSE	Pearson's r
linear regression	0.7762	0.8810	0.6017
support vector machine	0.8723	0.9339	0.5526
decision tree regressor	0.6089	0.7803	0.7234
random forest regressor	0.5901	0.7681	0.7337
gradient boosting regressor	0.6181	0.7862	0.7197

All correlations in this table are statistically significant ($p < 0.01$).

is very useful in online evaluation for identifying a problematic turn in a dialogue, thus allowing the system to adjust its recommendation or dialogue policy to avoid total dissatisfaction of the user by recovering from errors during the conversation.

At the turn level, the aim is to estimate the quality of the response from the annotated turn-level dialogue aspects, thus, we utilize graded satisfaction prediction in this task. We compare the performance of various regression models in estimating a user's response quality rating given the relevance and interestingness ratings for the current turn and report the results in Table 5. Evidently, all models perform comparatively well in estimating the user rating of each response. We note that ensemble models seem to learn a good representation of the aspects and improve their predictive performance compared to single models. The performance of traditional machine learning models is a clear indication that turn-level aspects can be used to estimate the quality of a response in cases where we do not have the user's turn-level satisfaction rating.

We also report the correlation coefficient between the predicted labels and the ground truth labels for each model. Among the six models we experimented with, RFR achieves the highest correlation coefficient (0.7337) followed closely by DTR at 0.7234. Our analysis of the predicted labels reveals that in most cases, the models predict accurately or close to the ground truth label for satisfactory dialogues compared with dissatisfactory dialogues. Identifying turns where the system fails is a difficult task due to label imbalance, as the majority of the turns are rated as satisfactory. It is worth noting that identifying dissatisfactory turns is more important in order for CRSs to adjust their interaction policy and avoid total user dissatisfaction.

In summary, extensively experimenting with the dialogue aspects as features, we conclude that both relevance and interestingness are important in predicting the quality of a response with CRS. We note that the random forest regressor achieves a high correlation coefficient of 0.7337 compared to other models. Thus, in cases where we do not have access to user's response quality ratings, we can rely on dialogue aspects such as relevance and interestingness to estimate the quality of a response.

7.2 Dialogue-Level Satisfaction

To show how effective the proposed dialogue aspects are in predicting user satisfaction, we report the results for several classical machine learning models on user satisfaction prediction. First, we predict overall user satisfaction from turn-level satisfaction ratings (see Table 6). Second, we experiment with a combination of turn- and dialogue-level aspects separately (see Table 7). Finally, to show the effectiveness of our proposed dialogue aspects, we predict user satisfaction from all the proposed dialogue aspects (see Table 8.)

Table 6 shows the performance of several machine learning models in predicting user satisfaction from turn-level satisfaction ratings. We report the evaluation metrics for both the Sat and

Table 6. Performance of Machine Learning Methods with a Variant, Predicting User Satisfaction Using Turn-Level Satisfaction Ratings, where the Best precision (Prec), recall (Rec), and F1-score (F1) for both the satisfactory (Sat) and dissatisfactory (DSat) Class are in Bold

	Prec		Rec		F1		
Models	Sat	DSat	Sat	DSat	Sat	DSat	Spearman's ρ
logistic regression	**0.93**	**1.00**	**1.00**	**0.56**	**0.96**	**0.71**	**0.7177**
support vector machine	0.91	0.67	0.96	0.44	0.93	0.53	0.4823
decision tree classifier	0.92	0.42	0.86	**0.56**	0.89	0.48	0.3734
random forest classifier	0.92	0.50	0.90	**0.56**	0.91	0.53	0.4383
Gaussian naive Bayes	0.92	0.62	0.94	**0.56**	0.93	0.59	0.5217
gradient boosting classifier	0.92	0.50	0.90	**0.56**	0.91	0.53	0.4383

All correlations in this table are statistically significant ($p < 0.01$).

Table 7. Performance Comparison of Machine Learning Methods with a Variant Predicting User Satisfaction from Turn-Level Dialogue Aspects vs Dialogue-Level Aspects, where the Best precision (Prec), recall (Rec) and F1-score (F1) for Both satisfactory (Sat) and dissatisfactory (DSat) Class are in Bold

	precision		recall		F1-score		
Models	Sat	DSat	Sat	DSat	Sat	DSat	Spearman's ρ
Turn-level Aspects							
support vector machine	0.86	**0.75**	**0.96**	0.38	0.91	0.50	0.4583
random forest classifier	0.90	0.56	0.88	0.62	0.89	0.59	0.4789
gradient boosting classifier	0.94	0.67	0.91	0.75	0.92	**0.71**	**0.6286**
Dialogue-level Aspects							
support vector machine	0.91	0.67	**0.96**	0.44	**0.93**	0.53	0.4823
random forest classifier	**0.96**	0.58	0.90	**0.78**	**0.93**	0.67	0.6067
gradient boosting classifier	**0.96**	0.58	0.90	**0.78**	**0.93**	0.67	0.6067

All correlations in this table are statistically significant ($p < 0.01$).

DSat classes, except for the correlation coefficient so as to capture the performance of the models in predicting a dissatisfactory dialogue. This is because identifying a problematic dialogue is of more importance for system designers to improve the model's performance for the next interaction. We note that for the Sat class, the models perform better in Prec, Rec, and F1 metrics than the DSat class. In terms of F1-DSat and Spearman's ρ, logistic regression is the best-performing model. This model classifies all the predicted satisfactory dialogues accurately as it achieves a recall score of 1.00 compared to 0.56 for dissatisfactory dialogues. Apart from having the limitation of dataset size representing dissatisfactory dialogues, it indicates that it is challenging for the model to identify dialogues where the user is dissatisfied since most of the data represent positive dialogues. Thus, understanding dialogue aspects that can easily be used to identify problematic dialogues is useful.

Additionally, we note that predicting user satisfaction from turn-level satisfaction ratings does not lead to a good performance for the DSat class. This demonstrates that user satisfaction ratings at each turn are not optimal in estimating whether a whole dialogue is dissatisfactory or not. We hypothesize that all turns are not equally weighted by the users when determining their overall satisfaction. Our experiments on predicting user satisfaction from individual turns reveal that the last turn is more important compared to the other turns in predicting user satisfaction. This

Table 8. Performance of Machine Learning Methods with a Variant Predicting User Satisfaction Using Ratings from all the Proposed Dialogue Aspects where the Best precision (Prec), recall (Rec) and F1-score (F1) for both satisfactory (Sat) and dissatisfactory (DSat) Class are in Bold

	precision (Prec)		recall (Rec)		F1-score (F1)		
Models	Sat	DSat	Sat	DSat	Sat	DSat	Spearman's ρ
logistic regression	0.93	0.83	0.98	0.56	0.95	0.67	0.6379
support vector machine	**0.96**	0.87	0.98	0.67	0.96	**0.80**	0.7934
decision tree classifier	0.94	0.67	0.91	**0.78**	0.92	0.71	0.6067
random forest classifier	0.94	**1.00**	**1.00**	0.67	0.97	**0.80**	**0.7956**
Gaussian naive Bayes	**0.96**	0.58	0.90	**0.78**	0.93	0.68	0.6029
gradient boosting classifier	**0.96**	0.79	0.93	**0.78**	0.94	0.78	0.7385

All Correlations in This Table are Statistically Significant ($p < 0.01$).

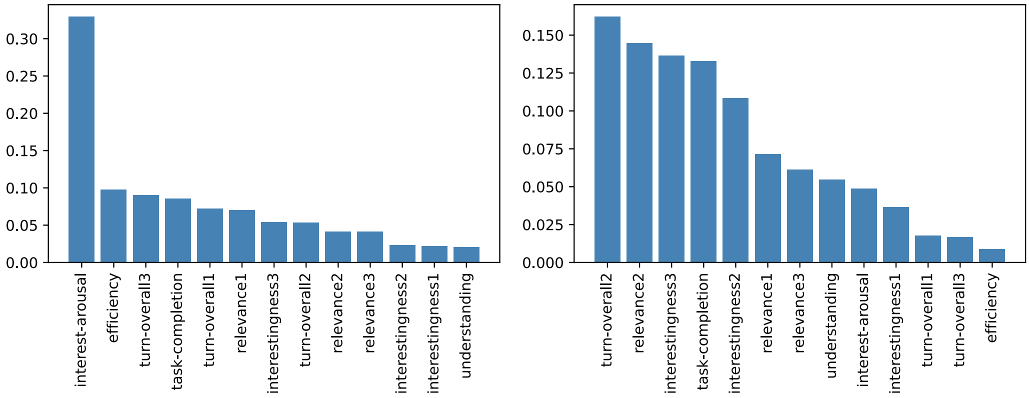
indicates that the ability of the system to have a successful last interaction impacts a user's overall impression.

In Table 8, we observe an increase in the performance of F1-DSat when we predict user satisfaction from all the annotated dialogue aspects. For precision, random forest performs better in the DSat class, both decision tree and GNB are the best-performing models in terms of recall, with random forest and SVM scoring a high F1-DSat. The predictions of random forest have a high correlation score with the ground truth labels, followed closely by SVM predictions. Although we do not experiment with neural architectures to allow us to model the dialogue context, all models indicate a comparative performance in predicting user satisfaction from dialogue aspects with moderate correlation scores. Thus, this implies that traditional machine learning approaches can be leveraged in user satisfaction prediction and we can rely on dialogue aspects ratings to predict user satisfaction and get comparative results without context modeling and additional implicit features.

Taking the three best-performing models from Table 8 (SVM, RFC, and GBC), we experiment with predicting user satisfaction using turn- and dialogue-level aspects and report the results in Table 7. GBC performs better in terms of F1-DSat for both the turn and dialogue levels, 0.71 and 0.67, respectively. All models perform better for precision, recall, and F-1 for the Sat class. We note a superior performance when predicting user satisfaction with the dialogue level aspects compared to the turn level aspects suggesting dialogue level aspects benefit the models more in identifying satisfactory dialogues. The DSat class seems to benefit more from the turn-level aspects when combined with turn-level satisfaction as we observe a high F1-DSat from this level. It is worth noting that, though we observe a high F1-DSat when predicting user satisfaction from the turn-level aspects, GBC and RFC from the dialogue-level aspects (see Table 7 row 5) achieve a high recall score for the DSat class showing their capability to accurately classify the predicted dialogues as dissatisfactory compared to the methods using turn-level features. We also report the correlation coefficients in Table 7 and note a comparative performance for GBC in both turn and dialogue level aspects.

Feature analysis. Since we experiment with several combinations of the aspects, we treat the aspects as our input features and conduct a feature importance analysis using RFC. As we report our result per class (i.e., Sat and DSat) we also report the importance of each feature based on each class, in addition to overall satisfaction prediction.

Figure 7 shows the significant percentage of features for (a) the Sat class and (b) the DSat class. The ability of the system to arouse a user's interest to watch an unseen movie is the most significant feature for Sat class. We note a five percent gap between the most significant feature (*Interest arousal*- 16%) and the second most important feature (*turn-overall3* at 11%). Closely followed by *turn-overall1, task-completion* and *relevance1*. This indicates that in order for a CRS to improve a



(a) Satisfactory (Sat)

(b) Dissatisfactory (DSat)

Fig. 7. Bar plots showing the importance of the dialogue aspects as input features for predicting, (a) satisfactory (Sat) class and (b) dissatisfactory (DSat) class using RFC model. For turn-level aspects, we represent the performance of the three turns toward user satisfaction prediction where *relevance* $1-3$,*interestingness* $1-3$, and *turn-overall* $1-3$ are labels at turn $1-3$.

user's interaction experience, it should create a good impression to the user at the start and end of a conversation.

We see that a user's overall impression in turn two is the most significant feature in predicting user dissatisfaction for the entire dialogue, as shown in Figure 7(b). Followed closely by *relevance2*, *interestingness3*, *task-completion* and *interestingness2*. Out of the top five features, we note that 3 of them are rated at the second turn that is *turn-overall2*, *relevance2*, *interestingness2*. In most dialogues, we examined, recommendations start at this turn after preference elicitation in turn one. If a system fails to capture a user's preference in the first turn, in most cases, it leads to an irrelevant recommendation being made, resulting in overall dissatisfaction. In order to improve the performance of the system at turn two, the system should be more understanding toward a user's request and preference. The features, *efficiency*, *turn-overall3*, *turn-overall1*, and *interestingness1* are the least significant in the prediction of the DSat class.

7.3 Summary

In general, we note that combining features from both the utility and user experience dimensions leads to a better user satisfaction measurement. In both the Sat and DSat classes, turn- and dialogue-level aspects are important. For Sat, the strong signal is interest arousal, which is measured at the dialogue level, whereas turn-level satisfaction at response two (*turn-overall2*) is the strongest in the DSat class. Evidently, we can conclude that features from both the turn and dialogue levels are important in determining satisfactory and dissatisfactory dialogues with CRSs. Therefore, based on results from Tables 6 and 8, we show that relying on only dialogue-level aspects to predict user satisfaction is as effective as using turn-level satisfaction ratings.

8 DISCUSSION AND LIMITATIONS

In this section, we present an analysis of our key findings and their significance, on understanding and predicting user satisfaction in CRS motivated by our experimental results. Furthermore, we examine the limitations of our research, primarily based on the methodology employed throughout this study. We delve into more details below.

8.1 Discussion

In this work, we have first focused on understanding user satisfaction with CRSs, generally categorized as a goal-oriented dialogue system. Although goal-oriented dialogue systems are ideally expected to optimize toward task accomplishment, in this study, we show that a system's behavior during interaction has an influence on their overall satisfaction during interactions at both the turn and dialogue levels. The interestingness aspect, however, does not show a high correlation with turn-level satisfaction. We hypothesize that when asked to scrutinize a CRS response explicitly on interestingness, annotators tend to rate such responses less favorably than they would if they were rating the overall experience according to the established rating process. Though this aspect is highly researched for non-task-oriented dialogue systems [7, 50, 51], from both the annotations and open-ended comments, we find that engaging with users in the form of chit-chat has both positive and negative effects on their overall satisfaction. If a user is already happy with a provided recommendation, more engagement can lead to further *interest arousal*, and hence more satisfaction; however, if the system fails to meet the user's expectations, it can have a negative effect. This is in line with [61], who stress the importance of finding the right amount of chit-chat in a goal-oriented dialogue.

Providing relevant recommendations throughout a dialogue is crucial for user satisfaction, but it does not tell the whole story. When a system makes relevant recommendations, they certainly lead to a satisfactory dialogue, but when the responses are both relevant and interesting most users tend to rate their experience as very satisfactory for both levels. This indicates that a CRS that can make relevant recommendations alongside generating natural responses that are interesting is more likely to result in an improved user's overall interaction experience. Thus, system designers and dataset creators should consider optimizing these two aspects during the design and development of CRS systems and datasets.

Our analysis of the justifications that support a user's overall satisfaction rating reveals new aspects that can affect users' satisfaction. In line with our quantitative analysis and related work [42, 47], many annotators mention the importance of a good user experience in the final turns of a conversation. Success in the last interaction has an implication on task completion, interest arousal, and on overall user satisfaction. When a system accomplishes its predefined goal, users tend to utter responses such as "Thank you for the suggestion!" "It was nice chatting with you." While utterances such as "But you did not get me something to watch" and "Such a waste of my time" indicate an inability of the system to fulfill a user's need. Therefore, in various cases, we can rely on the last user interaction to assess the system's ability to fulfill or not fulfill a user's need. It is also worth mentioning that other aspects such as repeated utterances and recommendations negatively impacted the user experience.

In general, we note that the UX dimension (*interestingness*, *understanding*, *interest arousal* and *efficiency*) of a CRS plays a very important role toward user satisfaction. The ability of a CRS to make relevant recommendations and accomplish a user's goal could lead to overall satisfaction, however, a system that demonstrates to be more engaging and understanding has a higher chance of satisfying users. This indicates the need to jointly optimize turn- and dialogue-level metrics and for a fine-grained model of user satisfaction that incorporates multiple aspects.

8.2 Limitations

In this work, we rely on external assessors to judge user satisfaction based on the user's utterances and reactions to the system's responses. While we have observed a high level of agreement for most dialogues, we have also noticed disagreement between annotators on some. This limitation could introduce a potential gap between the assessors' ratings and the subjective satisfaction levels of users in real-world scenarios. Additionally, interpretation biases among assessors can affect the reliability of turn and dialogue-level ratings. Therefore, it is essential to conduct this study with actual users so as to collect a set of fine-grained annotations from real users [49].

At the turn level, due to the substantial annotation effort required, following [51], we sample three responses from each dialogue for annotation. While this approach may not capture the full picture in a dialogue, we believe that our sampling strategy provides meaningful insights into what aspects influence turn-level satisfaction. Investigating the optimal way of selecting responses to annotate from each dialogue may provide additional useful findings, but this was not our concern in this study. Therefore, we think there is a rich research gap to solve the significant annotation effort required in dialogue annotations when all the turns are annotated.

9 CONCLUSION AND FUTURE WORK

In this article, we have focused on a user-oriented approach to understanding user satisfaction in conversational recommendations. We have conducted a study to assess the influence of multiple dialogue aspects on overall user satisfaction. Through a carefully designed annotation process, we have collected external assessors' feedback ratings on six dialogue aspects (*relevance*, *interestingness*, *understanding*, *task completion*, *interest arousal*, and *efficiency*) and user satisfaction at the turn and dialogue level. With this data, we have investigated the relationship between several dialogue aspects and user satisfaction. Furthermore, we have adopted several machine learning methods to predict response quality and overall user satisfaction with different feature combinations.

Combining both the qualitative and quantitative methods, our results indicate that: (i) Relevant recommendations are necessary but not sufficient for high user satisfaction feedback. Therefore, several aspects should be considered in estimating a user's overall satisfaction with a CRS. (ii) In the absence of response quality ratings, we can rely on turn-level aspects to estimate the user's rating for each response. And (iii) user satisfaction can be predicted more accurately with combined dialogue aspects as features unlike only using turn-level satisfaction ratings.

In addition to understanding how several dialogue aspects influence a user's overall satisfaction with a CRS, our findings also have implications for the design and evaluation of CRSs. Our results show that predicting user satisfaction with aspects representing the utility of a CRS (relevance and task completion) performs poorly compared to predicting with a combination of all aspects. Thus, in order to achieve high user satisfaction, the design of CRSs should not only be optimized toward goal accomplishment but also a good user interaction experience.

Our experimental results with traditional machine learning methods indicate a strong performance. We have not experimented with neural network architectures in this study as it is not the main focus of our work, and we leave this to future work. Furthermore, other dialogue features such as dialogue context, intent, and system-user action could be modeled in a neural architecture as they have proven to improve user satisfaction prediction. Since our study involves a small sample dataset, we plan to verify our findings on a larger scale and with diverse data collected from actual users interacting with the system. Collecting a large-scale dataset can be achieved in an automatic way by leveraging existing predictive models to capture key patterns by training them with explicit ratings or in an unsupervised way. Apart from that, techniques such as user simulation can be used to provide annotated user feedback within dialogues, thus increasing the amount of data to be annotated [5], where this feedback can include explicit ratings on the dialogue aspects allowing for the collection of ground truth data for training and automatic evaluation at scale.

Though the focus of our study is to uncover the relationship between various dialogue aspects and user satisfaction, we believe our findings can provide insights into the factors that contribute to increased user satisfaction in CRS and can serve as a basis for future research and system development. We, therefore, encourage future research to investigate the practical implications of our findings by looking at the impact of increasing dialogue aspects on user satisfaction through experimental studies or user-centered evaluations using tools such as CRSLab [73] to compare different CRS methods.

For future work, we are interested in integrating large language models in the annotation process to further enhance the accuracy, richness, and scale of the annotated dataset. We hypothesize that their advanced contextual understanding and semantic analysis capabilities will benefit the annotations. In particular, following [21], we expect that the annotations on the recommended items will more closely align with user preferences and intents expressed in the conversation.

APPENDIX

A INSTRUCTIONS FOR ASSESSORS

Tables 9–11 show the annotation instructions given to the assessors during the human quality annotation process. Figure 8 shows a sample interface that was used for dialogue-level annotation. In Table 12, we show a dialogue example with assessors annotations. These instructions and examples are a sample of what was shown to the assessors.

Table 9. Annotation Instructions Given to the Workers (1/3)

Annotation instructions	
In this task, your goal is to rate how well an intelligent SYSTEM (like Siri or Alexa) converses with a USER. The USER is looking for some movies and the SYSTEM tries to understand what the USER likes to finally give some suggestions to the USER. You will rate the quality of the provided SYSTEM responses and the overall dialogue.	
Turn-level annotation	
Relevance (1–4)	This means the response is appropriate to the previous utterance and a movie was mentioned that fulfills a user's goal, that is the user liked it, has seen it, or agreed to watch it. (1) Not applicable: there is no movie recommended to the user in the response (2) Irrelevant: the SYSTEM recommends a movie, but the user does not like the movie and mentions this fact in their response (3) Can't say: the SYSTEM recommends a movie, but the user does not express any opinions. So it's impossible to say whether the user likes the movie or not (4) Relevant: the SYSTEM recommends a movie and the user expresses a positive opinion in their utterance
Interestingness (1–3)	This means: the SYSTEM suggested a movie in the response accompanied by some small talk which would make a user want to continue interacting with the SYSTEM. (1) Not interesting: the SYSTEM makes small talk that is generic, dull, or only states a movie name (2) Somewhat interesting: the SYSTEM makes small talk that is specific to the provided context but does not make any recommendation (3) Interesting: the SYSTEM recommends a movie while making small talk
Turn-overall (1–5)	What is your overall impression of the system response? (1) Terrible: the SYSTEM does not understand the user's interest and does not fulfill it and the user expresses a negative opinion in their utterance (2) Bad: the SYSTEM understands the user's interest but fails to fulfill it and the user expresses a negative opinion in their utterance (3) Ok: the SYSTEM understands the user's interest and partially fulfills it and the user does not express any opinion in their utterance (4) Good: the SYSTEM understands the user's interest and fulfills it and the user expresses curiosity in their utterance (5) Excellent: the SYSTEM understands the user's interest and fulfills it and the user expresses a positive opinion in their utterance

Table 10. Annotation Instructions Given to the Workers (2/3)

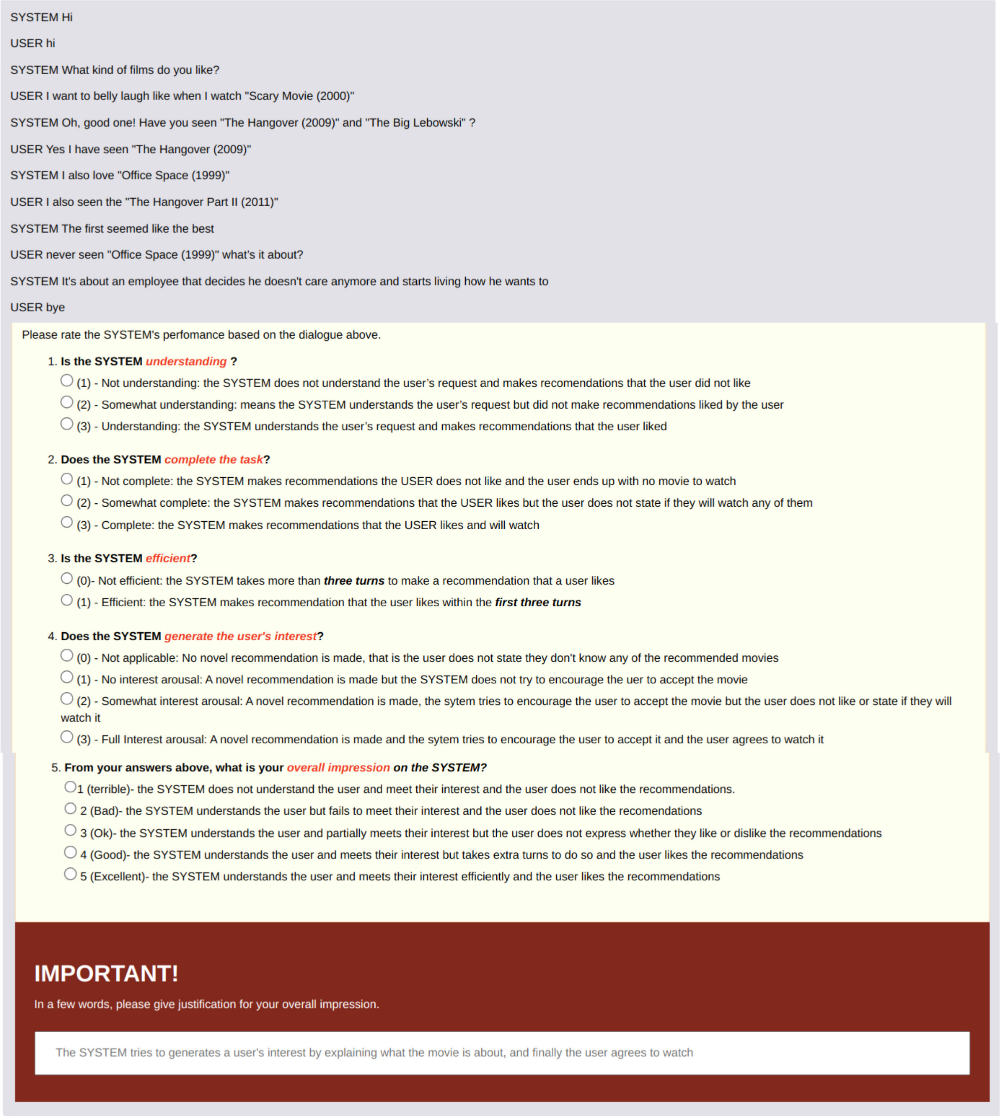
Annotation instructions (Cont'd)	
Dialogue-level annotation	
Understanding (1–3)	This means: the SYSTEM understands the user's request and makes a recommendation meeting their interest. (1) Not understanding: the SYSTEM does not understand the user's request and makes recommendations that the user did not like (2) Somewhat understanding: means the SYSTEM understands the user's request but did not make recommendations liked by the user (3) Understanding: the SYSTEM understands the user's request and makes recommendations that the user liked
Task completion (1–3)	This means: the SYSTEM makes recommendations that either the user "likes" or "has seen" and agrees to watch one of the recommendations by the end of the conversation. (1) Not complete: the SYSTEM makes recommendations the USER does not like and the user ends up with no movie to watch (2) Somewhat complete: the SYSTEM makes recommendations that the USER likes but the user does not state if they will watch any of them (3) Complete: the SYSTEM makes recommendations that the USER likes and will watch
Interest arousal (1–4)	This means: the SYSTEM makes a novel recommendation and tries to encourage the user to like and watch it by giving more explanation about the movie. (1) Not applicable: no novel recommendation is made, that is the user does not state they don't know any of the recommended movies (2) No interest arousal: a novel recommendation is made but the SYSTEM does not try to encourage the user to accept the movie (3) Somewhat interest arousal: a novel recommendation is made, and the system tries to encourage the user to accept the movie but the user does not like or state if they will watch it (4) Full Interest arousal: a novel recommendation is made and the system tries to encourage the user to accept it and the user agrees to watch it
Efficiency (0–1)	This means: the SYSTEM makes recommendations that meet the user's interest within the first three turns. (1) Not efficient: the SYSTEM takes more than three turns to make a recommendation that the user likes (2) Efficient: the SYSTEM makes recommendations that the user likes within the first three turns. A turn: in this case are two consecutive utterances from the SYSTEM and the USER.

Table 11. Annotation Instructions Given to the Workers (3/3)

Annotation instructions (Cont'd)	
Dialogue-level annotation	
Dialogue-overall	What is your overall impression of the system? (1) Terrible: the SYSTEM does not understand the user and meet their interest and the user does not like the recommendations. (2) Bad: the SYSTEM understands the user but fails to meet their interest and the user does not like the recommendations (3) Ok: the SYSTEM understands the user and partially meets their interest but the user does not express whether they like or dislike the recommendations (4) Good: the SYSTEM understands the user and meets their interest but takes extra turns to do so and the user likes the recommendations (5) Excellent: the SYSTEM understands the user and meets their interest efficiently and the user likes the recommendations

Table 12. Sample Turns Annotated in the Task with Worker Ratings for Relevance, Interestingness and Turn-Overall Impression

Turns	Utterances
1	SYSTEM hello how can i help you this morning USER hello im looking for some newer movies no preference SYSTEM Okay how about “Thor: Ragnarok (2017)”? Or “Justice League (2017)”? USER I haven't seen either one of those
	relevance: [3, 3, 4]; interestingness: [1, 1, 3]; Turn-overall impression: [3, 3, 4]
2	SYSTEM They are both great films USER they sound great I will try those out on our movie night” SYSTEM you should check them out on DVD's And how about “Split (2017)”? USER I havent seen that one either
	relevance: [3, 3, 4]; interestingness: [3, 3, 3]; Turn-overall impression: [3, 4, 5]
3	SYSTEM good film! USER sounds good SYSTEM “REC (2007)”? Its about a zombie apocalypse you should check it out! USER sound really good I will check them out! thanks for the recommendations
	relevance: [4, 4, 4]; interestingness: [3, 3, 3]; Turn-overall impression: [5, 5, 5]



SYSTEM Hi

USER hi

SYSTEM What kind of films do you like?

USER I want to belly laugh like when I watch "Scary Movie (2000)"

SYSTEM Oh, good one! Have you seen "The Hangover (2009)" and "The Big Lebowski" ?

USER Yes I have seen "The Hangover (2009)"

SYSTEM I also love "Office Space (1999)"

USER I also seen the "The Hangover Part II (2011)"

SYSTEM The first seemed like the best

USER never seen "Office Space (1999)" what's it about?

SYSTEM It's about an employee that decides he doesn't care anymore and starts living how he wants to

USER bye

Please rate the SYSTEM's perfomance based on the dialogue above.

1. **Is the SYSTEM *understanding* ?**
   - ○ (1) - Not understanding: the SYSTEM does not understand the user's request and makes recomendations that the user did not like
   - ○ (2) - Somewhat understanding: means the SYSTEM understands the user's request but did not make recommendations liked by the user
   - ○ (3) - Understanding: the SYSTEM understands the user's request and makes recommendations that the user liked
2. **Does the SYSTEM *complete the task*?**
   - ○ (1) - Not complete: the SYSTEM makes recommendations the USER does not like and the user ends up with no movie to watch
   - ○ (2) - Somewhat complete: the SYSTEM makes recommendations that the USER likes but the user does not state if they will watch any of them
   - ○ (3) - Complete: the SYSTEM makes recommendations that the USER likes and will watch
3. **Is the SYSTEM *efficient*?**
   - ○ (0)- Not efficient: the SYSTEM takes more than ***three turns*** to make a recommendation that a user likes
   - ○ (1) - Efficient: the SYSTEM makes recommendation that the user likes within the ***first three turns***
4. **Does the SYSTEM *generate the user's interest*?**
   - ○ (0) - Not applicable: No novel recommendation is made, that is the user does not state they don't know any of the recommended movies
   - ○ (1) - No interest arousal: A novel recommendation is made but the SYSTEM does not try to encourage the uer to accept the movie
   - ○ (2) - Somewhat interest arousal: A novel recommendation is made, the sytem tries to encourage the user to accept the movie but the user does not like or state if they will watch it
   - ○ (3) - Full Interest arousal: A novel recommendation is made and the sytem tries to encourage the user to accept it and the user agrees to watch it
5. **From your answers above, what is your *overall impression on the SYSTEM?***
   - ○1 (terrible)- the SYSTEM does not understand the user and meet their interest and the user does not like the recommendations.
   - ○ 2 (Bad)- the SYSTEM understands the user but fails to meet their interest and the user does not like the recomendations
   - ○ 3 (Ok)- the SYSTEM understands the user and partially meets their interest but the user does not express whether they like or dislike the recommendations
   - ○ 4 (Good)- the SYSTEM understands the user and meets their interest but takes extra turns to do so and the user likes the recommendations
   - ○ 5 (Excellent)- the SYSTEM understands the user and meets their interest efficiently and the user likes the recommendations

IMPORTANT!

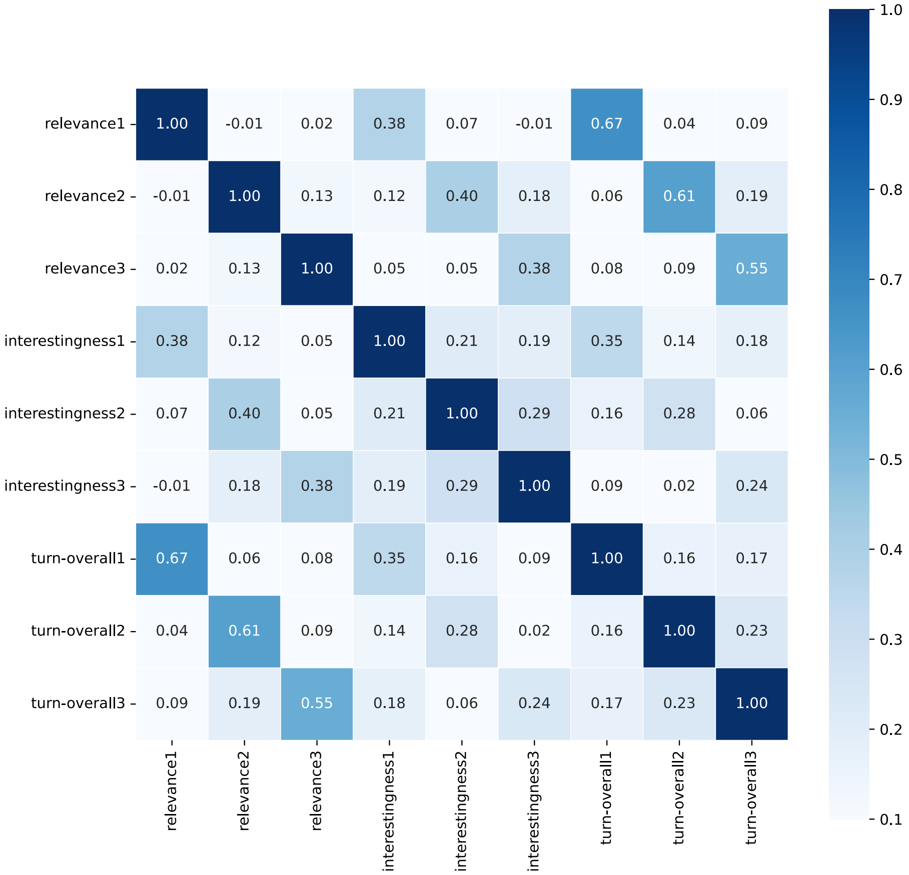
In a few words, please give justification for your overall impression.

The SYSTEM tries to generates a user's interest by explaining what the movie is about, and finally the user agrees to watch

Fig. 8. Interface used by annotators to annotate dialogues at the dialogue level showing the annotation guidelines for each of the four dialogue aspects, overall user satisfaction, and annotators open comment box

B DATA AND CORRELATION ANALYSIS

In Figures 9–11, we report the correlation analysis for each individual turn at turn-level, dialogue-level aspects correlation and correlation between turn and dialogue levels aspects, respectively.



	relevance1	relevance2	relevance3	interestingness1	interestingness2	interestingness3	turn-overall1	turn-overall2	turn-overall3
relevance1	1.00	-0.01	0.02	0.38	0.07	-0.01	0.67	0.04	0.09
relevance2	-0.01	1.00	0.13	0.12	0.40	0.18	0.06	0.61	0.19
relevance3	0.02	0.13	1.00	0.05	0.05	0.38	0.08	0.09	0.55
interestingness1	0.38	0.12	0.05	1.00	0.21	0.19	0.35	0.14	0.18
interestingness2	0.07	0.40	0.05	0.21	1.00	0.29	0.16	0.28	0.06
interestingness3	-0.01	0.18	0.38	0.19	0.29	1.00	0.09	0.02	0.24
turn-overall1	0.67	0.06	0.08	0.35	0.16	0.09	1.00	0.16	0.17
turn-overall2	0.04	0.61	0.09	0.14	0.28	0.02	0.16	1.00	0.23
turn-overall3	0.09	0.19	0.55	0.18	0.06	0.24	0.17	0.23	1.00

Fig. 9. Correlation heatmap for turn-level aspects, for the three annotated responses.

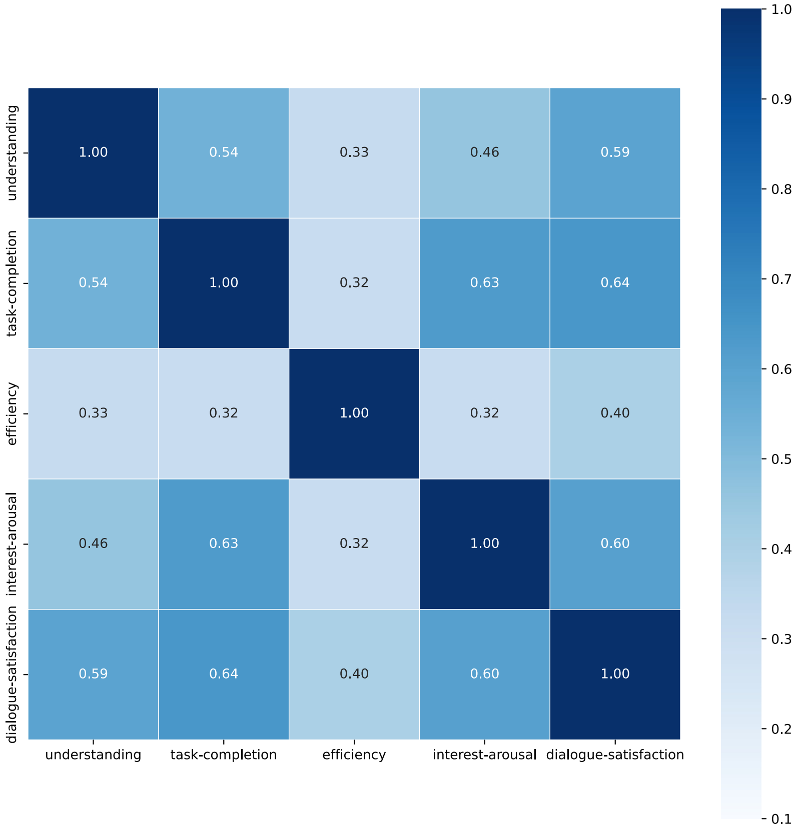
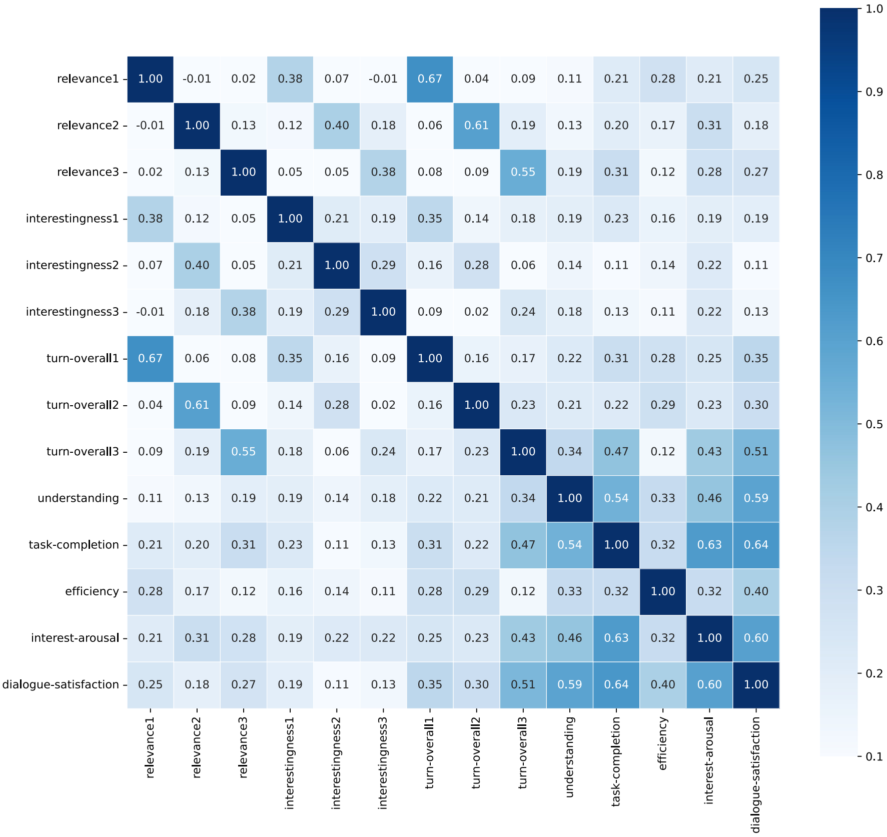


Fig. 10. Correlation heatmap for dialogue-level aspects.



	relevance1	relevance2	relevance3	interestingness1	interestingness2	interestingness3	turn-overall1	turn-overall2	turn-overall3	understanding	task-completion	efficiency	interest-arousal	dialogue-satisfaction
relevance1	1.00	-0.01	0.02	0.38	0.07	-0.01	0.67	0.04	0.09	0.11	0.21	0.28	0.21	0.25
relevance2	-0.01	1.00	0.13	0.12	0.40	0.18	0.06	0.61	0.19	0.13	0.20	0.17	0.31	0.18
relevance3	0.02	0.13	1.00	0.05	0.05	0.38	0.08	0.09	0.55	0.19	0.31	0.12	0.28	0.27
interestingness1	0.38	0.12	0.05	1.00	0.21	0.19	0.35	0.14	0.18	0.19	0.23	0.16	0.19	0.19
interestingness2	0.07	0.40	0.05	0.21	1.00	0.29	0.16	0.28	0.06	0.14	0.11	0.14	0.22	0.11
interestingness3	-0.01	0.18	0.38	0.19	0.29	1.00	0.09	0.02	0.24	0.18	0.13	0.11	0.22	0.13
turn-overall1	0.67	0.06	0.08	0.35	0.16	0.09	1.00	0.16	0.17	0.22	0.31	0.28	0.25	0.35
turn-overall2	0.04	0.61	0.09	0.14	0.28	0.02	0.16	1.00	0.23	0.21	0.22	0.29	0.23	0.30
turn-overall3	0.09	0.19	0.55	0.18	0.06	0.24	0.17	0.23	1.00	0.34	0.47	0.12	0.43	0.51
understanding	0.11	0.13	0.19	0.19	0.14	0.18	0.22	0.21	0.34	1.00	0.54	0.33	0.46	0.59
task-completion	0.21	0.20	0.31	0.23	0.11	0.13	0.31	0.22	0.47	0.54	1.00	0.32	0.63	0.64
efficiency	0.28	0.17	0.12	0.16	0.14	0.11	0.28	0.29	0.12	0.33	0.32	1.00	0.32	0.40
interest-arousal	0.21	0.31	0.28	0.19	0.22	0.22	0.25	0.23	0.43	0.46	0.63	0.32	1.00	0.60
dialogue-satisfaction	0.25	0.18	0.27	0.19	0.11	0.13	0.35	0.30	0.51	0.59	0.64	0.40	0.60	1.00

Fig. 11. Correlation heatmap for turn and dialogue aspects.

ACKNOWLEDGMENTS

We thank our associate editor and reviewers for the constructive feedback that they gave on an earlier version of this paper.

This research was supported by the Dreams Lab, a collaboration between Huawei, the University of Amsterdam, and the Vrije Universiteit Amsterdam, 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, and by project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21, which is (partly) financed by the Dutch Research Council (NWO).

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REPRODUCIBILITY

To facilitate reproducibility of our work, we are sharing our data at https://github.com/Clemenciah/Understanding-User-Satisfaction-Data under the MIT license.

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. DOI: https://doi.org/10.1002/asi.21300

[2] Azzah Al-Maskari, Mark Sanderson, and Paul Clough. 2007. The relationship between ir effectiveness measures and user satisfaction. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, 773–774. DOI: https://doi.org/10.1145/1277741.1277902

[3] Omar Alonso and Ricardo Baeza-Yates. 2011. Design and implementation of relevance assessments using crowdsourcing. In *Advances in Information Retrieval*. Paul Clough, Colum Foley, Cathal Gurrin, Gareth J. F. Jones, Wessel Kraaij, Hyowon Lee, and Vanessa Mudoch (Eds.). Springer, Berlin,, 153–164.

[4] Avishek Anand, Lawrence Cavedon, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational search (dagstuhl seminar 19461). *Dagstuhl Reports* 9, 11 (2020), 34–83. DOI: https://doi.org/10.4230/DagRep.9.11.34

[5] Krisztian Balog and ChengXiang Zhai. 2023. User simulation for evaluating information access Systems. CoRR abs/2306.08550 (2023). https://doi.org/10.48550/arXiv.2306.08550 arXiv:2306.08550.

[6] Praveen Kumar Bodigutla, Lazaros Polymenakos, and Spyros Matsoukas. 2019. Multi-domain conversation quality evaluation via user satisfaction estimation. CoRR abs/1911.08567 (2019). arXiv:1911.08567 http://arxiv.org/abs/1911.08567

[7] Praveen Kumar Bodigutla, Aditya Tiwari, Spyros Matsoukas, Josep Valls-Vargas, and Lazaros Polymenakos. 2020. Joint turn and dialogue level user satisfaction estimation on multi-domain conversations. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3897–3909. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.347

[8] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

[9] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Wadsworth.

[10] Wanling Cai and Li Chen. 2020. Predicting user intents and satisfaction with dialogue-based conversational recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP'20)*. Association for Computing Machinery, New York, NY, 33–42. DOI: https://doi.org/10.1145/3340631.3394856

[11] Georges Casella. 2002. *Statistical Inference* (2nd ed.). Duxbury/Thomson Learning.

[12] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: Survey and emerging trends. *User Modeling and User-Adapted Interaction* 22, 1 (2012), 125–150.

[13] Cyril W. Cleverdon. 1967. The cranfield tests on index language devices. In *Aslib Proceedings*, Vol. 19. 173–192.

[14] Cyril W. Cleverdon. 1991. The significance of the cranfield tests on index languages. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, 3–12. DOI: https://doi.org/10.1145/122860.122861

[15] William S. Cooper. 1971. A definition of relevance for information retrieval. *Information Storage and Retrieval* 7, 1 (1971), 19–37. DOI: https://doi.org/10.1016/0020-0271(71)90024-6

[16] Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the 9th Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 376–380. DOI: https://doi.org/10.3115/v1/W14-3348

[17] 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, 1 (2020), 755–810.

[18] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alexander J. Smola, and Vladimir Vapnik. 1996. Support Vector Regression Machines. In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, Michael Mozer, Michael I. Jordan, and Thomas Petsche (Eds.). MIT Press, 155–161. http://papers.nips.cc/paper/1238-support-vector-regression-machines

[19] 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, 3806–3812. DOI: https://doi.org/10.18653/v1/N19-1381

[20] Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication* 50, 8 (2008), 630–645. DOI: https://doi.org/10.1016/j.specom.2008.04.002

[21] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR'23)*. Association for Computing Machinery, New York, NY, 39–50.

[22] Sarah E. Finch and Jinho D. Choi. 2020. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 236–245. Retrieved from https://aclanthology.org/2020.sigdial-1.29

[23] Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29, 5 (2001), 1189–1232.

[24] 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, 2666-6510 (2021), 100–126. DOI: https://doi.org/10.1016/j.aiopen.2021.06.002

[25] Chongming Gao, Shiqi Wang, Shijun Li, Jiawei Chen, Xiangnan He, Wenqiang Lei, Biao Li, Yuan Zhang, and Peng Jiang. 2022. CIRS: Bursting filter bubbles by counterfactual interactive recommender system. *ACM Transactions on Information Systems* 42, 1 (2022), 1–27.

[26] Jianfeng Gao, Michel Galley, and Lihong Li. 2019. Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval* 13, 2-3 (February 2019), 127–298. DOI: https://doi.org/10.1561/1500000074

[27] Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. Association for Computational Linguistics, 82–89. DOI: https://doi.org/10.18653/v1/W19-2310

[28] 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 *Proceedings of the Interspeech 2019*. 1891–1895. DOI: https://doi.org/10.21437/Interspeech.2019-3079

[29] 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, 1183–1192. DOI: https://doi.org/10.1145/3269206.3271802

[30] Ahmed Hassan. 2012. A semi-supervised approach to modeling web search satisfaction. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'12)*. Association for Computing Machinery, New York, NY, 275–284. DOI: https://doi.org/10.1145/2348283.2348323

[31] Trevor Hastie, Jerome H. Friedman, and Robert Tibshirani. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. DOI: https://doi.org/10.1007/978-0-387-21606-5

[32] Lishan Huang, Zheng Ye, Jinghui Qin, Liang Lin, and Xiaodan Liang. 2020. GRADE: Automatic graph-enhanced coherence metric for evaluating open-domain dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 9230–9240. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.742

[33] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A survey on conversational recommender systems. *ACM Computing Surveys* 54, 5, Article 105 (May 2021), 36 pages. DOI: https://doi.org/10.1145/3453154

[34] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446. DOI: https://doi.org/10.1145/582415.582418

[35] Jiepu Jiang, Ahmed Hassan Awadallah, Xiaolin Shi, and Ryen W. White. 2015. Understanding and predicting graded search satisfaction. In *Proceedings of the 8th ACM International Conference on Web Search and Data Mining* (Shanghai, China) *(WSDM'15)*. Association for Computing Machinery, New York, NY, 57–66. DOI: https://doi.org/10.1145/2684822.2685319

[36] Shaojie Jiang, Svitlana Vakulenko, and Maarten de Rijke. 2023. Weakly supervised turn-level engagingness evaluator for dialogues. In *Proceedings of the 2023 ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR)*. Association for Computing Machinery, New York, NY, 258–268. DOI: https://doi.org/10.1145/3576840.3578319

[37] Yucheng Jin, Li Chen, Wanling Cai, and Pearl Pu. 2021. Key qualities of conversational recommender systems: From users' perspective. In *Proceedings of the 9th International Conference on Human-Agent Interaction* (Virtual Event, Japan) *(HAI'21)*. Association for Computing Machinery, New York, NY, 93–102. DOI: https://doi.org/10.1145/3472307.3484164

[38] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A. Konstan, Loren Terveen, and F. Maxwell Harper. 2017. Understanding how people use natural language to ask for recommendations. In *Proceedings of the 11th ACM Conference on Recommender Systems* (Como, Italy) *(RecSys'17)*. Association for Computing Machinery, New York, NY, 229–237. DOI: https://doi.org/10.1145/3109859.3109873

[39] Diane Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval* 3, 1–2 (Jan. 2009), 1–224. DOI: https://doi.org/10.1561/1500000012

[40] 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) *(WSDM'14)*. Association for Computing Machinery, New York, NY, 193–202. DOI: https://doi.org/10.1145/2556195.2556220

[41] 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, 45–54. DOI: https://doi.org/10.1145/2911451.2911521

[42] 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 Conference on Human Information Interaction and Retrieval, CHIIR* (Carrboro, North Carolina, USA) *(CHIIR'16)*. Association for Computing Machinery, New York, NY, 121–130. `DOI`:https://doi.org/10.1145/2854946.2854961

[43] Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved dialogue evaluation with optimized questions and multi-turn comparisons. CoRR abs/1909.03087 (2019). arXiv:1909.03087 http://arxiv.org/abs/1909.03087

[44] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Proceedings of the Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. 9748–9758. Retrieved from https://proceedings.neurips.cc/paper/2018/hash/800de15c79c8d840f4e78d3af937d4d4-Abstract.html

[45] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. Retrieved from https://aclanthology.org/W04-1013

[46] 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, 2122–2132. `DOI`:https://doi.org/10.18653/v1/D16-1230

[47] 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*. Association for Computing Machinery, 1141–1150.

[48] Mengyang Liu, Yiqun Liu, Jiaxin Mao, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. "Satisfaction with failure" or "unsatisfied success": Investigating the relationship between search success and user satisfaction. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW'18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1533–1542. `DOI`:https://doi.org/10.1145/3178876.3186065

[49] Hongyu Lu, Weizhi Ma, Min Zhang, Maarten de Rijke, Yiqun Liu, and Shaoping Ma. 2021. Standing in your shoes: External assessments for personalized recommender systems. In *Proceedings of the SIGIR 2021: 44th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1523–1533.

[50] Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, 225–235. Retrieved from https://aclanthology.org/2020.sigdial-1.28

[51] 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. `DOI`:https://doi.org/10.18653/v1/2020.acl-main.64

[52] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (Philadelphia, Pennsylvania) *(ACL'02)*. Association for Computational Linguistics, 311–318. `DOI`:https://doi.org/10.3115/1073083.1073135

[53] Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR*. Association for Computing Machinery, 25–33. `DOI`:https://doi.org/10.1145/3295750.3298924

[54] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 353–360. `DOI`:https://doi.org/10.18653/v1/W19-5941

[55] 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. `DOI`:https://doi.org/10.1609/aaai.v34i05.6394

[56] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2011. Introduction to recommender systems handbook. In *Recommender Systems Handbook*. Springer, 1–35.

[57] Tefko Saracevic. 1996. Relevance reconsidered. In *Information Science: Integration in Perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science, Copenhagen (Denmark)*.

[58] Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the CMU Let's go bus information system. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, 3369–3373. Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/333_Paper.pdf

[59] 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, 1702–1723. DOI: https://doi.org/10.18653/v1/N19-1170

[60] 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* (Madrid, Spain) *(SIGIR'22)*. Association for Computing Machinery, New York, NY, 2018–2023. DOI: https://doi.org/10.1145/3477495.3531798

[61] 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, 1570–1583. DOI: https://doi.org/10.18653/v1/2021.naacl-main.124

[62] 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. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, 2499–2506. DOI: https://doi.org/10.1145/3404835.3463241

[63] Andrew H. Turpin and William Hersh. 2001. Why batch and user evaluations do not give the same results. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA). Association for Computing Machinery, 225–231. DOI: https://doi.org/10.1145/383952.383992

[64] 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:1801.03625. Retrieved from https://arxiv.org/abs/1801.03625

[65] 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, 271–280. DOI: https://doi.org/10.3115/976909.979652

[66] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *KDD'22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Aidong Zhang and Huzefa Rangwala (Eds.). ACM, 1929–1937. DOI: https://doi.org/10.1145/3534678.3539382

[67] Jason D. Williams and Steve Young. 2004. Characterizing task-oriented dialog using a simulated ASR chanel. In *Proceedings of the Interspeech 2004*. 185–188. DOI: https://doi.org/10.21437/Interspeech.2004-114

[68] Raymond E. Wright. 1995. Logistic regression. In *Reading and Understanding Multivariate Statistics*. L.G. Grimm and P. R. Yarnold (Eds.). American Psychological Association, 217–244.

[69] Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 5676–5689. DOI: https://doi.org/10.18653/v1/2021.acl-long.441

[70] Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021. D-Score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, 2329-9290 (2021), 2502–2516. DOI: https://doi.org/10.1109/TASLP.2021.3074012

[71] Shuo Zhang and Krisztian Balog. 2020. Evaluating conversational recommender systems via user simulation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, 1512–1520. DOI: https://doi.org/10.1145/3394486.3403202

[72] 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, 2204–2213. DOI: https://doi.org/10.18653/v1/P18-1205

[73] Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. CRSLab: An open-source toolkit for building conversational recommender system. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL 2021—System Demonstrations, Online, August 1-6, 2021*, Heng Ji, Jong C. Park, and Rui Xia (Eds.). Association for Computational Linguistics, 185–193. DOI: https://doi.org/10.18653/v1/2021.acl-demo.22

[74] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion. In *KDD'20: The 26th ACM SIGKDD*

Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020, Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (Eds.). ACM, 1006–1014. DOI: https://doi.org/10.1145/3394486.3403143

[75] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 4128–4139. DOI: https://doi.org/10.18653/v1/2020.coling-main.365

Received 20 January 2023; revised 23 June 2023; accepted 4 September 2023