Conversational review-based explanations for recommender systems: Exploring users' query behavior

Diana C. Hernandez-Bocanegra
University of Duisburg-Essen
Duisburg, Germany
diana.hernandez-bocanegra@uni-due.de

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

Providing explanations based on user reviews in recommender systems (RS) can increase users' perception of system transparency. While static explanations are dominant, interactive explanatory approaches have emerged in explainable artificial intelligence (XAI), so that users are more likely to examine system decisions and get more arguments supporting system assertions. However, little attention has been paid to conversational approaches for explanations targeting end users. In this paper we explore how to design a conversational interface to provide explanations in a review-based RS, and present the results of a Wizard of Oz (WoOz) study that provided insights into the type of questions users might ask in such a context, as well as their perception of a system simulating such a dialog. Consequently, we propose a dialog management policy and user intents for explainable review-based RS, taking as an example the hotels domain.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Human-centered computing** → **User studies**; **Natural language interfaces**.

KEYWORDS

Recommender systems, explanations, argumentation, conversational agent, user study

ACM Reference Format:
Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *3rd Conference on Conversational User Interfaces (CUI '21), July 27–29, 2021, Bilbao (online), Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3469595.3469596

1 INTRODUCTION

Customer reviews have been increasingly used for explaining decisions made by recommender systems (RS), due to their wealth of detailed information on positive and negative aspects of items, which cannot be obtained directly from ratings. Although review-based explanations can be useful in improving the perception of efficacy and trust in RS, these are almost always presented in a static manner, often as an aggregation of opinions, limiting users in exploring the diverse views and arguments expressed in the reviews. On the other hand, interactive methods may positively influence user perception of RS [22], by allowing the user to request, for example, further elaboration of the claims made by the system. However, explanatory methods that allow users to scrutinize and customize explanations through interaction are largely unexplored, or lack sufficient empirical evidence [56]. Additionally, most interactive approaches in RS and, in a wider scope, in explainable artificial intelligence (XAI), are based on point and click options. However, recent developments in natural language processing (NLP) and natural language generation (NLG) enable a more flexible interaction, where users could indicate, in their own words, their explanation needs.

In particular, we aim to explore the feasibility and implications of using conversational approaches to explanations in review-based RS, and in particular the use of conversational agents (CA), given their ability to enable two-way natural language communication, opening up the range of possible questions a user can ask the system, which could contribute to a better understanding and acceptance of explanations by users, as prescribed by conceptual models of explanation based on dialogue [23, 62]. Although user interfaces inspired by human-to-human conversation have been developed and used for a long time to assist users in a wide range of tasks [46], little is known about how a CA should be conceptualized or designed in the context of XAI, and in particular, in explainable RS. Thus, we aim to explore:

RQ1: How to design a dialog management policy to implement a CA with explanatory purposes in review-based RS?

In this paper, we focus on the analysis of conversation patterns within an explanatory process. In this regard, [43] have drawn attention to the social and communicative aspect of explanation ("someone explains something to someone" [23]) and how an interactive and conversational approach could contribute to increasing user understanding in XAI approaches. While a general theoretical model of explainable recommendations has not yet been established, we propose to analyze explanations through the lens of argumentation theory. A first category of argumentation models seeks to define logical structures containing assertions, supporting evidence, refutations, among others [7]. A second category involves dialectical approaches [63], focusing on the exchange of arguments and supporting (or contradictory) information within a dialogue between two parties.

Thus, our goal is to explore the modeling of explanations in review-based RS as an argumentative dialogue, and how this can be facilitated by a conversational user interface. However, the above

CUI '21, July 27–29, 2021, Bilbao (online), Spain

ACM ISBN 978-1-4503-8998-3/21/07…$15.00
https://doi.org/10.1145/3469595.3469596

requires a close understanding of how a user would formulate questions in this particular setting. Particularly we aim to answer:

RQ2: How do review-based RS users communicate their explanation needs using a CA?

To this end, we conducted a WoOz study [26], taking as an example the hotels domain, since it represents an interesting mix between search goods (with attributes on which complete information can be found before purchase [49]) and experience goods (which cannot be fully known until purchase [49]). Such a product evaluation could benefit from third-party opinions [27, 49], potentially rich in argumentative information that can be used for explanatory purposes in RS. The results of our analysis provided a basis to formulate a dialog management policy for explainable review-based RS, and to draw attention to the challenges involved in implementing such an approach. The contributions of this paper can be summarized as follows:

- We propose a dialog management policy for explanations as conversational argumentation in review-based RS, inspired by dialog models and argument theories.
- We modeled the intents that can be used for the implementation of a CA for explanatory purposes in the hotels domain, based on a WoOz study, and analyzed to what extent follow-up questions were formulated.
- Participants' perception of a simulated system was evaluated in terms of system transparency, trust and effectiveness, as well as satisfaction with the explanation, sufficiency, confidence and persuasiveness.

2 RELATED WORK

2.1 Review-based and argumentative explanations

Explanations can bring several benefits to RS, by increasing users' perception of transparency, effectiveness, and trust [58]. Review-based explanatory RS integrate ratings and reviews to generate both predictions and explanations (e.g. [6, 64, 68]), usually presented as summaries of the positive and negative opinions on different aspects (e.g. [48]). Moreover, exploitation of reviews can facilitate the generation of argumentative explanations, [20], in which system claims (user will find a recommended item useful) are supported by evidence found and consolidated from reviews.

While argumentative approaches have already been applied to explanations, these are mainly based on the static display of the arguments, as in [4, 11, 20, 30, 66], where little can be done to indicate to the system that additional information is still needed to fully understand and accept the explanations. In contrast, interactive and conversational approaches to explanations seek to grant users further control over explanatory components [22, 56], in order to promote a better understanding of the rationale behind system predictions, based on the idea of an exchange of questions and answers between the user and the system, as would occur in a human explanatory conversation [43].

2.2 Conversational explanations

Accordingly, formal explanation dialogues have been conceptually formulated as theoretical support to the design of conversational explanation approaches [2, 14, 39, 54, 60]. Interactive and conversational explanations have been already addressed in the field of explainable artificial intelligence (XAI), although to a much lesser extent compared to static explanations [1], and mostly focused on the influence that features or data points have on machine learning predictions. For example, [56] proposed a system that provides explanations as an interactive dialogue that resembles a natural language conversation supporting why-questions, to facilitate the understanding of machine learning classification outcomes, e.g. the rejection of a credit loan. However, this approach differs from ours in that we use non-discrete and non-categorical sources of information, subjective in nature and unstructured, which are nevertheless rich in arguments that can be used to answer questions of a subjective nature. Similarly, [54] defined a protocol to provide conversational argumentative explanations in RS, however it restricts the possible user interactions to a limited set of possible questions a user may ask, while we explore possibilities for users to express their explanatory needs in their own words. Finally, despite the potential benefit of using dialog models to increase users' understanding of intelligent systems [43, 65], their practical implementation in RS (and in XAI in general) still lacks sufficient empirical evaluation [39, 43, 56], thus, it is still unclear how conversational explanatory interfaces should be conceived and designed, so that they actually improve users' perception of RS.

Consequently, we set out in this paper to explore the design of a dialog management policy for conversational explanations in RS, exploiting the potential benefits of a dialog system (particularly a CA or chatbot), where users can indicate their explanatory needs in their own words, in the form of questions. Our work differs from the traditional approach to CA in the hotel domain, which focuses on processes like customer service and booking assistance [10], and to conversational RS (e.g. [13, 67]) which aim to collect user preferences to generate recommendations through dialog. We aim, on the other hand, to explore the implications and effects of using CA to explain RS rationale, which remains largely unexplored [24]. A model of social explanations for movie recommendations was proposed by [51], in one of the few works on the subject. However, according to their approach, it is the system who leads the conversation, providing justifications for recommendations even when they are not explicitly requested by the user, whereas according to our proposal, the user would have the active role, being enabled to ask the questions that lead to an argumentation by the system.

2.3 Question answering (QA)

Our work is closely related to QA systems, which aim to answer questions posed by users in natural language, by using techniques like information retrieval (IR) or NLP, on various types of web documents or in knowledge bases. While most of QA systems are designed to respond factoid, definition, or list questions by offering excerpts from documents or list of items consistent with user's query, much less work has been devoted to advanced "how-to", "why", evaluative, comparative, and opinion questions [34, 44], that require usually the aggregation and comparison of multiple items over different pieces of information. Lipton [35] defines *explanation* as an answer to a *why-question* , however, other types of questions

can also be answered by explanations, i.e. how? what? [43], the latter being one that could be answered with a factoid sentence, for which we aim to support both factoid and advanced question types. Additionally, and in contrast to the common QA approach where the system replies to a series of standalone questions, interactive QA involves a dialog interface enabling related, follow-up and clarification questions [53].

Nevertheless, our approach differs from most QA methods, especially those based on IR, because in our case, responses should not be generated solely on the basis of information sources, but should be consistent and reflect the mechanism used to generate the recommendations. Additionally, to answer complex questions (e.g., "why"), our approach involves a focus on the most relevant aspects for users, to provide concise and relevant statements that aggregate information from different reviews. To this end, our approach relies on the user profile inferred by the RS algorithm, especially when no explicit features are addressed in users' questions. On the other hand, implicit user preferences are not taken into account in most QA approaches, which stems from their use of IR methods, where the relevance of a document is estimated based on how much its content is related to the query [45]. Additionally, we propose to follow an argumentative explanation structure to generate responses, which could improve users' perception of RS, as evidenced using the interactive, although not fully conversational approach proposed by[22]. Although argumentation has already been exploited in QA [47], it has been mostly used to extract high quality answers by means of argument mining, whereas very few approaches exploit argumentation as a way of presenting explanations in response to user queries [3].

2.4 Users' utterances on explanation needs

The design of adequate conversational explanations requires a proper understanding of possible user requests [33], which may vary according to the type of application, the context and user characteristics. [32] collected a dataset consisting of written conversations between humans with a movie recommendation goal, however, no explanatory inquiries like "Why do you recommend X?" are addressed . Furthermore, [8] collected a QA dataset for several domains (including hotels), which can be used to generate answers not limited to factoid questions, but also to subjective ones (e.g. "How is it the location?"). However, questions and answers only address one item at a time, leaving out comparison queries; moreover, the dataset is not oriented as such to an explanatory dialogue. On the other hand, [33] noted that a question-oriented framework offers a feasible way to conceive interactive explanations, and proposed a XAI question bank, consisting of inquiries that users might typically ask about AI algorithms. However, as it is the case for most XAI interactive approaches, this question bank was intended for explanation needs of users with expert knowledge in AI, whereas no similar question bank definition has been developed, to our knowledge, for end users and, in particular, for RS. Consequently, we conducted a user study using the WoOz [26] method, to capture the possible questions users would ask in the context of RS explanations, particularly in the hotel domain.

2.5 WoOz paradigm

WoOz studies allow for the incremental design of conversational interfaces, and involve the simulation of a human-machine interaction, in which a member of the research team (the *wizard*) simulates the response actions of the system, through a computer-mediated interface, a technique that has been widely adopted for HCI prototyping [15, 40]. The use of this type of technique allows to validate how users would interact with a conversational interface, and to evaluate the feasibility of dialog based systems that have not yet been fully implemented, as was done for example by [53] to design an open domain interactive QA system, or by [5] for the design of a conversational framework to support recipe recommendation.

3 EXPLANATIONS AS CONVERSATION

Our proposal is based on previous work reported in [22], where the effect of different levels of interactivity for accessing explanatory information was tested, without considering a CA perspective as such. Such approach was inspired by dialog-based explanation models [39, 61, 62] and the argumentative scheme by [19], and regards an explanation as an interactive argumentation, that is, an explanation consisting of a cyclic sequence of *argumentation attempts* made by the system in response to *argument requests* made by the user, as a way to challenge or critique system argumentation, or to inquire for further arguments, using why, what, or how-questions (Figure 1). Argumentation attempts include premises (a general reason to accept a claim that a recommended item is worthy to be chosen) and backing (specific information or additional evidence to support the claim, e.g. percentage of positive opinions about an aspect), among others.

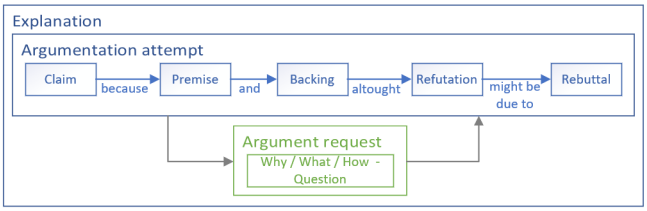

Figure 1: Simplified scheme for explanations as interactive argumentation in review-based RS [22].

Despite the positive perception by users of a system that implemented such a scheme, its components were not directly derived from observed natural human conversation, leading to the following constraints: 1) it only offers answers to a limited set of questions, 2) it does not consider comparative questions, e.g. "why is X better than Y?", 3) nor factoid questions, e.g. "does this hotel include breakfast?", 4) nor questions regarding users' own profile, or algorithm details. In consequence, we extended this scheme to support a wider range of questions that could be written by users in their own words, and used it as basis for the valid moves of the wizard in our study (Figure 2). Refutation and rebuttal components from proposal in [22] were left out from current proposal, for the sake of brevity of responses by the system (following guidelines by [46]), and will be evaluated in future work.

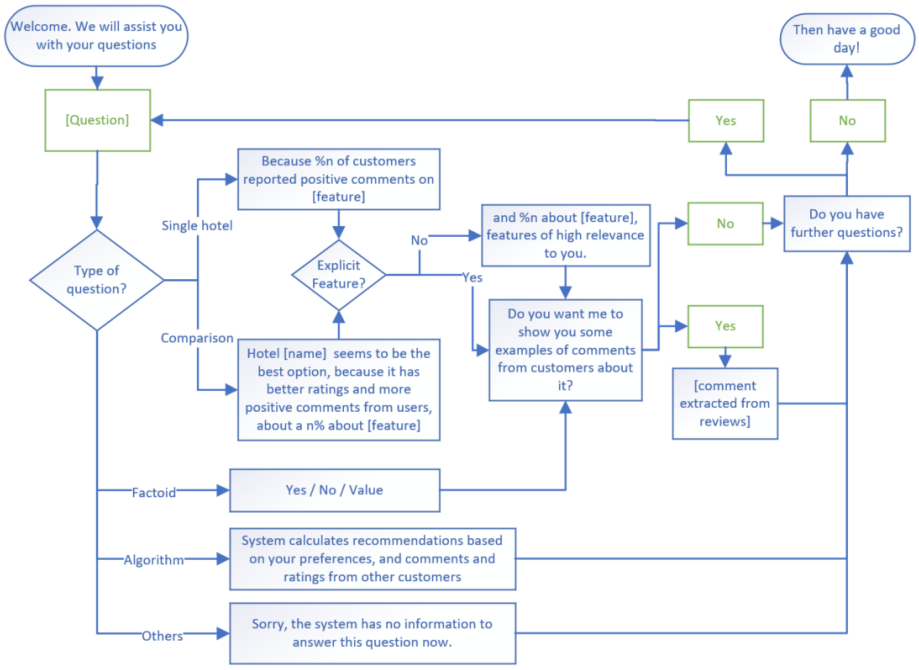

Figure 2: Scheme for conversational explanations used in WoOz experiment. Blue boxes represent utterances by the system, green boxes the utterances by users.

4 USER STUDY

We conducted a WoOz study to explore how users would express their explanatory needs, to a CA in a review-based RS, with hotels as an example domain. All subjects were assigned to the same experimental condition, and were instructed to interact with the RS, and to write their questions about the reasons for the recommendations, which were replied by the wizard (played by our main researcher). We used the scheme described in the previous section (Fig 2), as the guideline for the wizard, aiming to portray a structured conversation similar across participants. Particularly, we hypothesize that users will ask questions of the types *why?*, *how?*, and *what?*, as well as factoid, comparative, and evaluative questions, at the feature level as well as at the general level. Further details about the study are described below.

4.1 Participants

We recruited 20 participants (10 female, mean age 34.65 and range between 20 and 69) through Amazon Mechanical Turk. We restricted the execution of the task to workers located in the U.S, with a HIT (Human Intelligence Task) approval rate greater than 98%, and a number of HITs approved greater than 500. Participants were informed in consent form and instructions about payment rejection (if no effective interaction with the system) which could be checked using system logs, and responses to validation questions in questionnaires (e.g. "recommendations were based on: Opinions of celebrities, True/False", "The purpose of this question is to check attentiveness, please mark Disagree"). We discarded participants with less than 5 (out of 7) correct answers, or no effective interaction with the wizard. The responses of 12 of the 32 initial participants were then discarded for a final sample of 20 subjects. Participants were rewarded with $2 plus a bonus up to $0.40 depending on the quality of their response to the question "Why did you choose this hotel?" set at the end of the survey, aiming to achieve a more motivated choice by the participants, and to encourage an effective interaction with the system. Time devoted to the task by participants (in minutes): M=12.99, SD= 2.24.

4.2 Procedure

We informed participants that a list of hotels reflecting the results of a hypothetical search and within the same price range would be presented (i.e no filters to search hotels were offered), and that they could consult the general hotel information (photos, reviews, etc., by clicking on "Info Hotel"), but also freely enter any question of interest about one or more hotels in the chat box located on the right of the hotel list. We underlined that the chat box was designed to explain the reasons for the recommendations, in order to prevent the user from asking questions about other processes, such as booking assistance. We presented a cover story, to establish a common starting point in terms of travel motivation, asking participants to imagine the planning of a vacation trip, as in pre-COVID19 times, and that they had to decide which hotel to stay at. We requested the 5 most important hotel aspects to the participant, ranked in order of importance, to calculate personalized recommendations. We then presented the system showing a list of 6 recommended hotels (sorted by predicted recommendation score) and the "chatbox" (Figure 3). A debrief was provided at the end, indicating the main objective of the study.

4.3 Ethic concerns

The WoOz technique relies on deception: participants are supposed to believe they are interacting with a system, so researchers can have a better perception of what users would do when interacting with a real machine. Such a set up raises some ethical concerns given the necessary deception [15]. Following guidance from [17, 50, 57], we took the following actions to mitigate negative effects due to the study deception:

- We avoided an explicit mention of a "full automated" system or chatbot, instead we referred to a "chat box", where they could type their questions.
- We disclosed in the debrief that the responses were written by a human, and that the participants could request for the withdrawal of their responses in case they consider that the procedure went against their expectations, with payment being processed anyway.
- The main study researcher played the wizard, following a pre-established dialog flow, to avoid statements out of project scope that could harm or make participants uncomfortable.

4.4 Dataset and implemented system

Dataset. We used the ArguAna dataset [59], (hotel reviews and ratings from TripAdvisor; sentiment and explicit features annotated sentence wise), and the aspect annotation done by [22], in order to provide aspect based arguments.

Explainable RS. We used the review-based RS developed by [22], which implements the matrix factorization model proposed by [68], in combination with sentiment-based aspect detection methods, using the state of art NLP model BERT [16].

Conversational interface. We used Flask-SocketIO, a Socket.IO integration for Flask applications [18], to allow communication between participants and the wizard. Figure 3 depicts the interface presented to participants

Support system. To obtain the desired benefits, the wizard had to produce responses as fast and consistently as possible, so that users still feel they were interacting with a machine. This can only be achieved if the wizard uses a suitable support system [15], that provides, beyond canned sentences, appropriate answers consistent with participants preferences and the information they can obtain in their own system view. Thus, we added a module to the RS to quickly generate the answers, so the wizard could copy and paste them in the conversational interface.

Personalization mechanism: To reduce implications of the *cold start* problem [55] (system does not have enough information about the user to generate an adequate profile and thus, personalized recommendations), participants were asked for their aspects of most importance, and the RS selected the user with the highest preference similarity within the rating matrix of the RS algorithm to generate predictions.

4.5 Questionnaires

System perception. We evaluated system perception based on explanatory aims defined by Tintarev [58]. We focused on the subset effectiveness and trust, for which a significant effect of interactive options to explain was found in [22], and on transparency, and on transparency, for which an effect of conversational features is expected, as predicted by the dialogue models of explanation [62]. We utilized items for transparency [52] (user understands why items were recommended), effectiveness [28] (internal reliability Cronbach's α = 0.94, system is useful and helps the user to make better choices), and trust [42] (α = 0.92, user trusts system recommendations).

Explanations perception. We used single items from [29], which involve aspects related to explanations rather than the overall system: explanation confidence (user is confident that she/he would like the recommended item), explanation satisfaction (user would enjoy a system if recommendations are presented this way), and persuasiveness (explanations are convincing), and from [22] for sufficiency (explanations provided are sufficient to make a decision). All items were measured with a 1-5 Likert-scale (1: Strongly disagree, 5: Strongly agree).

4.6 Data Analysis

We first manually classified utterances into categories: questions and no-questions, the latter including e.g. greetings or gratitude statements. We categorized every question according to the dimensions: *scope*, *comparison*, *assessment* and *detail*, following an inductive category formation [41], i.e. we started with one category and benchmarked each question against the criteria of the category. Following that, we either classified the question into the existing category or created a new one. This step involved two independent annotators, who came to a Cohen's kappa = 0.91, almost perfect agreement intercoder reliability [31].

We checked whether questions were standalone questions, or follow-up questions, validating the presence of anaphoras ("a linguistic form whose full meaning can only be recovered by reference to the context") and ellipsis ("an omission of part of the sentence, resulting in a sentence with no verbal phrase") [53]. We used criteria from [53] and [9] to classify anaphoras (pronoun or possessive adjective, and noun phrase anaphora), and ellipsis.

Finally, we evaluated questions according to the feasibility of their automated response, and classified them according to possible methods that could be used to do so.

5 RESULTS

We collected a total of 20 dialogues and 105 utterances (M=5.20 utterances per participant, SD=2.48). 81 of the utterances were questions (M=4.05 questions per participant, SD=2.14). The average question length is 46.70 characters (SD=30.86). We observed that the conversations adhered to the explanatory objective, and not to other purposes, such as, booking process.

5.1 Intents and entities

We identified that users' intents could be classified into two main types: *domain-related* intents (regarding hotels and their features), and *system-related* intents (regarding the algorithm, or the system input). In turn, domain-related intents could be categorized according to the following dimensions:

- *Scope*: Whether the question refers to a single item (*single*), a limited list of items (*tuple*), or to no particular item (*indefinite*).

Figure 3: User interface presented to participants in WoOz study.

- *Comparison*: Whether the question is (*comparative*) or not (*non-comparative*). We adopt the comparative sentence definition by [25] "expresses a relation based on similarities or differences of more than one object", including superlatives and relations like "greater" or "less than".
- *Assessment*: Whether the question refers to the existence or characteristics of item features (*factoid*), to a subjective assessment of the item or its features (*evaluation*), or to system reasons to recommend an item (*why-recommended*).
- *Detail*: Whether the question inquires for an specific aspect or feature (*aspect*), or for the overall item (*overall*).

Consequently, the intent of a single domain question could be defined as a combination of the 4 dimensions. Table 1 contains examples for every dimension / value, Figure 4 depicts the distribution of questions regarding every dimension, and Table 2 contains examples of intents, and their frequency in the collected utterances. It is important to note that all but one of the questions could be correctly classified as system-related intent, namely: "why are there so few reviews?".

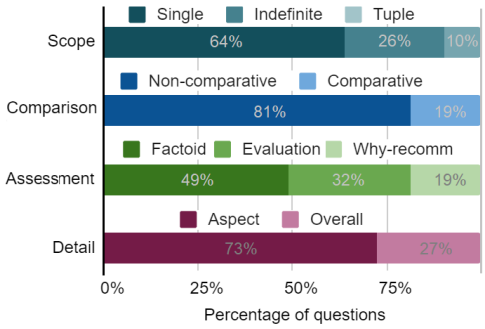

Figure 4: Distribution of questions according to each dimension of domain-related intents.

All questions of domain intent regarded the entities: *hotel* and *hotel feature*.

5.2 Follow-up questions

Figure 5 shows the distribution of standalone and follow-up questions. A special case are inquiries that could work as both types. Such is the case for comparative questions under the value "Indefinite" of dimension *scope*, which may refer to the best among all possible options (e.g. "which is the best hotel?") or, if a subset of options was previously discussed, as a follow-up, (e.g. "I am choosing between the Riley and the Evelyn. Which is the best hotel overall?").

Additionally, Figure 5b shows the distribution of follow-up question types: pronoun or possessive adjective anaphoras (e.g. " I'm looking for facts about current internet service - is **it** unchanged or upgraded?"), noun phrase anaphora (e.g. "When was the last time **the Hotel** underwent a remodel?), and ellipsis (e.g. "what are the checking in times for hotel owen? **and hotel evelyn?**"). We noted that pronouns and noun phrases in anaphoras referred only to hotels names or hotel features.

Moreover, 11% of utterances contained non-question sentences aiming to establish a context for a subsequent question, e.g. "I like the ambiance of the Hotel Evelyn, how were the reviews for that?". Finally, only 2.4% of utterances contained more than one question.

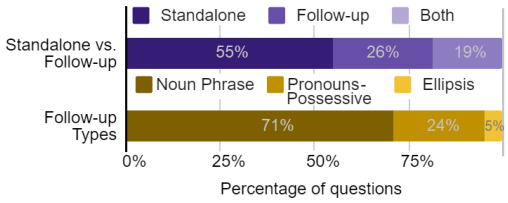

Figure 5: Distribution of follow-up questions.

5.3 Methods for generating answers

The number of questions that could be answered with different types of methods is shown in Figure 6. Some could be replied by using several methods, e.g. "How close is Hotel Julian to the city

Table 1: Example of domain-related intents classified by dimension.

Dimension	Value	Question is about	Example
Scope	Single	A specific item	How is the food at **Hotel Evelyn**?
	Tuple	Two or more items	Are either hotel **owen** or **evelyn** near station?
	Indefinite	No specific item (s)	Which hotel has the best views?
Comparison	Comparative	Relation of similarities or differences of more than one object.	what is **difference between** hotel evelyn and hotel james? Which hotel has the **best** views?
	Non-comp	No comparison	How close is Hotel Owen to the subway?
Assessment	Factoid	Facts, item having features or not	**Does** Hotel Owen have TV service?
	Evaluation	How hotel is evaluated (subjectively)	**How** is the food at Hotel Evelyn?
	.	Which hotel/feature is better/best	**Which** hotel has the **best** view?
	Why-recomm	Reasons of recommendations	**Why** is Hotel Julian my top recommendation?
Detail	Aspect	A specific aspect or feature	Why is it Hotel Julian in a good **location**?
	Overall	No specific aspect or feature	How good is hotel Julian?

Table 2: Most frequent domain intents (combination of dimensions values) sorted by number of questions per intent (desc.)

Scope	Comparison	Assessment	Detail	Example	# Qs	Type of initial response
Single	Non-comp	Factoid	Aspect	Does Hotel Julian have a pool?	29	Y/N or value
Single	Non-comp	Why-recomm	Overall	Why is Hotel Julian my top recommendation?	14	Because [Argument backing]
Single	Non-comp	Evaluation	Aspect	How is the food at Hotel Evelyn?	8	[Argument claim], because [Argument backing]
Indefinite	Comparative	Evaluation	Aspect	Which hotel has the best customer service?	7	Hotel X, because [Argument backing]
Indefinite	Non-comp	Factoid	Aspect	Do any of the hotels provide free breakfast?	6	Y/N or value
Tuple	Non-comp	Factoid	Aspect	what are the checking in times for hotel owen and hotel evelyn?	4	Y/N or value
Indefinite	Comparative	Evaluation	Overall	Which hotel has the best reviews?	4	Hotel X, because [Argument backing]
Indefinite	Non-comp	Evaluation	Aspect	what rooms would be good for parents with children?	3	Hotel X, because [Argument backing]
Tuple	Comparative	Evaluation	Overall	What is difference between hotel evelyn and hotel james?	2	Hotel X has better comments on [feature x] and [feature y].

centre?" could be replied both using hotel metadata, or a QA method to retrieve answers from users comments. Additionally, according to our proposed scheme, a question like "How is the food at Hotel Evelyn?" could be replied by presenting an aggregation of opinions, and by providing an example of such opinions extracted from reviews. Finally, 17% of the questions could not be directly replied to by any of our contemplated methods, e.g. "how was the price of the hotel decided?", given that price is not decided directly by the RS. Although we intended to provide approximate answers to questions such as "Has Hotel Evelyn made any upgrades to its internet/wi-fi service since some of its reviews were written?", such as "X% of customers reported positive opinions about wi-fi", it may not be enough to satisfy very curious users, as in this case, where we got the counter-response: "That doesn't answer my question".

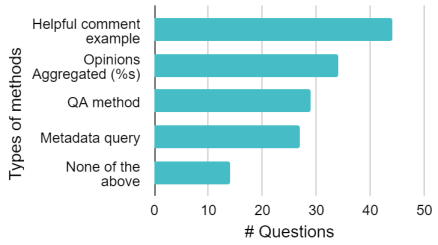

Figure 6: Number of questions that could be responded by different types of methods.

5.4 Perception of system and explanations by users

Figure 7 shows the distribution of users' perception, and the distribution of topics addressed in suggestions and comments provided by participants at the end of the study, in their own words.

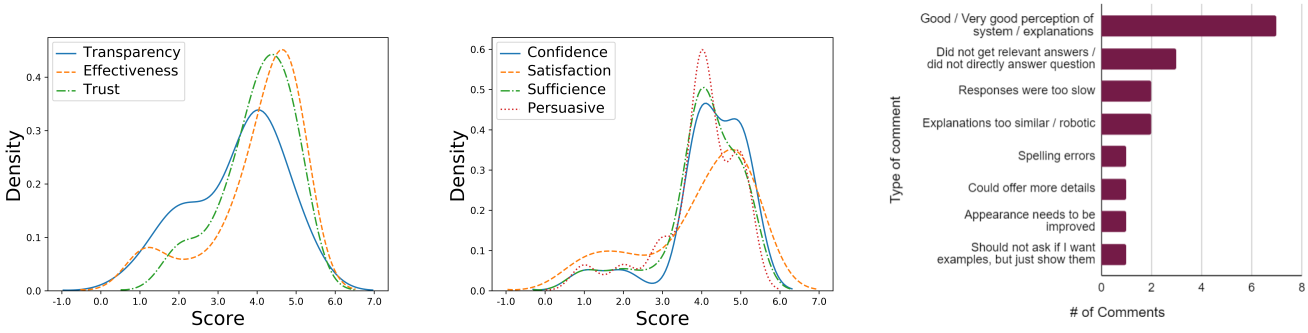

Figure 7: Kernel density estimates of participants' scores for perception of system (left) and explanations (middle); higher score values indicate a more positive perception. Distribution of comments and suggestions from participants (right)

6 DISCUSSION

Suitability of the approach. We consider that our proposed scheme and the WoOz study setup have been useful and effective for our purpose of exploring the use of CA to explainable review-based RS, given the predominantly positive perception of RS and its explanation by participants - especially in terms of effectiveness and trust -, and the observation that collected conversations adhered to the explanatory objective as expected, i.e. no questions regarding other processes were asked, like hotel booking.

Moreover, we observed that users actively expressed their needs for explanation, taking the lead in formulating their own questions (not expecting the system to choose what to explain) and challenging the system's attempts at argumentation when the answers provided did not satisfy their need. We believe that an implementation of our dialog management policy might contribute to a better perception of the RS, in comparison to interfaces providing only static explanations, or interactive but with a very limited set of possible questions to be answered, since 1) it allows for greater active control (voluntary actions that can influence the user experience [36]), which might be beneficial in environments involving information needs and a clear goal in mind [36], and 2) the two-way communication enabled, which might contribute to a better acceptance and understanding of arguments, as predicted by dialog models of explanation [23, 62].

Types of questions. As expected, participants asked both factoid questions and evaluative and why-recommended questions. Although not handled by the method our work is based on (matrix factorization model that integrates reviews [68]), the input from factoid questions could be handled as wish conditions, and lead to changes in recommendations' appearance (highlighting those that match the desired conditions) or to recalculate recommendations' ranking, as is done in critique-based RS. This has been proven to be beneficial to user experience [12, 37, 38], thus we believe it may also be useful to integrate it into our approach, once the factoid response does not remain a flat answer for a single item, but can be applied to the entire set of options, to facilitate comparisons to make a final decision.

Comparing our collected inquiries with the prototypical questions from XAI question bank by [33], we found that their why-questions had a similar objective to the our why-recommended: to ask for reasons why certain predictions have been provided. However, we also observed substantial differences in regard to other types of questions:

- Input questions (e.g. "what kind of data does the system learn from?") were asked only once in our study.
- No questions were asked about output (e.g., "what does the system output mean")
- Neither on performance (e.g. "how accurate or reliable are predictions?").
- We noted that how-questions asked mostly "how the opinions are" rather than asking about the overall logic of the system.
- No "What if?" questions were asked. However, factoid questions might implicitly ask such questions (e.g. "Which hotel has a gym?" could be considered as "what if the system takes into account that 'gym' is an important feature to me?").

These differences could be explained by the context of the task to be performed, and the type of users involved (general public vs. AI experts). However, it was somehow surprising to us that all but one of the questions referred to the system itself, its algorithm, or the input used for predictions. We believe that this may have been due to:

- Users might have perceived that the recommendations matched their preferences and that they had generally positive opinions, i.e., they did not receive very strange recommendations that raised their suspicions.
- Decisions in the chosen domain (hotels) are not as sensitive as in the medical or credit lending domains, where understanding the system logic or input influencing the prediction is critical to the acceptance of the system arguments.
- The perspective and opinion of others might be more relevant than details about their own inferred profile, as reported by [21] for opinionated explanations in a hotel RS, which seems to be the case when evaluating experience goods like hotels, a process characterized by a greater reliance on word-of-mouth [27, 49]. .

In regard to the *scope* dimension, we observed a dominance of single item questions. Although some authors consider that explanations mainly respond to contrast questions ("why P rather than Q?") [23, 35, 43], we observed that the comparative questions with

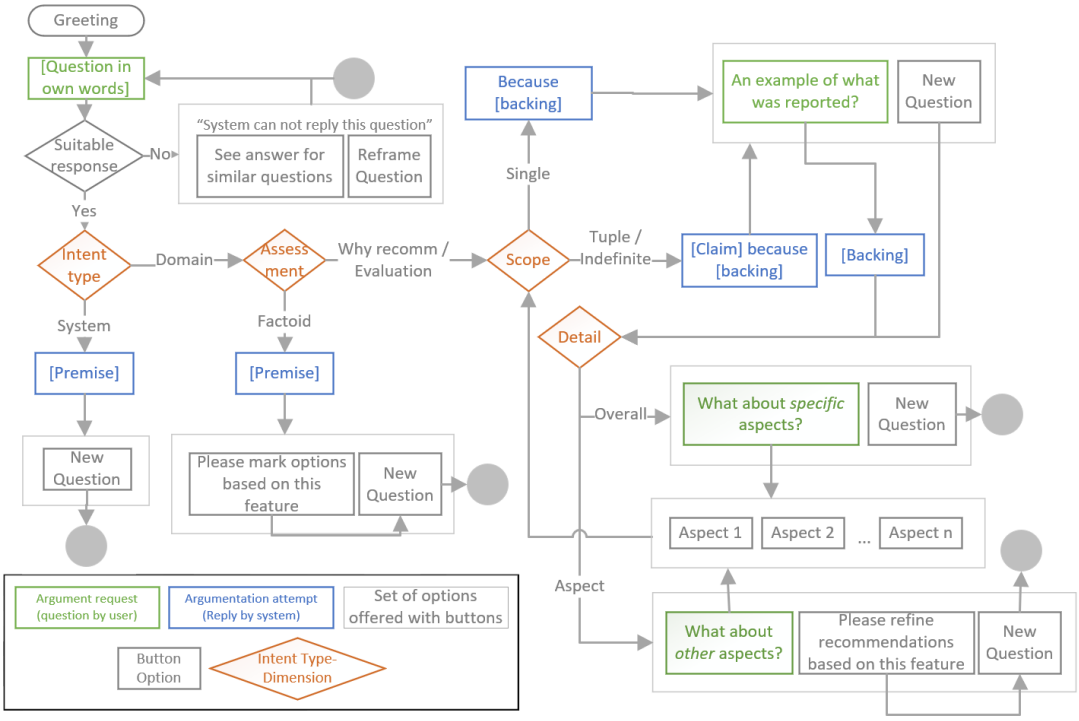

Figure 8: Proposed dialog management policy for conversational explanations in review-based RS

non-explicit items to be compared (indefinite) clearly outnumber those that do make them explicit (tuple). This involves an important implementation challenge, given that the methods proposed to answer this type of questions, and which are capable of processing several opinions on multiple items, are very scarce [44]. Furthermore, it should be noted that many of these questions also did not indicate specific features for evaluation (the fourth most frequent type of intention, see Table 2), so not only calculating the answer is challenging, but also how to communicate it briefly.

In this regard, we observed that while most of the questions were aspect-based utterances, an important portion also asked for overall assessment of the hotel(s). Here, an adequate balance must be maintained between relevance of the response (information about user's preferred features should be provided) and brevity. Guidelines from [46] recommend responding with concise utterances in the first place, and then enable the possibility to expand the information if needed, which could be facilitated by providing the option to choose specific aspects to dive into further details. System could also use this implicit indication of preferences to recalculate recommendations, as discussed for factoid questions.

Additionally, as expected, users not only generated standalone questions, but also follow-up questions, which confirms our expectation that an interactive QA approach would be appropriate to keep track of context and previously referred entities. Although creating a system able to respond to all possible questions is yet unrealistic [46], we suggest acknowledging "the system cannot answer this question" when the exact request cannot be answered. However, we suggest enabling the option of getting a response on a related feature, provided that the questioned aspect is within a reasonable range of similarity to those addressed by the system (e.g. "criminality rate" is related to the aspect "safety").

Finally and based on our observations, we adapted the scheme for interactive RS explanations proposed by [22], and extended it to support conversational argumentation, as well as system actions that could be triggered during the conversation. Figure 8 shows the proposed dialog management policy.

7 CONCLUSION AND FUTURE WORK

In this paper we have explored the design of a dialog management policy for explanations as conversational argumentation in review-based RS, conducted a WoOz study to assess the types of questions users might ask, and modeled user intents that could be used for the implementation of an explanatory CA in a hotel RS.

While the results obtained allowed us to gain a first insight into the type of questions that users would ask in the context under study, we acknowledge that a larger sample of participants would allow us to establish with more certainty the range of possible questions and reactions to the system's responses by users. Thus, we plan as future work, to continue with the implementation of the proposed methods for both automatic recognition of intents and generation of responses, as well as the implementation of the proposed policy within a dialog system, so that conversations can be collected on a larger scale. The above would also allow us to assess users' perception of the proposed solution, as compared, for example, to RS with static, or interactive but non-conversational explanations.

ACKNOWLEDGMENTS

This work was funded by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI 18*. 1–18.

[2] Abdallah Arioua and Madalina Croitoru. 2015. Formalizing Explanatory Dialogues. *Scalable Uncertainty Management* (2015), 282–297.

[3] Abdallah Arioua, Madalina Croitoru, Laura Papaleo, Nathalie Pernelle, and Swan Rocher. 2016. On the Explanation of SameAs Statements Using Argumentation. *Scalable Uncertainty Management* (2016), 51–66. https://doi.org/10.1007/978-3-319-45856-4_4

[4] Roland Bader, Wolfgang Woerndl, Andreas Karitnig, and Gerhard Leitner. 2012. Designing an explanation interface for proactive recommendations in automotive scenarios. In *Proceedings of the 19th International Conference on User Modeling, Adaptation, and Personalization (UMAP'11)*. 92–104.

[5] Sabrina Barko-Sherif, David Elsweiler, and Morgan Harvey. 2020. Conversational Agents for Recipe Recommendation. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 73–82. https://doi.org/10.1145/3343413.3377967

[6] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 717–725.

[7] Jamal Bentahar, Bernard Moulin, and Micheline Belanger. 2010. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review* 33, 3 (2010), 211–259.

[8] Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing EMNLP*. 5480–5494. https://doi.org/10.18653/v1/2020.emnlp-main.442

[9] Marco De Boni and Suresh Manandhar. 2020. Implementing clarification dialogues in open domain question answering. *Natural Language Engineering* 11, 4 (2020), 343–361. https://doi.org/10.1017/S1351324905003682

[10] Dimitrios Buhalis and Emily Siaw Yen Cheng. 2020. Exploring the Use of Chatbots in Hotels: Technology Providers Perspective. *Information and Communication Technologies in Tourism* (2020), 231–242. https://doi.org/10.1007/978-3-030-36737-4_19

[11] Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. In *Artif. Intell.*, Vol. 170. 925–952.

[12] Li Chen and Pearl Pu. 2014. Critiquing-based recommenders: survey and emerging trends. 22, 1–2 (2014), 3085–3094.

[13] 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

[14] Oana Cocarascu, Antonio Rago, , and Francesca Toni. 2019. Extracting Dialogical Explanations for Review Aggregations with Argumentative Dialogical Agents. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*.

[15] Nils Dahlback, Arne Jonsson, and Lars Ahrenberg. 1993. Wizard of Oz Studies: Why and How. In *Proceedings of the 1st Int. Conference on Intelligent User Interface*. 193–200.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2019).

[17] Rebecca Eynon, Chris Davies, and Wayne Holmes. 2012. Wizard of Oz for Multimodal Interfaces Design: Deployment Considerations. In *Proceedings of the 8th International Conference on Networked Learning*. 66–73.

[18] Miguel Grinberg. 2020. Socket.IO. (2020). https://github.com/miguelgrinberg/Flask-SocketIO

[19] Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. In *Computational Linguistics 43*, Vol. 1. 125–179.

[20] Diana C. Hernandez-Bocanegra, Tim Donkers, and Jürgen Ziegler. 2020. Effects of Argumentative Explanation Types on the Perception of Review-Based Recommendations. In *Adjunct Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20 Adjunct)*.

[21] Diana C Hernandez-Bocanegra and Juergen Ziegler. 2020. Explaining Review-Based Recommendations: Effects of Profile Transparency, Presentation Style and User Characteristics. *Journal of Interactive Media* 19, 3 (2020), 81–200. https://doi.org/10.1515/icom-2020-0021

[22] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Effects of interactivity and presentation on review-based explanations for recommendations. In *arXiv:2105.11794*. http://arxiv.org/abs/2105.11794

[23] Denis J. Hilton. 1990. Conversational processes and causal explanation. 107, 1 (1990), 65–81.

[24] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2020. A Survey on Conversational Recommender Systems. 1–35. https://doi.org/abs/2004.00646

[25] Nitin Jindal and Bing Liu. 2006. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 06*. 244–251. https://doi.org/10.1145/1148170.1148215

[26] John F. Kelley. 1984. An Iterative Design Methodology for User-Friendly Natural Language Information Applications. In *Transactions on Office Information Systems*, Vol. 2. 26–41.

[27] Lisa Klein. 1998. Evaluating the Potential of InteractiveMedia through a New Lens: Search versus Experience Goods. In *Journal of Business Research*, Vol. 41. 195–203.

[28] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the User Experience of Recommender Systems. In *User Modeling and User-Adapted Interaction*. 441–504.

[29] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of 24th International Conference on Intelligent User Interfaces (IUI 19)*. ACM, 379–390.

[30] Béatrice Lamche, Ugur Adigüzel, and Wolfgang Wörndl. 2012. Interactive explanations in mobile shopping recommender systems. In *Proceedings of the 4th International Workshop on Personalization Approaches in Learning Environments (PALE'14), held in conjunction with the 22nd International Conference on User Modeling, Adaptation, and Personalization (UMAP'14)*. 92–104.

[31] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. Klagenfurt, Germany: SSOAR.

[32] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *32nd Conference on Neural Information Processing Systems, NeurIPS 2018*. 9725–9735.

[33] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* 9042 (2020), 1–15. https://doi.org/10.1145/3313831.3376590

[34] Nathalie Rose Lim, Patrick Saint-Dizier, , and Rachel Roxas. 2009. Some challenges in the design of comparative and evaluative question answering systems. In *In Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions - KRAQ 09*. 15–18. https://doi.org/10.3115/1697288.1697292

[35] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement* 27 (1990), 247–266.

[36] Yuping Liu and L J Shrum. 2002. What Is Interactivity and Is It Always Such a Good Thing? Implications of Definition, Person, and Situation for the Influence of Interactivity on Advertising Effectiveness. *Journal of Advertising* 31, 4 (2002), 53–64.

[37] Benedikt Loepp, Katja Herrmanny, and Juergen Ziegler. 2015. Blended Recommending: Integrating Interactive Information Filtering and Algorithmic Recommender Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI 15*. 975–984.

[38] Benedikt Loepp, Tim Hussein, and Juergen Ziegler. 2014. Choice-based preference elicitation for collaborative filtering recommender systems. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI 14*. 3085–3094.

[39] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2019*. 1–9.

[40] Bella Martin and Bruce Hanington. 2012. *Universal Methods of Design*. Rockport Publishers, Beverly, MA.

[41] Philipp Mayring. 2014. Qualitative Content Analysis: Theoretical Foundation, Basic Procedures and Software Solution. (2014). Klagenfurt, Germany: SSOAR.

[42] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. In *Information Systems Research*, Vol. 13.

[43] Tim Miller. 2018. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* (2018).

[44] Amit Mishra and Sanjay Kumar Jain. 2015. An Approach for Sentiment analysis of Complex Comparative Opinion Why Type Questions Asked on Product Review Sites. *Computational Linguistics and Intelligent Text Processing Springer LNCS* 9042 (2015), 257–271.

[45] Christof Monz. 2003. Document Retrieval in the Context of Question Answering. In *Proceedings of the 25th European conference on IR research*. 571–579.

[46] Robert J. Moore and Raphael Arar. 2018. Conversational UX Design: An Introduction. *Studies in Conversational UX Design* (2018), 1–16. https://doi.org/10.1007/978-3-319-95579-7_1 Springer International Publishing.

[47] Emanuela Moreale and Maria Vargas-Vera. 2004. A Question-Answering System Using Argumentation. *MICAI 2004: Advances in Artificial Intelligence* (2004), 400–409. https://doi.org/10.1007/978-3-540-24694-7_41

[48] Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. In *Intelligent User Interfaces (IUI 16)*, Vol. 2. 256–260.

[49] Philip J. Nelson. 1981. Consumer Information and Advertising. In *Economics of Information*. 42–77.

[50] Hans Dybkjaer Niels Ole Bernsen and Laila Dybkjaer. 1993. Wizard of Oz prototyping: How and when. In *In CCI Working Papers in Cognitive Science and HCI*.

[51] 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*. 135–143. https://doi.org/10.1145/3349537.3351899

[52] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems - RecSys 11*. 157–164.

[53] Silvia Quarteroni and Suresh Manandhar. 2008. Designing an interactive open-domain question answering system. *Natural Language Engineering* 15, 1 (2008), 73–95. https://doi.org/10.1017/S1351324908004919

[54] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, and Francesca Toni. 2020. Argumentation as a Framework for Interactive Explanations for Recommendations. In *Proceedings of the Seventeenth International Conference on Principles of Knowledge Representation and Reasoning*. 805–815.

[55] Andrew I Schein, Alexandrin Popescul, Lyle H Ungar, and David M Pennock. 2002. Methods and Metrics for Cold-Start Recommendations. In *Proceedings of SIGIR 2002*. 253–260.

[56] Kacper Sokol and Peter Flach. 2020. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. 34, 2 (2020), 235–250.

[57] Ronnie Taib and Natalie Ruiz. 2007. Wizard of Oz for Multimodal Interfaces Design: Deployment Considerations. In *Human-Computer Interaction. Interaction Design and Usability*. 232–241.

[58] Nava Tintarev. 2007. Explanations of recommendations. *Proceedings of the 2007 ACM conference on Recommender systems, RecSys 07* (2007), 203–206.

[59] Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In *15th International Conference on Intelligent Text Processing and Computational Linguistics*. 115–127.

[60] Douglas Walton. 2000. The Place of Dialogue Theory in Logic, Computer Science and Communication Studies. 123 (2000), 327–346.

[61] Douglas Walton. 2004. A new dialectical theory of explanation. 7, 1 (2004), 71–89.

[62] Douglas Walton. 2011. A dialogue system specification for explanation. 182, 3 (2011), 349–374.

[63] Douglas Walton and Erik C. W. Krabbe. 1995. *Commitment in dialogue: Basic concepts of interpersonal reasoning*. State University of New York Press, New York.

[64] Nan Wang, Hongning Wang, Yiling Jia, , and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 18*. 165–174.

[65] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. 62, 6 (2019), 70–79.

[66] Markus Zanker and Martin Schoberegger. 2014. An empirical study on the persuasiveness of fact-based explanations for recommender systems. In *Joint Workshop on Interfaces and Human Decision Making in Recommender Systems*. 33–36.

[67] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards Conversational Search and Recommendation: System Ask, User Respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 177–186. https://doi.org/10.1145/3269206.3271776

[68] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval*. 83–92.