A Model of Social Explanations for a Conversational Movie Recommendation System

Florian Pecune
florian.pecune@glasgow.ac.uk
University of Glasgow

Shruti Murali
shrutim@andrew.cmu.edu
Carnegie Mellon University

Vivian Tsai
viv@jhu.edu
Johns Hopkins University

Yoichi Matsuyama
matsuyama@pcl.cs.waseda.ac.jp
Waseda University

Justine Cassell
justine@cs.cmu.edu
Carnegie Mellon University

ABSTRACT

A critical aspect of any recommendation process is explaining the reasoning behind each recommendation. These explanations can not only improve users' experiences, but also change their perception of the recommendation quality. This work describes our human-centered design for our conversational movie recommendation agent, which explains its decisions as humans would. After exploring and analyzing a corpus of dyadic interactions, we developed a computational model of explanations. We then incorporated this model in the architecture of a conversational agent and evaluated the resulting system via a user experiment. Our results show that social explanations can improve the perceived quality of both the system and the interaction, regardless of the intrinsic quality of the recommendations.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**.

KEYWORDS

conversational recommendation system; explanation; socially-aware

ACM Reference Format:
Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A Model of Social Explanations for a Conversational Movie Recommendation System. In *Proceedings of the 7th International Conference on Human-Agent Interaction (HAI '19), October 6–10, 2019, Kyoto, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3349537.3351899

1 INTRODUCTION

People constantly seek recommendations when engaging in day-to-day activities, from eating at a restaurant to planning a vacation itinerary. Over the last decades, large effort has been put into optimizing different algorithms to deliver the most accurate–or relevant–recommendations [24, 36]. Researchers are also investigating *how* users interact with recommendation systems, as this significantly impacts overall experiences. In one of the first attempts to formalize the conversation between a recommendation system and its user, Carenini et al. provide a list of techniques that a recommendation system can use to query information or user feedback and describe the effect of these techniques on the system's accuracy and user's effort [4]. As people grow increasingly familiar with conversational agents, the scope of these conversational techniques has been extended, and the interactions between conversational recommendation agents and their users have consequently become more complex [1, 15].

Beyond the actual recommendation itself, several factors influence the user's perception of the system and overall experience. When presenting recommendations, both the modality (e.g., text vs. image) and the organization (single item vs. list of items) play a role [19]. Additionally, minimizing latencies within the system improves users' perceived quality of the recommendation [34]. The faster a recommendation is delivered, the more relevant it is perceived.

Another influential factor is the explanation the system gives to support its recommendations [32]. By revealing the reasoning behind a specific recommendation, a system can increase users' trust in the system; convince users to try or buy an item; or help users reach decisions faster. All of these aims can be interrelated, e.g., an explanation that increases transparency by clarifying how a recommendation was chosen could also increase users' trust in the system. Recent work attempts to classify the types of explanations found in a recommendation context, emphasizing the growing need to endow recommendation systems with the right explanation model [21, 22, 37].

In this paper, we present a conversational recommendation system that draws from the various explanations humans use with one another, and we describe the human-centered design we relied on to build such a system. First, we explored, annotated, and analyzed an existing corpus of interactions between two people discussing movies via phone. We specifically focused on the explanations and descriptions people gave as they depicted movies to their interlocutors. Next, we built a conversational agent that explains movie recommendations through the different social strategies we identified during our analyses. Finally, we evaluated our conversational agent through an experiment with real users. The main contributions of this work are thus (1) a model of social explanations, driven by our analyses of human-human interactions, for a conversational recommendation system, and (2) a subjective evaluation investigating the influence of our model on the perceived quality of both a conversational recommendation agent and the overall interaction.

HAI '19, October 6–10, 2019, Kyoto, Japan

ACM ISBN 978-1-4503-6922-0/19/10…$15.00
https://doi.org/10.1145/3349537.3351899

2 RELATED WORK

Many existing recommendation systems already generate explanations, and several attempts have been made to classify these explanations [21, 22, 37]. In *item-based explanations*, the system relies on the previous recommendation's outcome to justify the current recommendation: "I have recommended X because you previously liked/bought Y." *Feature-based explanations* use preferences that were specified by the user during the preference-elicitation process: "Your interest in Z suggests that you would like X." These two different types can be combined: an example of item and feature based explanations can be found in [25], where the system displays the features of previously liked movies to justify the current recommendation. In the domain of movie recommendations, a system can justify its decision by emphasizing a plot similarity [25] or an overlapping cast [30]. An evaluation comparing feature-based explanations, item-based explanations, and a combination of both shows that the hybrid explanation type was significantly more appreciated by users [30].

Both item-based and feature-based explanations are machine-centered and thus essentially reveal the system's decision-making process. Although they have a great impact on transparency, these explanations are tightly coupled with the types of features that the recommendation engine is relying on and may lack the persuasiveness and richness that humans often express when they recommend a specific item. Another important question regarding feature-based explanations is whether they should be personalized to match users' preferences. Research indicates that while personalization generally increases satisfaction, it can be detrimental to effective decision-making [33]. This shows how effectiveness and satisfaction aims can be discordant.

Human-based explanations take an alternative approach; here, the system relies on collaborative filtering to reference similar products: "People who liked X also liked Y." One such example is [13], in which the system recommends social software items such as social groups or communities and justifies its choice by showing the names of people in the group/community, as well as their relationship to the user. This relation could be "familiar" if the user was friends with the person, or "similar" if both shared similar interests. The authors' experiment shows that when these people were "familiar," users were more satisfied with the recommendations.

Human-based explanations can be merged with feature-based explanations by combining existing reviews with users' preferences [7, 14] to generate explanations: "You might want to watch X because Bob says that the storyline is amazing and I know that you are highly interested in plot. Here is his review: (...)." This approach thus uses third-party opinions to justify choices. However, reviews are sometimes extremely long, making them difficult to integrate when conversing with a user.

As recently demonstrated by [17], researchers would benefit from taking a more human-centered approach for the design of their recommendation systems, i.e., building systems able to express their "own" opinions. The authors' recommendation system, which used social conversational strategies such as self-disclosures and reciprocity in its recommendation process, significantly increased users' satisfaction and intention to seek future recommendations.

In this paper, we aim to build a conversational recommendation system that recommends movies by expressing its "own" opinions and experience through social explanations. We thus focus on the following research questions:

RQ-1: What are the types of social conversational strategies that humans use when they describe a movie they watched to someone?

RQ-2: Do social explanations used by a conversational recommendation agent to justify its recommendations influence the perceived quality of both the recommendations and the interaction?

3 MODEL OF EXPLANATIONS

To answer **RQ-1** and analyze the different strategies humans use when discussing movies, our first step was to find a corpus appropriate for our eventual goal of a conversational agent, i.e., a corpus of human-human spontaneous dyadic discussions about movies [27]. The Switchboard corpus [10] contains 2430 spontaneous dyadic conversations held over the phone. Around 500 American English speakers of both genders were recorded, totaling 240 hours of speech. Each speaker was prompted with a specific topic (out of 70 different ones) to discuss and instructed to hang up whenever they wanted. The maximum duration of the interactions was 10 minutes.

3.1 Data Annotation

For our purpose, we focused on the subpart of the Switchboard corpus that was dedicated to movie discussions; here, each user was prompted to "find out what the other caller thought about the last few movies they saw" and to talk about the movies they've seen lately. This subpart is comprised of 32 interactions (64 speakers). To annotate them, we isolated chunks of dialogue that pertained to a specific movie (totaling over 250 movie chunks), then classified the explanations within these movie chunks along different categories and subcategories. We developed a coding manual detailing each subcategory and used this manual to train two different annotators.

The annotators were asked to separately annotate the same subset of the corpus (25 movie chunks), labeling each utterance of the chunks with at most one of the subcategories from the coding manual. After each annotation round, we calculated Inter-Rater Reliability scores. If the agreement was too low (Cohen's $\kappa < 0.7$), we discussed agreements and disagreements with the two annotators, adapted our coding manual accordingly, and underwent another round of annotation using 25 new movie chunks. Once we reached acceptable agreement (Cohen's $\kappa > 0.7$) for each of the subcategories, each annotator labeled the utterances for one half of the entire corpus (126 movie chunks per annotator).

3.2 Explanation Categories

In the end, we clustered the annotated explanations into four categories (see Table 1 for examples):

Movie Features. These are similar to the feature-based explanations in [22]. The five subcategories are: the *cast* of the movie (either

Category	Subcategory	Example
Movie Feature	Cast [MF_C] *(4%)*	Julia Roberts has the lead role in the movie.
	Genre [MF_G] *(9%)*	It's sort of a suspense and thriller movie.
	Plot [MF_P] *(17%)*	It's about this boy who is left alone in his house and he sets traps for robbers trying to break in.
	Award [MF_A] *(2%)*	It was last year's number one movie at the box office.
	Other [MF_O] *(5%)*	It's a sequel.
Third-Party Opinion	Broad [TPO_B] *(4%)*	A lot of people said it was really good.
	Specific [TPO_S] *(3%)*	My friend watched that movie last night and said it was hilarious!
Personal Opinion	Positive [PO_POS] *(20%)*	I was really impressed with it.
	Analytic [PO_ANA] *(12%)*	I think it says a lot about the effects of technology on relationships.
	Structured [PO_SO] *(7%)*	The plot was pretty bad in my opinion, but the actors had a great chemistry.
Personal Experience	Anecdote [PE_A] *(3%)*	My friend and I kept talking during the movie because we couldn't believe what was happening.
	Logistics [PE_L] *(9%)*	I watched the Blu-Ray with my colleagues.
	Comparison [PE_C] *(5%)*	That movie reminds me of *The Dead Poets Society*. They have the same themes.

Table 1: Categories and subcategories of social explanations with their associated probability to follow a recommendation.

actor or director), the *genres*, the *plot*, the *awards* won so far, and *other* features (e.g., duration, location).

Third-Party Opinions. These are similar to the human-based explanations in [22] and can be either *broad* ("People said that (...)") or *specific* ("My uncle Bob told me that (...)").

Personal Opinions. We found three different subcategories: *positive* (i.e., positively valenced opinions); *analytic*, which regroups neutral interpretations such as, "I think they could have made the movie more realistic."; and *structured*, which represents explanations expressing more than one valenced opinion ("The plot was very bad, but the actors had a great chemistry").

Personal Experience. These experiences can be a specific *comparison* with another movie, similar to item-based explanations [22]; an *anecdote* linked to the movie; or a *logistics* explanation, which is formed based on "how" people watched the movie ("I rented it" or "I watched that one with a friend in theaters").

3.3 Discussion

When discussing favored movies, people tend to speak personally; our analyses show that more than half of the explanations involved either personal opinions (39%) or personal experiences (17%). Although humans commonly search for online reviews when independently seeking recommendations, speakers in the corpus rarely used third-party opinions to justify their choices (only 7% of the overall explanations). The positive opinions were either broad (e.g., "It was a fun movie to watch") or targeted specific elements (e.g., "The characters played well off one another"). People used two-sided strategies such as *structured opinion* to highlight that both negative and positive aspects of the movie were considered. Such a technique is known to improve persuasiveness [26].

The movie features *cast* and *genre*, as well as *comparisons* were used in two different ways. People used them as transitions between two movies, initiating a recommendation via a feature (e.g., "Speaking of violent movies, I'd recommend *Pulp Fiction* (...)"). Using similarities as a means of transitioning ensures dialogue coherence, which is consistent with the work on topic shifting presented in [29]. People also used movie features in a negative way to contrast with their positive opinion about a specific movie (e.g., "I don't like Stephen King, but *Misery* is an excellent movie!"). As described in [3], negative framing makes arguments more convincing.

Logistics explanations and *anecdotes* were almost always paired with one or more personal opinions. As described in [20], interactional remembering signals the transition from a balanced turn-taking mode to a more narrative, one-sided mode.

4 CONVERSATIONAL AGENT ARCHITECTURE

We next investigated how a conversational agent that justifies its recommendations via social strategies would be perceived by its users. To do so, we built and deployed the conversational agent depicted in Fig.1. We detail below the different components of our architecture and describe the typical communication flow that occurs during one interaction turn.

Front-End. Our user interface runs on Unity Web Player. SARA, our animated virtual character, is displayed on the right of the interface, with voice generated by the Chrome Text-To-Speech (TTS) plugin. When our system recommends a movie, a corresponding poster is displayed on the left of the screen. To talk to our agent, users push a button displayed at the bottom of the interface. This push-to-talk button has three states: *available* (the user can push it to start talking), *processing* (the button has been pushed and the system is processing the user's speech), and *busy* (the user cannot press the button because the agent is not done speaking). Speech is processed by the Chrome Speech-to-Text (ASR) plugin and the textual transcriptions are then sent to our Multiuser Framework, a middleware in charge of handling simultaneous sessions and managing the communication between our backend modules.

Natural Language Understanding. The first component triggered is the natural language understanding (NLU) module, which extracts intentions and entities from users' utterances. Our NLU module leverages two different NLP libraries. Sempre [2] allows

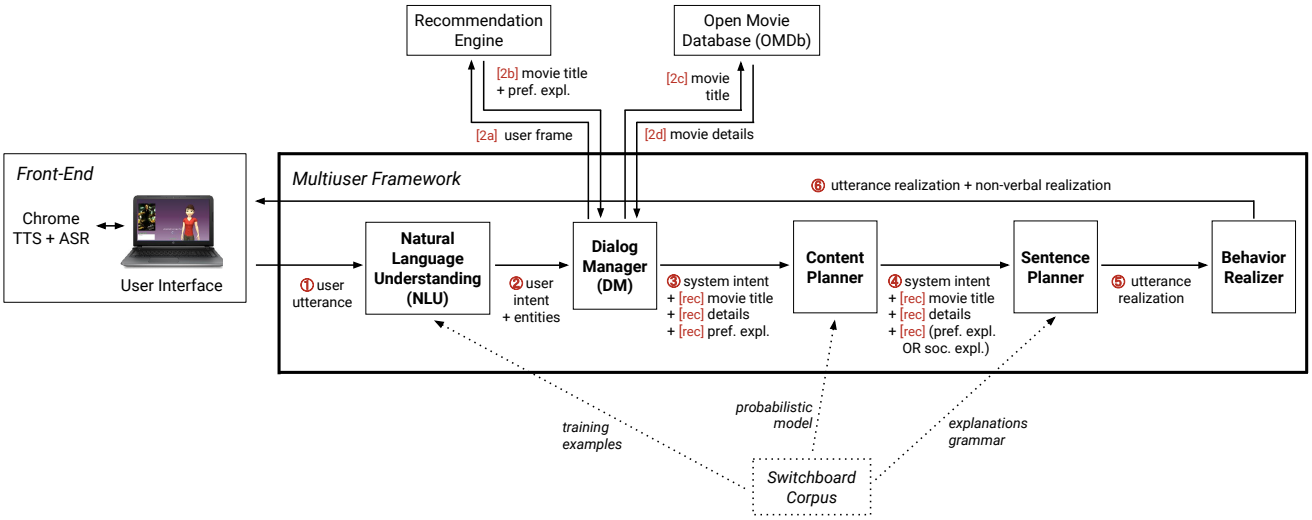

Figure 1: Architecture of our conversational recommendation agent, with recommendation-only items preceded by brackets.

us to classify utterances among eleven different intentions, using a model that was trained with examples from the Switchboard dataset, as well as datasets from previous experiments. The Stanford CoreNLP library [18] allows us to parse entities; we developed a fuzzy matching technique to map these named entities to a list of actors/directors/genres extracted from the Internet Movie Database (IMDb) [1]. For example, our NLU's output for the sentence, "I kind of like Tom Cruise" is `Inform(actor="tom$\_$cruise")`.

Dialog Management. Our dialog manager (DM) is designed as a finite state machine that takes the user intent and entity from the NLU as inputs; it then uses these to transition to each new state based on the current state of the dialog and a set of rules (see section 5.1 for an overview of the scenario). The DM stores the user's recognized entities in the *user frame*, which also stores two different lists of movies that were liked and disliked, respectively, by the user during the interaction.

Whenever our agent has to deliver a recommendation to its user, the DM queries the recommendation engine for a movie matching the information stored in the user frame (preferred actors/directors/genres, liked/disliked movies). The movie recommendation engine [6] then generates a relevant movie using a knowledge graph and personalized PageRank algorithm. In addition to the title of the movie, the recommendation engine returns a preference-based explanation in the form of the particular entity with the biggest weight in the final decision (e.g., "Tom Cruise" for the movie *Oblivion*). The DM uses the movie title to query the Open Movie Database (OMDb) API[2] for further details (e.g., synopsis, movie duration, awards, etc.).

Content Planning. All information obtained by the DM (i.e., system intent, movie title, movie explanation, movie details) is sent to our content planner, which selects the type of explanation (if any) to use when recommending the movie. Based on the experimental condition (see section 5.1), the content planner first decides whether to generate an explanation or to simply provide the movie title. If the former is true, it next decides whether to use the feature-based explanation generated by the recommendation engine or to pick a social explanation from our computational model.

For all social explanations, the content planner relies on a probabilistic model to select the specific subcategory (see Table 1). For instance, our agent has a 20% of chance of generating a positive personal opinion [PO_POS] ($p(PO\_POS) = .2$), but will generate a comparison [PE_C] only 5% of the time ($p(PE\_C) = .05$). The content planner also uses the model to determine a second, subsequent explanation, based on conditional probabilities from our annotated corpus (e.g., there is a $p(PO\_POS|PE\_L) = .01 = 1\%$ chance that a logistics explanation [PE_L] will be followed by a positive opinion [PO_POS]).

Response Generation. Once explanation types are selected, our sentence planner constructs a valid utterance expressing all previously determined details, as well the relationship between them. We followed the approach presented in [28] to generate explanations; each subcategory of explanations (see Table 1) was associated with a distinct context-free grammar developed from the annotated Switchboard examples. Examples of generated sentences can be found in Table 2. The final module to be triggered is BEAT [5], a nonverbal behavior realizer which adds and synchronizes nonverbal behavior with the utterance to generate a behavior plan in the Behavior Markup Language (BML) form [16]. The BML file is then sent to the front-end to animate our virtual character.

5 EXPERIMENT

To answer **RQ-2**, we designed an experiment investigating how the type of explanations delivered by our agent influenced the perceived quality of both the system and the interaction. We recruited

[1] https://www.imdb.com/
[2] https://www.omdbapi.com/

Dialog Phase	Intent name	Sentence
Introduction	*greeting()*	*Hi! Are you doing well today?*
	yes()	I am doing good how are you
	intro_task()	*I'm here to help you find a great film to watch. Are you in the mood for one?*
	yes()	always
	request(last_movie)	*I'm sure you have great tastes. What was the last movie you watched?*
	inform(last_movie)	acrimony
	request(opinion)	*What did you like about the film?*
	inform(opinion)	it was very dramatic
Interview	*request(genre)*	*Before we move on I'd like to know more about your preferences. Would you mind telling me your favorite genre?*
	inform(genre=comedy)	comedy
	request(director)	*Do you have a favorite director? I've been watching a lot of Steven Spielberg movies lately.*
	inform(director=wes anderson)	Wes Anderson
	request(actor)	*Last but not least. Tell me about your favorite actor.*
	no()	I don't really have a favorite actor
Delivery	*inform(movie=Goats (2012))*	*How about Goats (2012)? I really enjoyed that one **[PO_POS]**.*
	request(actors)	who's in it
	inform(actors=David Duchovny Vera Farmiga ...)	*The leading actors are David Duchovny, Vera Farmiga, Graham Phillips*
	yes()	that sounds interesting
	request(another)	*Would you like another recommendation?*
	yes()	yes I would
	inform(movie=Barbershop: The Next Cut (2016))	*How about Barbershop: The Next Cut (2016)? My friends and I saw it in theaters **[PE_L]**. It had a thoughtful message **[PO_ANA]**.*
	inform(watched=true)	I've actually seen this movie before and it's very good
	request(opinion)	*I'm curious to hear what you thought of it!*
	inform(opinion)	it was very entertaining and funny
	request(another)	*Would you like another recommendation?*
	yes()	yes I would
	inform(movie=L!fe Happens (2011))	*How about L!fe Happens (2011)? That's a comedy **[MF_G]**. Have you seen it before?*
	inform(watched=false)	I have not
	inform(plot)	*Here is the plot: A comedy centered on two best friends, Kim and Deena, who fight to maintain (...)*
	yes()	that sounds very interesting
	request(another)	*Would you like another recommendation?*
	no()	no I'm good thank you
	bye()	*Feel free to let me know if you ever need a movie recommendation. See you!*

Table 2: Textual transcription of a conversation between a user (romanized) and the version of our agent *(italicized)* implementing our model of social explanations. The subcategories of explanations that our system used during this interaction are in brackets.

participants through Amazon Mechanical Turk; to ensure the quality of the data collected, all participants had at least a 95% approval rating with more than 1,000 previous HITs validated.

Participants were first presented with a consent form informing them about the conditions of the experiment. Those who agreed to partake in the study were then presented with a short description of the task and information on how to interact with our conversational agent, SARA; they were then introduced to the scenario, which provided the context of the interaction (i.e., the participant is hosting a movie night with friends and looking for a movie to watch). Each participant was randomly assigned to a group according to the different independent variables and interacted with SARA following the scenario in section 5.1. Once they finished their interactions, participants took two surveys that evaluated the quality of the conversational agent and the quality of the interaction, respectively. The maximum task duration was set to 20 minutes. Participants received a compensation of $0.80 after they completed the surveys.

5.1 Stimuli

The interaction scenario is designed to follow the traditional interview/delivery structure proposed by [35] with an additional introductory phase (see Table 2). In this *introductory* phase, our agent first greets the user before introducing itself. It then asks the user for the last movie they watched and their opinion of it. The *interview* phase is comprised of a sequence of three questions that gather relevant preferences: the user's preferred genre, director, and actor (always in that order). In the *delivery* phase, the agent recommends a movie along with any explanations that the content planner may have selected. From there, one of the following occurs:

(1) If the user accepts the recommendation, the system updates the user frame accordingly and asks whether the user would like another movie title. If the user declines, the agent says goodbye and the interaction ends.
(2) If the user rejects the recommendation, the agent updates the user frame accordingly, then requests the reason behind the rejection before asking whether the user would like another movie title.
(3) If the user has already watched the movie, the agent requests their opinion of it before asking whether the user would like another movie title.
(4) If the user requests additional information (e.g., the movie's cast, genre, or plot), the system provides that information accordingly.

We identified two between-subject independent variables. The first is the recommendation type, **Rec-Type**, which has two levels: random recommendations (*rand-rec*), where the system does not store any of the user's preferences and thus delivers random recommendations; and personalized recommendations (*pers-rec*), where the personal assistant delivers tailored recommendations based on the stored user preferences. We sought to investigate how participants would react to our model of social explanations regardless of the intrinsic quality of the recommendation itself. The second between-subject variable is the type of explanations, **Expl-Type**, which is how the agent justifies each of its recommendations.

This has three levels: the no-explanation condition (*no-expl*); the preference-based explanations (*pref-expl*) condition, where the system uses the feature-based explanation obtained by the DM; and the social explanations condition (*soc-expl*), where the system uses the social explanation selected by the content planner.

5.2 Measurements

Participants were asked to answer two different questionnaires for this experiment: one measuring perceived quality of the conversational agent and the other measuring perceived quality of the overall interaction. For the former, we adapted the questionnaire used in [8]; this encompasses multiple aspects of a recommendation system's task performance and thus helped us analyze the potential trade-offs between the various independent variable conditions. The eight different items we used to measure task performance are listed in Table 3.

For the quality of the interaction, we relied on *rapport* [31], a notion commonly used in the domain of human-agent interactions to evaluate whether people are in sync with the system they are interacting with [11, 39]. The eight different items we used to measure rapport are listed in Table 4.

All answers for both questionnaires were on a 5-point Likert scale (anchors: 1 = completely disagree, 5 = completely agree).

5.3 Hypotheses

We hypothesized the following:

- **H1-a**: The type of recommendation (**Rec-Type**) delivered by the conversational agent will have a main effect on the agent's perceived quality. More specifically, the quality of the agent when delivering random recommendations (*rand-rec*) will be perceived as lower than the quality when delivering personalized recommendations (*pers-rec*).
- **H1-b**: The type of explanations (**Expl-Type**) used by the conversational agent will have a main effect on the agent's perceived quality. More specifically, the quality of the agent when using social explanations (*soc-expl*) will be perceived as higher than the quality when using preference-based explanations (*pref-expl*), which will in turn be perceived as higher than the quality when using no explanations at all (*no-expl*).
- **H2-a**: The type of recommendation (**Rec-Type**) delivered by the conversational agent will have a main effect on the perceived quality of the interaction. More specifically, the interactions with random recommendations (*rand-rec*) will be perceived as worse than the interactions with personalized recommendations (*pers-rec*).
- **H2-b**: The type of explanations (**Expl-Type**) used by the conversational agent will have a main effect on the perceived quality of the interaction. More specifically, interactions with social explanations (*soc-expl*) will be perceived as better than interactions with preference-based explanations (*pref-expl*), which will in turn be perceived as better than the interactions with no explanations at all (*no-expl*).

6 RESULTS

We collected 60 interactions, with 10 per condition. Our conversational agent recommended 140 movies in total, with an average of 2.33 recommendations per interaction (std = 1.5).

6.1 Quality of the conversational agent

We conducted a 2x3 factorial MANOVA (i.e., multivariate analysis of variance) with Rec-Type and Expl-Type as between-subject factors. The dependent measures were the eight questions presented in Table 3. The factorial MANOVA revealed two overall significant main effects of Rec-Type ($F(1, 54) = 6.3535$; $p < 0.0001$; Wilk's $\lambda = 0.48$) and Expl-Type ($F(2, 54) = 2.5508$; $p < 0.005$; Wilk's $\lambda 2 = 0.49$) on the perceived quality of the conversational agent. Both H1-a and H1-b are validated. The interaction between the two variables was not significant ($F(2; 54) = 0.689$; $p = 0.80$; Wilk's $\lambda = 0.80$).

Our follow-up analysis looked at univariate effects for each dependent measure with two-way ANOVAs and followed up with a post-hoc analysis when necessary. In Table 3, we report a summary of all means and standard errors (in parentheses) for the eight dependent variables. The differences between the means are marked according to their level of significance (* for $p < 0.05$, ** for $p < 0.005$ and *** for $p < 0.001$). We give more details about the follow-up analyses and discuss the results in the sections below.

6.1.1 Rec-Type vs. quality of the conversational agent. The type of recommendations delivered by the agent had a significant impact on the perceived quality of the system. Indeed, the results of the independent two-way ANOVAs showed a significant main effect of Rec-Type on all the dependent variables except for the perceived usefulness: decision confidence ($F(1, 54) = 48.672$; $p < 0.001$; $\eta^2 = 0.44$), user control ($F(1, 54) = 6.360$; $p < 0.05$; $\eta^2 = 0.09$), intention to return ($F(1, 54) = 9.371$; $p < 0.005$; $\eta^2 = 0.14$), perceived effort ($F(1, 54) = 17.184$; $p < 0.001$; $\eta^2 = 0.22$), intention to watch ($F(1, 54) = 28.767$; $p < 0.001$; $\eta^2 = 0.33$), recommendation quality ($F(1, 54) = 37.839$; $p < 0.001$; $\eta^2 = 0.35$), and transparency ($F(1, 54) = 4.382$; $p < 0.05$; $\eta^2 = 0.06$).

For all the questionnaire items, the agent was rated with higher scores when delivering personalized recommendations (*pers-rec*) than when delivering random ones (*rand-rec*).

6.1.2 Expl-Type vs. quality of the conversational agent. The results of the independent two-way ANOVAs showed a significant main effect of Expl-Type on four of the dependent variables: decision confidence ($F(2, 54) = 3.474$; $p < 0.05$; $\eta^2 = 0.06$), recommendation quality ($F(2, 54) = 7.703$; $p < 0.005$; $\eta^2 = 0.14$), perceived usefulness ($F(2, 54) = 6.677$; $p < 0.005$; $\eta^2 = 0.19$), and transparency:($F(2, 54) = 4.355$; $p < 0.05$; $\eta^2 = 0.12$)

The results of the post-hoc analyses (after Bonferroni correction) show that the agent was rated with a significantly higher score in decision confidence ($p < 0.05$) when using our model of social explanations (soc-expl) than when using preference-based explanations (pref-rec), and with a significantly higher score in recommendation quality ($p < 0.005$) and perceived usefulness ($p < 0.005$) compared to the two other levels of explanations (pref-expl and no-expl). The agent was rated with a significantly higher score ($p < 0.05$) when using preference-based explanations than when delivering recommendations without any explanation.

Dimensions	Subjective items	Rec-Type		Expl-Type		
		rand-rec	**pers-rec**	**no-expl**	**pref-expl**	**soc-expl**
Decision Confidence	The movies recommended to me during this interaction matched my interests.	2.07(±.90)***	3.80(±.77)***	2.90(±.97)	2.55(±.88)*	3.35(±.83)*
User Control	SARA allowed me to specify and change my preferences during the interaction.	3.23(±1.05)*	3.90(±.64)*	3.5(±.89)	3.2(±1.21)	4.0(±.71)
Intention to Return	I would use SARA to get movie recommendations in the future.	2.47(±1.34)**	3.50(±1.03)**	2.75(±1.06)	2.65(±1.27)	3.55(±1.35)
Perceived Effort	I easily found the movies I was looking for.	2.10(±1.04)***	3.33(±1.01)***	2.40(±.87)	2.55(±1.19)	3.20(±1.19)
Intention to Watch	I would watch the movies recommended to me, given the opportunity.	2.53(±1.22)***	3.07(±.53)***	3.20(±.88)	3.05(±1.06)	3.65(±1.02)
Recommendation Quality	I was satisfied with the movies recommended to me.	2.33(±1.05)***	3.93(±.57)***	2.85(±.82)**	2.70(±.91)**	3.85(±.97)**
Perceived Usefulness	SARA provided sufficient details about the movies recommended.	2.97(±1.14)	3.40(±1.01)	2.80(±1.15)**	2.80(±.98)**	3.95(±1.11)**
Transparency	SARA explained her reasoning behind the recommendations.	2.67(±1.37)*	3.40(±1.24)*	2.35(±1.34)*	3.60(±1.25)*	3.15(±1.25)

Table 3: Subjective questionnaire adapted from [8] to measure users' perceived quality of the system.

Dimensions	Subjective items	Rec-Type		Expl-Type		
		rand-rec	**pers-rec**	**no-expl**	**pref-expl**	**soc-expl**
Coordination	I felt I was in sync with SARA.	2.23(±.97)**	3.13(±1.04)**	2.25(±.85)*	2.60(±1.20)	3.20(±1.09)*
	I was able to say everything I wanted to say during the interaction.	3.13(±1.22)*	3.77(±1.00)*	3.05(±1.28)	3.70(±1.01)	3.60(±1.03)
Mutual Attentiveness	SARA was interested in what I was saying.	2.97(±1.14)*	3.70(±.66)*	3.35(±.98)	2.80(±1.07)*	3.85(±.93)*
	SARA was respectful to me and considered to my concerns.	3.30(±1.06)***	4.13(±.60)***	3.60(±.80)	3.50(±.82)	4.50(±.89)
Positivity	SARA was warm and caring.	3.10(±1.16)	3.47(±.89)	3.00(±.91)	3.25(±1.26)	3.60(±1.12)
	SARA was friendly to me.	4.10(±.74)	4.20(±.62)	3.95(±.80)	4.15(±.66)	4.35(±.67)
Rapport	SARA and I established rapport.	2.67(±1.10)*	3.27(±.80)*	2.75(±.86)	2.75(±1.28)	3.40(±1.03)
	I felt I had no connection with SARA.	3.50(±1.09)	2.90(±1.24)	3.50(±.86)	3.25(±1.39)	2.85(±1.33)

Table 4: Subjective questionnaire adapted from [39] to measure users' perceived quality of the interaction.

6.1.3 Discussion. As hypothesized, the type of recommendation had a significant impact on the perceived quality of the conversational agent. Participants were more satisfied with the agent when it delivered personalized recommendations matching their preferences, regardless of the type of explanation that was used.

Although the preference-based explanations helped participants better understand the reasoning behind the agent's recommendation, our model of social explanations helped them learn more details about the recommendation. One solution for solving this trade-off would be to combine these two types of explanations. Indeed, as noted in section 3.3, we noticed that humans often use feature-based explanations to link two successive recommendations before delivering more details on the current one (e.g., "Speaking of Tom Cruise movies, what about *Edge of Tomorrow*? I found it exceptionally well-made in every aspect, intriguing, exciting and even funny in the right way"). An alternative solution for the agent would be to frame its explanation negatively (e.g., "I don't like Tom Cruise, but I found *Edge of Tomorrow* exceptionally well-made in every aspect, intriguing, exciting and even funny in the right way"). However, expressing a disagreement towards one of the user's preferences might be harmful.

Unlike [17], we did not find any evidence that social explanations would increase users' intentions to return. However, participants who received social explanations were more satisfied with the recommendations and believed these recommendations better matched their preferences. These results show that a conversational agent able to give its "own" opinions and refer to its personal experiences is perceived as more convincing.

6.2 Quality of the interaction

We conducted a 2x3 factorial MANOVA with Rec-Type and Expl-Type as between-subjects factors. The dependent measures were the eight questions presented in Table 4. The factorial MANOVA revealed two overall significant main effects of Rec-Type ($F(1, 54) = 2.3955$; $p < 0.05$; Wilk's $\lambda = 0.71$) and Expl-Type ($F(2, 54) = 1.9608$; $p < 0.05$; Wilk's $\lambda 2 = 0.56$) on the perceived quality of the interaction. Both H2-a and H2-b are validated. The interaction between the two variables was not significant ($F(2; 54) = 1.2524$; $p = 0.24$; Wilk's $\lambda = 0.68$).

Similar to the previous section, we performed a follow-up analysis that looked at univariate effects for each dependent measure with two-way ANOVAs and followed with a post-hoc analysis when necessary. In Table 4, we report a summary of all means and standard errors (in parentheses) for the eight dependent variables. The differences between the means are marked according to their level of significance (* for $p < 0.05$, ** for $p < 0.005$ and *** for $p < 0.001$). We give more details about the follow-up analyses and discuss the results in the sections below.

6.2.1 Rec-Type vs. quality of the interaction. The results of the independent two-way ANOVAs showed a significant main effect of Rec-Type on five dependent variables: the two items measuring coordination ("I felt I was in sync with SARA" ($F(1, 54) = 9.663$; $p < 0.005$; $\eta^2 = 0.13$) and "I was able to say everything I wanted during the interaction" ($F(1, 54) = 4.292$; $p < 0.05$; $\eta^2 = 0.07$)), the two items measuring mutual attentiveness ("SARA was interested in what I was saying" ($F(1, 54) = 6.892$; $p < 0.05$; $\eta^2 = 0.10$) and "SARA was respectful to me and considered to my concerns" ($F(1, 54) = 12.255$; $p < 0.001$; $\eta^2 = 0.17$)), and one item measuring rapport ("SARA and I established rapport" ($F(1, 54) = 4.154$; $p < 0.05$; $\eta^2 = 0.06$)).

In all these cases, the agent was rated with higher scores when delivering personalized recommendations (*pers-rec*) than when delivering random ones (*rand-rec*).

6.2.2 Expl-Type vs. quality of the interaction. The results of the independent two-way ANOVAs showed a significant main effect of Expl-Type on two dependent variables: one item measuring coordination ("I felt I was in sync with SARA" ($F(2, 54) = 3.474$; $p < 0.05$; $\eta^2 = 0.10$)) and one measuring mutual attentiveness ("SARA

was interested in what I was saying" ($F(2, 54) = 4.714$; $p < 0.05$; $\eta^2 = 0.13$)).

The results of the post-hoc analyses (after Bonferroni correction) show that the agent was rated with a significantly higher score in the coordination item ($p < 0.05$) when using our model of social explanations (soc-expl) compared to when using no explanation (no-rec), and with a significantly higher score in the mutual attentiveness item ($p < 0.005$) compared to the pref-expl condition.

6.2.3 Discussion. Participants preferred interacting with a conversational agent delivering personalized recommendations. This result matches with the findings from [12], which explains that while the interview phase might help find more relevant items for users, the additional questioning might lead to disappointment if the recommendation does not meet the user's expectations. This also shows that in a recommendation context, a conversational agent's task-performance influences rapport through enhanced coordination and mutual attentiveness, regardless of the explanations it uses.

Regarding the explanations, participants felt they were more in sync with a conversational agent using social explanations and considered the agent as more interested in what participants were saying. This can be linked to the computational model of rapport proposed in [38]: disclosing topic related personal information improves both mutual attentiveness and coordination. This is also consistent with our above results showing that participants who received social explanations found their recommendations to be more relevant; participants felt that the conversational agent was more interested in what they were saying, which resulted in a better-informed recommendation.

7 CONCLUSION

In this paper, we presented the human-centered design implemented in our conversational recommendation agent. Our model of social explanations, constructed through careful annotation and analysis of a relevant corpus, leveraged observed probabilities for identified categories and subcategories of recommendations. This was incorporated in the form of a content planner within our conversational agent's architecture. Our user experiment evaluated the influence of these social explanations on the perceived quality of our system as well as the interaction; results indicate that they significantly improved both. Moreover, a system using social explanations was perceived as more in sync with its users and more interested in what they were saying. This aligns with [17] and emphasizes the need to endow conversational recommendation systems with social conversational strategies, as well as to build systems able to express personal opinions and experiences.

One potential extension of this work would be to overcome the limited size of our initial corpus by annotating a larger dataset of movie reviews using our explanation categories. That would allow us to refine our content planner and would provide us with more examples to generate natural sentences. Although endowing our agent with a human-like identity might seem inappropriate (e.g. users know the agent cannot watch movies in theaters), the results from [9] show that the type of identity revealed by a virtual character (human-like vs. artificial) does not influence people's perception.

Another way to improve the perceived quality of the system and/or interaction would be to optimize the interview phase as suggested by [15]. Although almost all of the participants had a preferred movie genre, only a few specified a favorite director. Soliciting too many specific preferences could be stress-inducing, and participants might consequently overthink their responses. Moreover, as described in [23], a conversational recommendation system using sentences that are too-long (compared to the user's utterances) will decrease the quality of the interaction, and the recommendations are less likely to be approved. We thus seek to extend our sentence planner such that it can adapt the length of its explanations based on the length of the user's sentences; this improved means of generation will likely result in "better" recommendations.

ACKNOWLEDGMENTS

This work was supported in part by funding from Oath and the IT R&D program of MSIP/IITP [2017-0-00255, Autonomous digital companion development]. We would also like to thank John Choi for his generous help, Yoo Jin Shin for refining annotations, and the members of Carnegie Mellon University's ArticuLab for their feedback and support.

REFERENCES

[1] Amos Azaria and Jason Hong. 2016. Recommender systems with personality. In *Proceedings of the 10th ACM conference on Recommender systems*. ACM, 207–210.
[2] J. Berant, A. Chou, R. Frostig, and P. Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Empirical Methods in Natural Language Processing (EMNLP)*.
[3] George Y Bizer, Jeff T Larsen, and Richard E Petty. 2011. Exploring the valence-framing effect: Negative framing enhances attitude strength. *Political psychology* 32, 1 (2011), 59–80.
[4] Giuseppe Carenini, Jocelyin Smith, and David Poole. 2003. Towards more conversational and collaborative recommender systems. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 12–18.
[5] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer, 163–185.
[6] Rose Catherine and William Cohen. 2016. Personalized recommendations using knowledge graphs: A probabilistic logic programming approach. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 325–332.
[7] Shuo Chang, F Maxwell Harper, and Loren Gilbert Terveen. 2016. Crowd-based personalized natural language explanations for recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 175–182.
[8] Li Chen and Pearl Pu. 2008. A cross-cultural user evaluation of product recommender interfaces. In *Proceedings of the 2008 ACM conference on Recommender systems*. ACM, 75–82.
[9] Setareh Nasihati Gilani, Kraig Sheetz, Gale Lucas, and David Traum. 2016. What kind of stories should a virtual human swap?. In *International Conference on Intelligent Virtual Agents*. Springer, 128–140.
[10] John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. IEEE, 517–520.
[11] Jonathan Gratch, Anna Okhmatovskaia, Francois Lamothe, Stacy Marsella, Mathieu Morales, Rick J van der Werf, and Louis-Philippe Morency. 2006. Virtual rapport. In *International Workshop on Intelligent Virtual Agents*. Springer, 14–27.
[12] Ulrike Gretzel and Daniel R Fesenmaier. 2006. Persuasion in recommender systems. *International Journal of Electronic Commerce* 11, 2 (2006), 81–100.
[13] Ido Guy, Naama Zwerdling, David Carmel, Inbal Ronen, Erel Uziel, Sivan Yogev, and Shila Ofek-Koifman. 2009. Personalized recommendation of social software items based on social relations. In *Proceedings of the third ACM conference on Recommender systems*. ACM, 53–60.
[14] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1661–1670.

[15] Michael Jugovac and Dietmar Jannach. 2017. Interacting with recommenders—overview and research directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 7, 3 (2017), 10.

[16] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *International workshop on intelligent virtual agents*. Springer, 205–217.

[17] SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103 (2017), 95–105.

[18] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*. 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010

[19] Theodora Nanou, George Lekakos, and Konstantinos Fouskas. 2010. The effects of recommendations' presentation on persuasion and satisfaction in a movie recommender system. *Multimedia systems* 16, 4-5 (2010), 219–230.

[20] Neal R Norrick. 2005. Interactional remembering in conversational narrative. *Journal of Pragmatics* 37, 11 (2005), 1819–1844.

[21] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (2017), 393–444.

[22] Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. 2012. A generalized taxonomy of explanations styles for traditional and social recommender systems. *Data Mining and Knowledge Discovery* 24, 3 (2012), 555–583.

[23] Florian Pecune, Jingya Chen, Yoichi Matsuyama, and Justine Cassell. 2018. Field Trial Analysis of Socially Aware Robot Assistant. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1241–1249.

[24] Ivens Portugal, Paulo Alencar, and Donald Cowan. 2018. The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications* 97 (2018), 205–227.

[25] Arpit Rana and Derek Bridge. 2018. Explanations that are Intrinsic to Recommendations. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 187–195.

[26] Derek D Rucker, Richard E Petty, and Pablo Briñol. 2008. What's in a frame anyway?: A meta-cognitive analysis of the impact of one versus two sided message framing on attitude certainty. *Journal of Consumer Psychology* 18, 2 (2008), 137–149.

[27] Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742* (2015).

[28] Matthew Stone, Christine Doran, Bonnie Webber, Tonia Bleam, and Martha Palmer. 2003. Microplanning with communicative intentions: The SPUD system. *Computational Intelligence* 19, 4 (2003), 311–381.

[29] Jan Svennevig. 2000. *Getting acquainted in conversation*. John Benjamins.

[30] Panagiotis Symeonidis, Alexandros Nanopoulos, and Yannis Manolopoulos. 2009. MoviExplain: a recommender system with explanations. *RecSys* 9 (2009), 317–320.

[31] Linda Tickle-Degnen and Robert Rosenthal. 1990. The nature of rapport and its nonverbal correlates. *Psychological inquiry* 1, 4 (1990), 285—-293.

[32] Nava Tintarev and Judith Masthoff. 2007. A survey of explanations in recommender systems. In *2007 IEEE 23rd international conference on data engineering workshop*. IEEE, 801–810.

[33] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 399–439.

[34] Vivian Tsai, Timo Baumann, Florian Pecune, and Justine Casell. 2018. Faster responses are better responses: Introducing incrementality into sociable virtual personal assistants. In *Proceedings of the 2018 International Workshop on Spoken Dialog System Technology*.

[35] Pontus Wärnestål, Lars Degerstedt, and Arne Jönsson. 2007. Interview and delivery: Dialogue strategies for conversational recommender systems. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*. 199–205.

[36] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)* 52, 1 (2019), 5.

[37] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192* (2018).

[38] Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International Conference on Intelligent Virtual Agents*. Springer, 514–527.

[39] Ran Zhao, Oscar J Romero, and Alex Rudnicky. 2018. SOGO: A Social Intelligent Negotiation Dialogue System. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. ACM, 239–246.

A ANNOTATION EXAMPLES

(a) Annotation Example 1

Speaker	Sentence	Annotation
A	So we went to see SOAPDISH and...	
B	Was it good?	
A	Oh, hysterical.	**[PO_POS]**
	We laughed so hard, it was just, you couldn't hear half the dialogue because everyone in the audience was laughing.	**[PE_A]**

(b) Annotation Example 2

Speaker	Sentence	Annotation
A	Have you seen the movie CLASS ACTION with Gene Hackman?	**[MF_C]**
B	No, I haven't yet.	
A	I saw it this weekend	**[PE_L]**
	and it is, uh, to me an outstanding movie.	**[PO_POS]**
	I thoroughly enjoyed it.	**[PO_POS]**
	He is, uh, an attorney and his daughter is an attorney and she has a suit against his company.	**[MF_P]**

(c) Annotation Example 3

Speaker	Sentence	Annotation
A	DANCES WITH WOLVES did not seem to have anything added. It was just a legitimate kind of film.	**[PO_ANA]**
	And that is the reason why I suppose it won so many Oscars,	**[MF_A]**
	because it really was good even though it is such a long movie.	**[PO_SO]**
	You know, they said, "Oh, people won't be interested in a three hours movie." But it certainly gotten good acclaim everywhere it has gone.	**[TPO_B]**

Figure 2: Three annotated movie chunks from the corpus.