Towards a Taxonomy of User Feedback Intents for Conversational Recommendations

Wanling Cai
Department of Computer Science
Hong Kong Baptist University
Hong Kong, China
cswlcai@comp.hkbu.edu.hk

Li Chen
Department of Computer Science
Hong Kong Baptist University
Hong Kong, China
lichen@comp.hkbu.edu.hk

ABSTRACT

Understanding users' feedback on recommendation in natural language is crucially important for assisting the system to refine its understanding of the user's preferences and provide more accurate recommendations in the subsequent interactions. In this paper, we report the results of an exploratory study on a human-human dialogue dataset centered around movie recommendations. In particular, we manually labeled a set of over 200 dialogues at the utterance level, and then conducted descriptive analysis on them from both seekers' and recommenders' perspectives. The results reveal not only seekers' feedback intents as well as the types of preferences they have expressed, but also the reactions of human recommenders that have finally led to successful recommendation. A taxonomy for feedback intents is established along with the results, which could be constructive for improving conversational recommender systems.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in interaction design**; *User models*; • **Information systems** → *Recommender systems*.

ACM RecSys 2019 Late-breaking Results, 16th-20th September 2019, Copenhagen, Denmark

KEYWORDS

Dialogue-based conversational recommender systems; user feedback; intent taxonomy

INTRODUCTION

In recent years, dialogue systems have become increasingly popular in our daily life, with applications in various domains such as education, healthcare, e-commerce, business, etc. They often mimic human-like behavior to converse with users for addressing their chit-chatting or information-seeking requirements [9]. Given that users often explicitly request recommendations when they communicate with a task-oriented dialogue system [9], more efforts have been put in integrating recommending approaches into the system, so called the *Dialogue-based Conversational Recommender System* (DCRS) [3]. However, most of existing systems have provided one-shot recommendation, with the focus on selecting most informative questions to ask users [3, 8]. The dialogue often ends when the system produces one or a list of recommendations to the user (see related work in the left bar). But in reality, users may not get the desired recommendation within a single turn, in which case it becomes important to allow users to freely provide their feedback on the recommendation, so that the system could help them to find the desired item in the subsequent interactions. Our work is actually motivated by the real dialogue that can occur between two persons [6]. For example, if a seeker does not like the recommended movies from the recommender, s/he can give feedback such as "*I don't like any of those movies, too much talking*" to refine her/his preferences. The user feedback issue has been studied in a broader area of recommender systems, such as *critiquing-based recommender systems* that elicit users' feedback in graphical user interfaces [2], but little work has been done on user feedback in natural language. Since the language-based feedback can be in diverse, free styles, it is meaningful to investigate how users express it (e.g., what intents they may have and what kinds of preferences they want to convey), which should be constructive for developing more dedicated preference elicitation and intent prediction strategies for DCRS. Therefore, we manually labeled a set of human-human dialogues (over 200) centered around movie recommendations [6] with our established taxonomy for user feedback intents, and analyzed the utterances starting from the point when a seeker did not like one recommendation till s/he accepted another one. The results analysis reveals not only the seeker's intents and preferences, but also the human recommender's responses that eventually helped the seeker find a satisfactory item. It is hence inspiring for boosting the human-like aspect of current dialogue systems.

Related Work

Current dialogue-based conversational recommender systems (DCRS) have mostly focused on question generation and selection before giving the recommendation. For instance, an active learning and bandit learning based conversational framework was proposed in [3], which aims to adjust question selection strategy in real time. [8] trained a deep policy network to decide when the system should conduct facet preference elicitation. However, little work on DCRS has explicitly studied the feedback issue that occurs when the user is not satisfied with the current recommendation. In the broader area, *critiquing-based recommender systems* [2] have been developed to elicit users' feedback in graphical user interfaces (GUI), for which several major types of critiquing are supported such as user-initiated critiquing and system-suggested critiques. However, this kind of system limits the way users can post their feedback since their interactions are restricted to traditional GUI elements (e.g., menu, form, button). The advantage of dialogue systems is that the interaction is not limited to a pre-defined procedure or a fixed set of attributes. But to the best of our knowledge, few studies have investigated users' goals, intents, and ways of expressing preferences when they interact with DCRS [4], not to mention their feedback on recommendations.

DIALOGUE-BASED RECOMMENDATION DATASET

In this section, we present our data selection, taxonomy definition, and data annotation procedure.

Data Selection

The original ReDail[1] dataset contains 11,348 human-human dialogues [6]. We first filtered out dialogues with less than 3 conversation turns[2]. We also removed those with inconsistent answers from seekers and recommenders regarding the post-conversation reflective questions, because it may be due to carelessness or dialogue ambiguity [6]. We then selected the dialogues containing at least two movies suggested by the recommender, among which one was not liked by the seeker while another subsequent recommendation was accepted by her/him. This process was mainly to capture the seeker's feedback on recommendation in case s/he was not satisfied with it, as well as how the human recommender responded to the seeker and helped her/him to find a satisfactory item later. As a result, we got 225 dialogues (see Table 1 with the statistics of our selected dialogue data).

[1]https://redialdata.github.io/website/

[2]One conversation turn denotes a consecutive utterance-response pair: Utterance is from seeker and response is from recommender.

Table 1: Statistics of our selected dialogue data (from ReDail [6])

Items	Dialogue data
# Conversations	225
# Human seekers	111 (# utterances: 1,537)
# Human recommenders	134 (# utterances: 1,565)
# Movies suggested	1,096
# Turns per dialogue	mean=6.64, min=3, max=19
# Words per utterance	mean=10.83, min=1, max=72

Taxonomy for User Feedback Intents

Based on literature survey, we first established an initial taxonomy to classify user feedback on recommendations, which basically covers all of the feedback types, such as the three types of feedback modality (i.e., similarity-based, quality-based, quantity-based) in critiquing-based recommender systems [2], the session-aware intents (i.e., add filter condition, see-more, negation) in task-oriented dialogue systems [9], and the follow-up query strategies (i.e., refine, reformulate, and start over) when users ask for recommendations [4]. Then, we refined the taxonomy by applying the open coding and theme identification approaches [7] to our dialogue data. Four new categories (i.e., *Inquire*, *Seen*, *Provide Details*, *Ask*) were added through keywords-in-context method; and some existing categories were modified or merged into seven categories (i.e., *Reject*, *Critique-Add*, *Critique-Compare*, *Critique-Feature*, *Restate*, *Restate with Further Constraints*, *Restate with Clarification*) based on the real dialogues through constant comparison method. We went through the standard classification procedure (i.e., *propose-annotate-refine*) three times, and finally came up with the taxonomy for user feedback intents (see Table 2). The data annotation work is shown in the left bar.

Data Annotation

Two annotators were involved into the labeling work. They were instructed to carefully read the taxonomy table before they started. For each utterance, the annotator was encouraged to choose all labels that s/he thinks can represent the seeker's intents. They first independently labeled 143 random dialogues. The inter-rater agreement across their intent labels is 0.87 (through Fuzzy Kappa [5]), which indicates satisfactory annotation quality and consistency. They then labeled the remaining dialogues, and met to discuss and resolve disagreements.

DATA ANALYSIS & RESULTS

Seeker Feedback Intents and Preference Expression

Feedback Intent Distribution. The feedback intent distribution is shown in Table 2, where we can see *Reject*, *Seen*, *Critique-Feature*, *Provide Details*, and *Inquire* more frequently occur than others, which suggests that the seeker may tend to explicitly express her/his negative opinions on a recommendation, and attempt to explain why s/he dislikes it as well as providing more preference info to the recommender. Relatively, some seekers are also inclined to critique the recommendation by adding further constraints, or start a new query if they feel it is difficult to receive a satisfactory result with the current query.

Table 2: A taxonomy for user feedback intents during the interaction with a dialogue-based recommender, and intent distribution in our dataset

User Feedback Intent (Code)	Description	Example	Percentage
Reject (REJ)	Seeker dislikes the recommended item.	"*I hated that movie. I did not even crack a smile once.*"	19.2%
Seen (SEE)	Seeker has seen the recommended item before.	"*I have seen that one and enjoyed it.*"	16.3%
Critique-Feature (CRI-F)	Seeker makes critique on specific features of the current recommendation.	"*That's a bit too scary for me.*"	11.8%
Provide Details (PRO)	Seeker provides detailed preferences for the item s/he is looking for.	"*I usually enjoy movies with Seth Rogen and Jonah Hill.*"	11.7%
Inquire (INQ)	Seeker wants to know more about the recommended item.	"*I haven't seen that one yet. What's it about?*"	10.9%
Critique-Add (CRI-A)	Seeker adds further constraints on top of the current recommendation.	"*I would like something more recent.*"	8.5%
Start Over (STO)	Seeker starts a new query.	"*Anything that I can watch with my kids under 10.*"	5.2%
Neutral Response (NRE)	Seeker does not indicate her/his preferences for the current recommendation.	"*I have actually never seen that one.*"	5.1%
Critique-Compare (CRI-C)	Seeker requests something similar to the current recommendation.	"*Den of Thieves (2018) sounds amazing. Any others like that?*"	2.9%
Answer (ANS)	Seeker answers the question issued by the recommender.	"*Maybe something with more action.*" (Q: "*What kind of fun movie you look for?*")	2.8%
Ask (ASK)	Seeker asks the recommender's personal opinions.	"*I really like Reese Witherspoon. How about you?*"	1.6%
Restate with Further Constraints (RES-CO)	Seeker restates her/his query with further constraints.	"*Do you have something that is a thriller but not too scary?*"	1.6%
Restate (RES)	Seeker completely restates her/his query.	"*Maybe I am not being clear. I want something that is in the theater now.*"	1.5%
Restate with Clarification (RES-CL)	Seeker restates her/his query with clarification.	"*I'm fine with any sort of horrors, jump scares, clowns, etc.*"	0.4%
Others (OTH)	The utterance cannot be categorized into any other categories.	"*Sorry about the weird typing.*"	0.4%

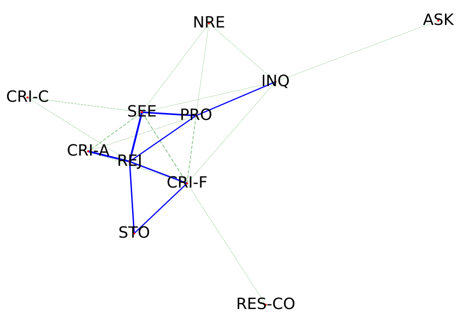

Figure 1: Seeker feedback intent co-occurrence. The edge with larger weight (co-occurrence frequency greater than 10) is in solid line and that with lower weight in dashed line.

[3] *Objective* refers to the user's criteria on the item's attributes (e.g., movie genre), *subjective* involves the user's emotional or opinionate preferences (e.g., "happy movie"), and *navigational* is in relation to a movie the user refers to (e.g., "Star Wars movies").

Intent Co-occurrence. We find 40.5% of utterances contain more than one intent label. The undirected graph of feedback intent co-occurrence weighted by the co-occurrence frequency is shown in Figure 1. It can be seen that *Reject* often co-occurs with *Critique-Feature*, *Critique-Add*, *Seen*, *Provide Details*, and *Start Over*, which may explain the reasons why some seekers reject a recommendation, i.e., because it does not satisfy their preferences for some specific features, miss values on some constraints they have not stated, or it was already seen by the seeker. Besides, rather than critiquing the current recommendation, some seekers try to provide more detailed preferences, or start a new query when they reject an item.

Preference Expression. We then analyzed how seekers actually express their preferences in the feedback, which is inspired by [4] that classifies user queries into three-level goals: *objective*, *subjective*, and *navigational*[3]. We refined this classification scheme by linking them to the concepts that the seeker may mention [1]: *Entity* (like a movie or a series of movies that can be with subjective or navigational goal), *attribute* (with objective or subjective goal), and *purpose* (the general uses of the item, e.g., "*Anything that I can watch with my kids under 10?*"). The results show that seekers more frequently express their preferences at the attribute level, which is much more often than the mentions of entity and purpose concepts (see Figure 2). Moreover, they like to express subjective opinions on entity when they mention it, but have more objective criteria for attributes (slightly higher than the proportion of attribute-level subjective goals).

Recommender Actions

From the human recommender's perspective, we investigated what actions s/he may carry out in response to the seeker's feedback. We first identified five major types of actions (see Table 3), and then

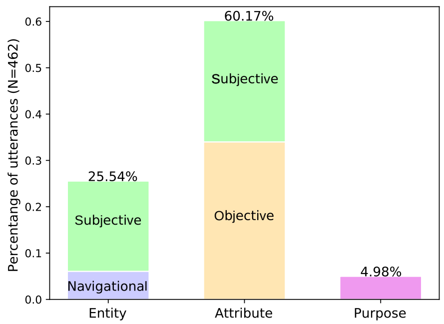

Figure 2: Seeker preference expression on the three concepts respectively: entity, attribute, and purpose. Note that our dataset consists of 462 utterances which start from the point when a seeker did not like one recommendation till s/he accepted another one.

Table 3: Recommender reactions to seeker feedback, and action distribution in our dataset

Action (Code)	Description	Percentage
Recommend (REC)	Recommender provides one or more recommendations.	43.8%
Explain (EXP)	Recommender explains why the item is recommended.	30.0%
Respond (RES)	Recommender responds to any other queries by the seeker.	12.4%
Answer (ANS)	Recommender answers the question from the seeker.	10.2%
Request (REQ)	Recommender requests for the seeker's preferences.	3.1%

asked annotators to label all recommenders' responses. From Table 3, we can see that, in nearly half of the cases, the recommender tends to recommend one or more other items when the seeker rejects the current one. In the other cases, the recommender tries to explain why the new recommendation would be good to the seeker, respond to the seeker's requests, answer the seeker's explicit question, or ask for the seeker's preferences.

FUTURE WORK

In this work, we established a taxonomy for user feedback intents and analyzed a set of human-human dialogues centered around movie recommendations. As the next step, we plan to label more dialogues to further validate the taxonomy. We also want to perform temporal analysis so as to reveal the frequent conversation patterns that may occur between seekers and recommenders. Based on the findings from our analysis, we intend to develop a dedicated user intent prediction model to predict users' intents given their utterances, which is believed as an important component that could help DCRS to track users' current states, refine their preference model, and then select an appropoiate action to respond to users.

REFERENCES

[1] Joyce Yue Chai, Malgorzata Budzikowska, Veronika Horvath, Nicolas Nicolov, Nanda Kambhatla, and Wlodek Zadrozny. 2001. Natural Language Sales Assistant - A Web-Based Dialog System for Online Sales. In *Proceedings of the Thirteenth Conference on Innovative Applications of Artificial Intelligence Conference*. 19–26.

[2] Li Chen and Pearl Pu. 2012. Critiquing-based Recommenders: Survey and Emerging Trends. *User Modeling and User-Adapted Interaction* 22, 1-2 (April 2012), 125–150. https://doi.org/10.1007/s11257-011-9108-6

[3] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 815–824. https://doi.org/10.1145/2939672.2939746

[4] 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 Eleventh ACM Conference on Recommender Systems (RecSys '17)*. 229–237. https://doi.org/10.1145/3109859.3109873

[5] Andrei P. Kirilenko and Svetlana Stepchenkova. 2016. Inter-Coder Agreement in One-to-Many Classification: Fuzzy Kappa. *PLOS ONE* 11, 3 (03 2016), 1–14. https://doi.org/10.1371/journal.pone.0149787

[6] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems 32*. 9748–9758.

[7] Gery W Ryan and H Russell Bernard. 2003. Techniques to Identify Themes. *Field Methods* 15, 1 (2003), 85–109.

[8] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 235–244. https://doi.org/10.1145/3209978.3210002

[9] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building Task-oriented Dialogue Systems for Online Shopping. In *Thirty-First AAAI Conference on Artificial Intelligence*.