Conversational Recommendation based on End-to-end Learning: How Far Are We?

Ahtsham Manzoor and Dietmar Jannach

University of Klagenfurt, Austria
email: ahtsham.manzoor@aau.at, dietmar.jannach@aau.at

Abstract

Conversational recommender systems (CRS) are software agents that support users in their decision-making process in an interactive way. While such systems were traditionally mostly manually engineered, recent works increasingly rely on machine learning models that are trained on larger corpora of recorded recommendation dialogues between humans. One promise of such end-to-end learning approaches therefore is that they avoid the knowledge-engineering bottlenecks of traditional systems. Recent empirical evaluations of such learning-based systems sometimes demonstrate continuous progress *relative* to previous systems. Therefore, it may not be entirely clear how usable these systems are on an *absolute* scale. To address this research question, we evaluated two recent end-to-end learning approaches presented at top-tier scientific conferences with the help of human judges. A first study showed that in both investigated systems about one third of the system responses were not considered meaningful in the given dialogue context, which questions the applicability of these systems in practice. In a second study, we benchmarked the two systems against a trivial rule-based approach, again with human judges. In this second study, the participants considered the quality of the responses of the rule-based approach significantly better on average than those of the learning-based systems. Overall, besides pointing to open challenges of state-of-the-art learning-based approaches, our studies indicate that we must improve our evaluation methodology for CRS to ensure progress in this field.[1]

Keywords: Conversational Recommender Systems; Evaluation; End-to-end Learning,

1. Introduction

Conversational Recommender Systems (CRS) are software agents that interactively support users in their information search or decision-making processes.

[1]This work significantly extends a previously published workshop paper [1].
This paper to appear in Computers in Human Behavior Reports, 2021.

In such systems, the recommendation process is usually structured as a multi-turn dialogue [2], where the system tries to elicit the user's needs and preferences, makes recommendations, possibly provides explanations, and processes the users' feedback on the recommendations. From the interaction perspective, such systems are therefore much richer than the *one-shot* recommendations that we receive today on e-commerce or media streaming sites.

Research on online CRS goes back to the mid-1990s. From a technical perspective, most early approaches had a web-based user interface with forms and buttons, and the required dialogue and recommendation logic was explicitly encoded in the system. Such *engineered* solutions typically relied on case-based reasoning, critiquing techniques, or constraint-based approaches [3, 4, 5, 6, 7, 8]. Only some of the earlier systems also aimed at offering a natural language interface [9]. These approaches were however hampered by the available natural language processing (NLP) technology of that time.

In recent years, we have observed an increasing interest in CRS, both in academia and industry. These developments are spurred by major technological advances in the area of natural language processing and the rise of deep learning. Furthermore, with the spread of chatbots and digital assistants like Apple's Siri, we are more and more used to interact with virtual agents in natural language. This change in the interaction *modality*—from forms and buttons to voice and written language—was accompanied by a change in the way the systems are built. Instead of relying on manually engineered knowledge, most recent systems adopt a *learning-based* approach, where the system aims to learn from data how to respond to actions and utterances by the users.

Learning in such systems can be limited to certain aspects, e.g., when the goal is to learn which question to ask next [10, 11]. However, some recent approaches also pursue the vision of an *end-to-end* learning system, where the amount of engineered knowledge is reduced to a minimum. Internally, such systems rely on machine learning models that are trained on a large corpus of recorded natural language dialogues.

Recent examples of such approaches are the DeepCRS [12] and the KBRD [13] system, which both rely on complex deep learning models to generate suitable responses to a user utterance in a given dialogue situation. Both systems were empirically evaluated also with the help of human judges, and in both cases the authors observed that their newly proposed system was considered favorably over existing ones. However, since these human judgments, at least in the case of DeepCRS, were *relative* statements regarding the response quality, it remains unclear if the generated responses would be considered meaningful, on average, on an *absolute* scale. If a system fails too often to provide a meaningful response, its usability in practice might be very limited, even when it outperforms previous systems.

The goal of our present work is to close this research gap and to shed light on the state-of-the-art in end-to-end learning approaches to conversational recommendation. To that purpose, we conducted two studies. In *Study-1*, we analyzed the responses generated by DeepCRS and KBRD on an absolute scale. Specifically, we asked human annotators to assess if the system responses represented

meaningful continuations of the dialogue and if the recommendations were plausible. Furthermore, we programmatically analyzed to what extent the systems generated new sentences, i.e., sentences that did not appear in the training data. In *Study 2*, our goal was to assess through a user study how the two recent learning-based systems compare to a basic engineered solution consisting of a dozen of simple intent detection rules and predefined answer templates.

Overall, the studies indicate that today's end-to-end learning solutions might make too many mistakes to be immediately usable in practice, even when used in relatively restricted dialogue situations. Specifically, *Study-1* revealed that at least one third of the system responses were not considered meaningful by the annotators. Furthermore, *Study-2* indicated that a very limited amount of manually encoded dialogue knowledge is sufficient to create a CRS that compares favorably to the learning-based systems. One main implication of our work therefore is that the way we currently evaluate end-to-end learning-based CRS has its limitations, and that additional evaluation approaches are needed to ensure progress in this area.

The paper is organized as follows. Next, in Section 2, we discuss DeepCRS and KBRD in more detail, and provide more background on the evaluation of CRS. Section 3 and Section 4 report details about *Study-1* and *Study-2* respectively. In Section 5, we finally discuss the implications of our research.

2. Previous Work

In this section, we will first discuss technical details of DeepCRS and KBRD, which are representatives of end-to-end learning systems investigated in this paper. Afterwards, we will briefly review current evaluation practices for conversational recommender systems.

2.1. Architecture of DeepCRS and KBRD

DeepCRS [12] consists of multiple components. To represent the observed dialogue acts (utterances), a hierarchical recurrent encoder based on HRED [14] is used. A switching mechanism based on [15] connects the decoder with the recommendation module. After every dialogue act, the system checks if a movie entity is included. If this is the case, an RNN module is instantiated for each movie, which is responsible for classifying the user's sentiment or feedback on that entity. The outcomes of the sentiment analyses are then used as input to an autoencoder-based recommendation module, discussed in [16], which was pre-trained on MovieLens data. Based on the sentiments for the movies, the recommendation module computes and provides recommendation(s). Overall, the entire process of generating the dialogue utterances, and thereby the recommendations, is learning-based. In terms of additional knowledge besides the recorded dialogues and a list of movie titles, the system only leverages community ratings from MovieLens for the recommendation task.

KBRD [13] is also a multi-component system. Instead of HRED, the authors rely on the encoder-decoder Transformer framework [17] based on sequence-to-sequence learning. This is done because Transformer has shown to perform

better than HRED in several tasks such as machine translation [17, 18], natural language generation [19, 20], and question answering (Q&A) [21, 22]. Differently from DeepCRS, KBRD's recommendation module is not only based on titles of movies, but also uses an external knowledge graph from DBpedia [23]. The structural and relational information contained in the knowledge graph is encoded into entity hidden representations using Relational Graph Convolutional Networks. Both the mentioned entities in the dialogue history and the extracted features (e.g., fantasy, sci-fi, etc.) are provided as input to the recommender to compute and provide recommendation(s). The intuition of the recommendation approach is to retrieve items that are close to the concepts that are mentioned in the seeker utterances.

2.2. Evaluation of CRS

The authors of DeepCRS analyze their system in different dimensions. First, they assess the performance of the sentiment analysis component; second, they make an offline experiment to compute the accuracy of the predictions of the recommender on a recent MovieLens dataset; third, they run a user study for overall dialogue quality assessment. In the user study, 10 judges are asked to each rank all recommendation utterances in 10 complete dialogues (560 rankings in total). Specifically, the task was to rank the quality of (a) the true human utterance, (b) the utterance provided by DeepCRS, and (c) the utterance created by HRED. While the human answer was considered best on average, the results showed that DeepCRS was generally better ranked than HRED.

KBRD was also evaluated both in an automatic way and with the help of humans. In the automatic evaluation, *perplexity* and *distinct n-gram* were used to assess the linguistic quality of the generated utterances. *Recall* was used to assess recommendation quality. In the human evaluation, ten annotators with knowledge in linguistics were asked to score the system responses in 100 randomly sampled dialogues. Specifically, they were asked to score the system responses with respect to the consistency with the dialogue history. In this comparison, the responses by DeepCRS (termed ReDial in their paper) were used as a baseline. The results show that KBRD's average score (1.99 on a 1-3 scale) was better than DeepCRS, which had a score of 1.73.

Generally, CRS can be evaluated in several dimensions [2]. Similar to a non-interactive recommender system, we can assess their *effectiveness of task support*. With the help of a user study, we can for example investigate if the system helped users to find relevant items, if it is generally considered useful, and if users have the intention to use the system again in the future [24, 25, 26]. Specific aspects such as the quality of the recommendations can also be investigated through offline studies. Second, we can study the *efficiency of the task support*, e.g., by measuring the number of required dialogue turns until the user accepts a recommendation or by asking study participants regarding the perceived effort when using the system [27, 28]. Third, we can assess the *quality of the conversation* as well as usability aspects. The quality of the conversations can for example be assessed both with computational linguistic metrics or in terms of the subjective quality perceptions of users. Finally, we

can evaluate individual subcomponents, like the ones responsible for intent or entity recognition or for sentiment analysis [29, 30, 31].

The authors of DeepCRS and KBRD relied on several of these evaluation approaches, and they relied both on computational measures (e.g., for recommendation quality) and subjective measures (when asking judges to rank or score system utterances). In terms of the human evaluation, DeepCRS was however only assessed in a relative way, i.e., compared to the quality of the HRED system. In the evaluation of the KBRD system, on the other hand, the quality of the recommendations was actually not considered explicitly in the human evaluation. The average score given by the human judges in the case of KBRD was about 2 on a 1-3 scale. While this might be generally good, the reported average number cannot tell us in how many cases and in which situations the system failed to produce a reasonable answer.

In this present work, our goal is to close this research gap. Through the studies that we present next, we try to understand the mentioned aspects, e.g., how often and in which situations do the systems fail, in more detail. Furthermore, we investigate how an entirely manually engineered system—as the opposite of an end-to-end learning system–would fare in a comparison.

3. Study-1: Independent Evaluation of DeepCRS and KBRD

In *Study-1*, our main goal is to understand how good the responses and recommendations of DeepCRS and KBRD are on an absolute scale, and not relative to a baseline system of unclear quality. Furthermore, since in one of the original papers [13], also computational measures were applied to assess the linguistic quality of the generated responses, we are interested to know to what extent these generated-responses are different from utterances that appeared in the training data.

3.1. Experiment Design

In the study, we relied both on human judgments and automated processes for our analyses. We analyzed the following three aspects in detail:

(1) The meaningfulness of system responses in the given context.

(2) The quality of the recommendations.

(3) The originality of the responses with respect to the training data.

We denote our analyses regarding aspects (1) and (2) as *Study-1a* and the analysis of aspect (3) as *Study-1b*.

Study Material. As a study material, we used 70 dialogues, which we randomly sampled from the ReDial dataset[2]. We then applied DeepCRS and KBRD

[2]We share all study material online to ensure reproducibility of our work: `https://github.com/ahtsham58/rbCRS-chbr`

to generate responses to the *seeker* requests in these dialogues. Through this process, we obtained 758 system responses, 399 by DeepCRS and 359 by KBRD. This discrepancy between the systems is caused by the fact that KBRD did not always return a response.[3]

To create the responses, we used the code provided by the authors of DeepCRS and KBRD. We fed the original human-to-human dialogues into the system and extended them in a way that they not only contain the responses by the human recommender, but also the responses generated by the learning system. Table 1 shows a fragment of the result of the response generation process for one dialogue when using DeepCRS.

...	
SEEKER:	2001 : a space odyssey might be a great option. anything else that you would suggest ?
HUMAN:	you can try planet of the apes the older one is quite suspenseful and family friendly .
DeepCRS:	star wars : the force awakens is also a good one return of the jedi all good movies

Table 1: Verbatim output of the Response Generation Process (DeepCRS Example).

Experiment Procedure (Study-1a). In *Study-1a*, which aimed to answer the questions regarding the meaningfulness of the responses and the quality of the evaluations, we relied on three human judges. The judges were provided with the previously created material, i.e., with files containing the dialogues as shown in Table 1, and they were tasked to annotate each system response on a binary scale. Specifically, they were instructed to label each system response as being meaningful or not, given the original human-to-human dialogue situation up to that response. In case the system response contained a recommendation—remember that some responses might be questions to the seeker or simply chit-chat—the judges had to state if the recommendation is meaningful or not, given the stated preferences by the seeker so far. Finally, the judges were asked to indicate which of the responses contained *phatic* expressions, i.e., chit-chat.

In order to not bias the judges, we did not provide specific instructions to them regarding what makes a response or recommendation meaningful. With respect to the assessment of the recommendations, they were asked to look up the mentioned movies they did not know on the website IMDb.com.

All three judges were doctoral students in computer science in Austrian universities. One was one of the authors of this paper, the other two are working on different areas and have no previous experience regarding conversational

[3]We contacted the authors regarding this topic, and we were informed that when the model does not learn anything new compared to the previous utterance, it does not generate any response.

recommender systems. The responses of each system, DeepCRS and KBRD, were evaluated independently by exactly two judges and the second opinion, for each system, is obtained by one of the authors. Despite the fact that we did not provide detailed instructions on how to interpret the term "meaningful", the reviewer agreement was generally very high, with 92.73% for DeepCRS and 95.82% for KBRD.

Experiment Procedure (Study-1b). For the assessment of the originality of the generated responses, we wrote a computer program, which compared the responses returned by each system with the training data. The program counted both exact matches as well as sentences that were found in almost identical form in the training data. We used a very strict interpretation of what "almost identical" means. Specifically, to be considered as almost identical, a generated response and a sentence in the training data had to consist of the exact same set of words, and these words had to appear with the same frequency in the sentence. In other words, in almost identical sentences, only the order of some words or segments was usually exchanged.

3.2. Results

Study-1a. The results of *Study-1a* are shown in Table 2. The numbers in the table represent the averages of the two annotators for each system. Note that the labeling exercise was done independently by the two annotators for each system before we computed the averages. Remember also that the agreement between reviewers was very high.

- Looking at how the generated sentences were labeled, we observe that 31% (DeepCRS) and 42% (KBRD) of the system responses were *not* considered meaningful by the judges. When the system response was generated in the context of chit-chat, the systems were more successful, with 85% (DeepCRS) and 87% (KBRD) of the responses being considered meaningful. Overall, however, only 5 out of 70 dialogues (7%) were free of problems for both systems.

- In terms of the quality of the recommendations, 40% (DeepCRS) and 45% (KBRD) of them were *not* considered meaningful. In 36% (DeepCRS) and 28% (KBRD) of the dialogues, not one meaningful recommendation was observed. Moreover, in 10% and 8.5% of the total dialogues, no recommendation was made at all by DeepCRS and KBRD respectively.

Overall, while the systems could respond in a meaningful way in the majority of cases, the results in some ways appear a bit disappointing, even more so because the works were published at highly selective conferences. As a result, the question arises if these systems could be safely deployed in practice. Despite the quite complex machine learning models and the relatively large training dataset, both systems failed to react properly in many dialogue situations. Various types of issues occurred, including not being able to understand the user intent, repeated questions, abrupt endings of the conversation, or bad recommendations.

	DeepCRS	KBRD
Number of dialogues	70	70
Generated sentences (overall)	399	359
Sentences labeled as meaningful	277 (69%)	209 (58%)
Sentences labeled as *not* meaningful	122 (31%)	150 (42%)
Dialogues without problems	5	5
Chit-chat sentences	132	88
Chit-chat labeled as meaningful	112 (85%)	77 (87%)
Number of recommendations	106	119
Recs. labeled as meaningful	63 (60%)	66 (55%)
Nb. dialogues with no meaningful recs.	25 (36%)	20 (28%)
Nb. dialogues with no rec. made.	7 (10%)	6 (8.5%)

Table 2: Analysis of Dialogue and Recommendation Quality

The success rates in the chit-chat context may appear promising. However, many of the system responses are rather trivial, like returning a greeting or saying goodbye after a response to the seeker after a recommendation.

In terms of the subjective quality assessments of the recommendations, we were expecting that we would more often observe disagreement between the judges. It turned out, however, that the agreement was also very high in this dimension, with 93% for DeepCRS and 96% for KBRD. Looking at the evaluations, we found that many bad recommendations were easy to agree on, for example when the system recommends "The Secret Life of Pets", an animated comedy film, after the seeker mentioned that s/he liked "Avengers: Infinity War", a superhero film. Also, the judges sometimes found repeated recommendations in a single dialogue, which were considered bad recommendations as well.

Study-1b. The statistics for *Study-1b* on the originality of the generated responses are shown in Table 3. The observations can be summarized as follows.

- First of all, it is apparent that both systems respond with identical sentences in several dialogues. DeepCRS for example only uses 46 different sentences when generating its 399 responses.
- More importantly, however, is that both systems mostly return one of the existing sentences from the training data, i.e., they almost never create a new sentence. DeepCRS even exclusively returns responses that were previously found in the data. KBRD, on the other hand, while creating variations of training sentences sometimes, in 11 cases also returns broken sentences.

Overall, both systems rather share the characteristics of retrieval-based approaches than of language generation based ones. While this is not a problem

	DeepCRS	KBRD
Generated sentences	399	359
Unique sentences	46	159
Identical in training data	44	87
Almost identical in training data	2	59
New sentences	0	5
Broken sentences	0	11

Table 3: Characteristics of Generated Sentences

per se, it seems questionable to apply metrics in the evaluation that assess linguistic aspects of the returned responses. Since both systems almost exclusively return sentences that were originally written or spoken by humans, the metrics do not measure any quality that is specific to the CRS implementation.

4. Study-2: Comparison with a Minimal Rule-Based CRS

The goal of *Study-2* was to explore which level of quality can be achieved compared to DeepCRS and KBRD when using a minimal rule-based CRS. This comparison therefore involves two extremes of building a CRS. On the one hand, we have two advanced learning-based systems that do not require any knowledge engineering and where the quality of the system should improve when more data is available. On the other hand, we have an entirely static system, which is built on domain knowledge. We are aware of the fact that such a comparison comes with limitations, and we make no claims regarding which approach is better in general.[4] In contrast, the main question we seek to answer with *Study-2* is how much explicitly encoded dialogue knowledge is needed to achieve a quality perception that matches those of the studied end-to-end learning systems.

4.1. Experiment Design

In the first part of *Study-2*, termed *Study-2a*, we conducted an online user study (N=58), where we asked participants to score the responses of three different CRS (DeepCRS, KBRD, our rule-based CRS) in a given dialogue situation. Afterwards, in *Study-2b* we compared in more detail in which situations the different CRS failed. Also for this study, we used a set of randomly selected 70 dialogues.[5].

Study-2a. Previously, in *Study-1a*, we were analyzing, on a binary scale, how many of the responses of two learning-based systems were considered meaningful by two human judges. In the follow-up study *Study-2a* presented here, our

[4] We also do not explore any hybrid system here, which might be able to combine the best of both worlds.

[5] These dialogues were thus not the same as the ones used in *Study-1*

aims were to *(i)* obtain more fine-grained feedback on the quality of the system responses and to *(ii)* investigate how a minimal rule-based CRS, dubbed RBCRS from here on, compares to the learning-based systems. Moreover, in *Study-2a* we were involving a larger set of human judges, where each subject assessed a certain subset of the dialogue situations.

Study-2a was conducted as an online study with a specifically designed web application. The tasks for the participants were as follows. In a first step, the participants were informed about their task: assessing the responses by three different CRS and answering questions in a follow-up questionnaire. After obtaining informed consent, participants were forwarded to a screen as shown in Figure 1.

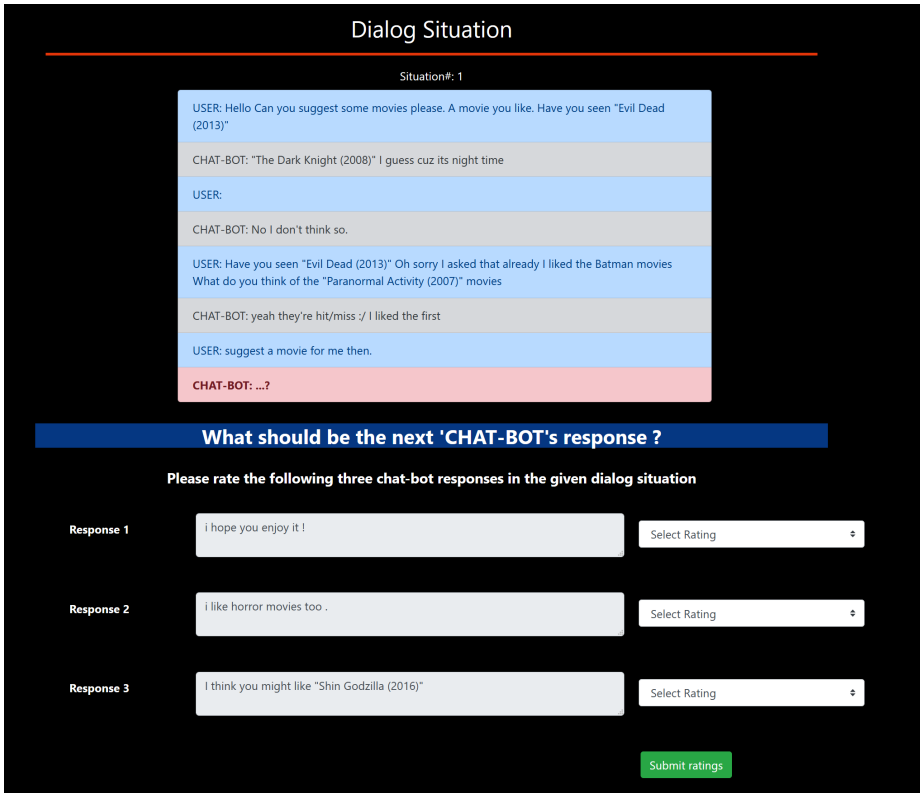

Figure 1: Response Rating User Interface in *Study-2a*

On this screen, the participants were presented with a dialogue situation selected from the 70 previously determined dialogues. The specific situation in each dialogue was also randomly selected to make sure that we collect feedback by the participants for different stages in a dialogue. Below the dialogue fragment, the responses to the last seeker utterance, as produced by DeepCRS, KBRD, and RBCRS were presented in random order. Participants then had to

provide a score for each response individually using a five-item response scale that ranged from "entirely meaningless" to "perfectly meaningful". This scoring task was repeated 10 times, i.e., each participant rated responses for 10 dialogue situations. However, one of the dialogue situations was used as an attention check and was discarded while analyzing results. Afterwards, users were asked a number of questions regarding *(i)* their demographics and experience with chatbots and *(ii)* their impressions regarding the quality of the generated responses in general. The exact questions regarding the latter point are shown in Table 4. The answer options ranged from "strongly disagree" to "strongly agree".

	Questions
Q1	I found the presented dialogues natural.
Q2	The presented dialogue situations look realistic.
Q3	I could imagine that such dialogues also happen between humans.
Q4	Considering only the best responses found in each dialogue, I would find the chatbot useful.
Q5	Considering only the best responses found in each dialogue, I would probably use such a movie recommendation chatbot in the future

Table 4: Questions from the Post-Task Questionnaire

Study-2b. Given the data collected in *Study-2a*, we aimed to investigate the extent of three typical types of errors that we observed in some preliminary experiments with the learning-based CRS.

- First, there are two situations where the CRS are apparently not able to properly maintain the history of the ongoing dialogues: *(i)* when they repeat a previous response and *(ii)* when they make the same recommendation twice or even more often in a dialogue. Note that such problems can easily go unnoticed in *Study-2a* when the human judges mainly focus on the last seeker utterance and system response in a given dialogue.
- Second, we observed that the CRS sometimes failed to properly respond to specific questions by a seeker such as "Is Reese Witherspoon starring in this movie?."

The assessment of the frequency of such situations was done by manual inspection of the dialogues used in *Study-2a*. Again, two human judges were involved, one being an author of this present paper and the other one a doctoral student in computer science not working on recommender systems.

4.2. Implementation Details of the Rule-Based CRS

The rule-based RBCRS system used in the study consists of two main components: the Recommendation Engine and the Response Selector.

The Recommendation Engine. The recommendation engine implements two similar-item retrieval strategies:

- The *latent factors* approach first applies matrix factorization once on the underlying movie rating dataset. Upon a recommendation request, it returns movies that are similar to the most recent *seeker-mentioned* movie. The similarity between the movies is computed based on their latent factor representations (embeddings), see also [32].
- The *genre-based* approach is used in situations where the system has detected that the seeker looks for movies of a certain genre. The retrieval of movies in this approach is based on matching the genres of the movies.

To avoid the recommendation of obscure items, see also [33], the Recommendation Engine applies a popularity filter to the results before returning a recommendation. Specifically, it was ensured that any recommended movie was among the 15% most popular movies in the underlying dataset.

The Response Selector. Depending on the assumed intent of the seeker, this component selects a response from a predefined catalog of templates. In case the response is a recommendation, it parameterizes the template with a movie title. The system used in *Study-2a* supports 5 main intents, see below; some of them have subintents. For each intent, there is exactly one corresponding system response [34]. For each system action, we selected a small set of textual alternatives from the ReDial dataset. When the response is determined, the system randomly picks one of these alternatives, which prevents us to use one single phrase over and over.

We list the set of system actions in Table 5. The selection of the response, which corresponds to intent detection, is based on simple keyword matching rules. Overall, the code for heuristically choosing the response consists of a few dozen if-statements in Python code. However, we make sure that any type of seeker query will be categorized in one of the defined intent types. Therefore, unlike the KBRD system, which sometimes fails to generate a response, our system always returns a response. Note that for the case of seeker questions, we do not actually try to provide answers in detail. Rather, our CRS either answers in a confirming and socially pleasing way, e.g., "Yes, you should try it out.", or it answers that it does not know the answer or has not seen a certain movie, in case one is mentioned.

4.3. Results

The results for *Study-2a*, in which we conduct a user study to compare different algorithms, and *Study-2b*, in which we analyze in which situations the compared systems fail, are as follows.

Study-2a. Overall, 137 subjects participated in the study and used the provided online application to score the responses generated by the three systems (DeepCRS, KBRD, rbCRS). We recruited 103 of them through Amazon Mechanical Turk, and 34 trough emails to personal contacts (or own *circle*).

System Action	Description
Chit-Chat	Three different types of chit-chat responses are supported, responding to greetings such as "How are you?", returning a goodbye, or responding to a "thank you" utterance.
Ask Preference	This action is usually taken after the seeker enters the conversation with a greeting.
Recommend	The system selects this action when the user asks for recommendations. The choice of the algorithm (genre or latent factor based) depends on whether a genre is detected or not.
Respond to Question	Different types of user questions are supported. The system tries to detect if *(a)* the seeker asked the recommender about the opinion for a movie or *(b)* if the seeker utterance ended with any other question. In the first case, the system responds positive and encouraging; in the second case the system answers in a way that it does not know.
Encourage User	This action is taken when the seeker seems to have settled on a choice.

Table 5: List of System Actions (Corresponding to Supported User Intents) in rbCRS.

Of these 137 participants, 69 (around 50%) passed the attention check that we included in the study[6]. Moreover, to further ensure the reliability of the collected responses, we manually scanned the provided scores for problematic patterns that indicate limited attention. One problematic pattern, for example, was detected when two of the systems responded very similarly (e.g., "Hi" or "Hi there") but received completely different scores. To avoid cherry-picking of responses, we entirely removed participants who exhibited such patterns.[7] After this process, we ended up with 58 participants for which we are highly confident that they carefully accomplished the task. The summary statistics regarding the considered and discarded cases are shown in Table 6.

The detailed demographics of the participants are shown in Table 4. The average task completion time for the participants was about 11 minutes. The crowdworkers were compensated with 1.6 USD for their work.

After the experiment and the cleaning process, we had valid feedback for 522 dialogue situations, i.e., 9 per participant. Recall that one of the 10 responses was the attention check. The average scores and standard deviations are shown in Table 8. On average, the best scores were recorded for our minimal engineered system rbCRS. Also, the standard deviation is lowest for this method. The

[6]Specifically, in one of the 10 scoring tasks, we artificially created a non-meaningful system response. We considered the attention check failed whenever a participant rated the meaningless response with 4 or 5 on the 1-5 scale.

[7]The specific rules for removing participants and all cases that were ruled out this way are documented in the online material.

	Circle	Mturk	Total
Number of participants (cases)	34	103	137
Nb. participants passing the attention check	21	48	69
Nb. manually discarded cases	3	8	11
Nb. valid cases	18	40	58

Table 6: Statistics about Participants and Discarded Cases (*Study-2a*)

Demographic Feature	Scale	Circle	MTurk	Total
Gender	Male	14	27	41
	Female	4	13	17
Age	18-25	7	1	8
	25-30	7	12	19
	30-35	4	11	15
	35-45	0	7	7
	45-70	0	9	9
English fluency level	Beginner	0	1	1
	Intermediate	2	0	2
	Fluent	9	38	47
	Advanced	7	1	8
Education level	high school or less	0	9	9
	Bachelor's	6	22	28
	Master's	10	8	18
	Doctorate	2	0	2
	Other	7	1	8
Movie watching frequency	Everyday	3	7	10
	Several times a week	0	17	17
	Once in a week	7	11	18
	Once every few weeks	3	3	6
	Less frequent	5	2	7

Table 4: Demographic Details of Participants

differences between RBCRS and the second-ranked KBRD system is statistically significant according to a Student's t-test (p=0.002).

A histogram that visualizes the distribution of the scores is shown in Figure 2. The histogram shows the differences mainly come from the extreme values, where RBCRS more often receives very good scores and less frequently makes major mistakes which lead to the lowest score of 1.

Looking at the responses to the post-task questionnaire, we found that the participants generally thought positive of the presented dialogues and chatbot answers, see Table 9. Regarding the naturalness and realism of the dialogues (Q1–Q3), the average values (mean, median, and mode) are around or at 4, on a 1-5 scale. This result seems expected: on the one hand, the dialogues actually happened between users; on the other hand, the dialogues are not fully natural, because the crowdworkers received very specific instructions, e.g., regarding

	DeepCRS	KBRD	RBCRS
Average score	3.07	3.52	**3.77**
Std. deviation	1.47	1.40	**1.29**

Table 8: Results of Scoring of System Responses (*Study-2a*)

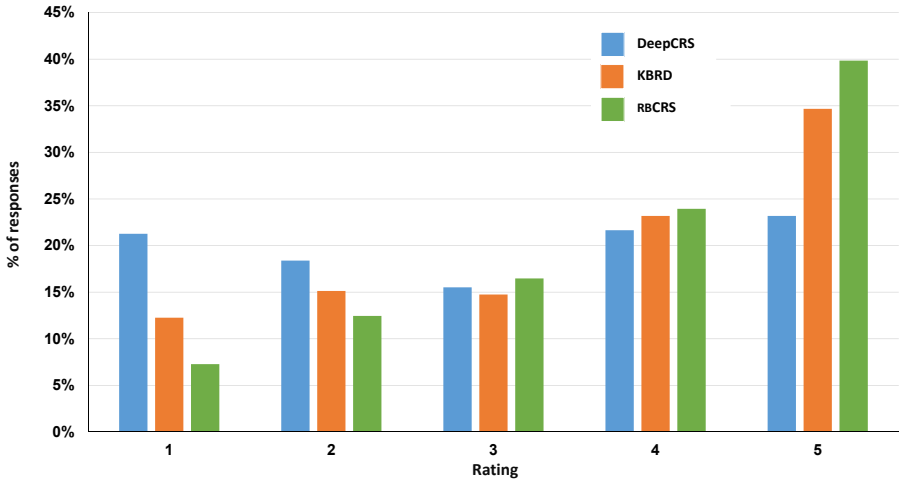

Figure 2: Distribution of Scores (*Study-2a*)

how many movies they should mention. Our interpretation therefore is that our study participants did notice that the dialogues are to some extent artificial. This points to a potential limitation of the ReDial dataset.

Regarding the perceived usefulness and the intention-to-reuse of the *best* CRS in the comparison (Q4 and Q5), the average responses were also encouraging, again with values around 4 for both questions. According to the results shown in Figure 2, the best system was RBCRS, which provided at least reasonable responses (rated with 3 or higher) in about 80% of the cases. However, remember that the study participants did not know which response was generated by which CRS, and the order of the presentation of the responses was randomized. Therefore, we cannot know to what extent the overall impression was influenced by each of the studied CRS.

Study-2b. In *Study-2b*, we finally look closer at the outcomes of *Study-2a*, aiming to identify in which situations the systems failed and if there are certain types of problems that can go unnoticed by the study participants when they focus too much on the last system response in a dialogue.

In our first analysis, we investigated how often each system made repetitive statements or repetitive recommendations in the *same dialogue*. As a background, we consider recommending the same movie twice in one conversation or repeating an utterance as problematic or at least detrimental to the perceived

Questions	Mean (Std)	Median	Mode
Q1 (Dialogue naturalness)	3.86 (0.54)	4	4
Q2 (Realism of dialogues I)	4.05 (0.57)	4	4
Q3 (Realism of dialogues II)	4.03 (0.79)	4	4
Q4 (Usefulness of best bot)	4.03 (0.74)	4	4
Q5 (Intention to re-use best bot)	3.78 (0.89)	4	4

Table 9: Post-task Questionnaire Statistics

dialogue quality. Table 10 shows that repetitive statements occur sometimes both for DeepCRS and KBRD, with a slightly higher frequency for KBRD. Note that the recommendation rules in RBCRS by design exclude movies that were already recommended in an ongoing dialogue. Generally, the number of repetitions in the end-to-end learning systems is not too high, given that we consider 522 dialogue situations. Nonetheless, it is an indication that both systems can have difficulties maintaining the state of the dialogue. In this context, remember that we involved two independent annotators in this process, like in *Study-1a*. The annotator agreement for the problematic cases was 100%.

	DeepCRS	KBRD
Repeated utterance	4	14
Repeated recommendation	19	32

Table 10: Within-Dialogue Repetitions in End-to-End Learning Systems

In the second analysis, we studied how often one of the learning systems did not return a proper sentence or did not generate a response at all. Table 11 shows the results. Again, note that RBCRS by design always returns a response from a predefined set. The annotator agreement was again 100%.

The results show that only KBRD sometimes has problems returning a valid response. Examples of broken responses are half-finished sentences or sentences that repeat individual statements more than once in a single utterance. Regarding the problem of KBRD not returning a response at all, we were in exchange with the authors of KBRD, as mentioned above. In 7 of the 70 dialogues the KBRD system apparently did not generate a response at the very end of the dialogue when there was nothing new to learn for the system (e.g., from a phatic utterance by the seeker). These missing final responses might in all observed cases not be too detrimental to the perceived system quality; sometimes, however, the dialogues end a bit abruptly when the system does not react anymore.

	DeepCRS	KBRD
Broken responses	0	17
No response	0	7

Table 11: Broken Responses or No Responses in End-to-End Learning Systems

Finally, we looked at how often the involved systems were not able to react to a specific question by the seeker, e.g., regarding the involvement of a certain actor in a movie. The results are shown in Table 12 and indicate the number of cases where *at least one annotator* considered a response not suited. All three systems, including ours, are not well prepared for such questions. The number of problems reported here are still relatively low, because such situations do not occur often in the ReDial dataset as mentioned above. There are basically two forms of problems that we observe. Either the system does not answer the seeker's question, probably because it went unnoticed, or it returns an unsuited general or evasive answer. In particular in this latter case, when the system responds with a generality, the annotators did disagree sometimes. A typical example for an annotator agreement is when the seeker asked a question about the content of a movie, e.g., "*What is the movie [movie name] about?*". Our system, might for example answer "*I have not seen it yet.*". In such cases, one annotator found the response appropriate, whereas the other found that it is not a satisfactory response.

	DeepCRS	**KBRD**	**RBCRS**
Failure to respond to question	18	13	15
Annotator agreement	15	11	9

Table 12: Number of Cases where System Failed to Respond to Seeker Questions.

5. Implications

In this section, we summarize the main findings and implications of our studies and discuss potential limitations of our research.

5.1. Implications

Generally, despite the mentioned limitations of today's systems, we believe that substantial progress was made in end-to-end learning approaches to conversational recommendation. More and more datasets are becoming available today, and the analyzed systems (DeepCRS and KBRD) could answer in a reasonable way in many situations despite the limitations of datasets like ReDial. Nonetheless, our analyses also show that building pure learning-based[8] conversational recommender systems remains difficult. We identify two main areas in which our research may have implications.

First, more work seems to be required to make the behavior more predictable and to ensure that the responses stay within certain guiding rails. This might be particularly important in certain application domains, e.g., when recommending high-involvement products in the banking or insurance domains, or when

[8]KBRD actually uses a structured item database in the background, but this database is only used for the purpose of finding matching items.

giving medical advice in a CRS. Moreover, in our analyses we found that certain problems could be relatively easily avoided, such as the repetition of responses and recommendations within a dialogue. Implementing such rules and heuristics however require the manual incorporation of additional knowledge, even though on a very general level.

The integration of additional sources of knowledge, e.g., about item features as done in KBRD, seems generally promising. Given the limitations of datasets like ReDial, it seems difficult to learn which actor was involved in which movie, as this information simply does not appear in the training data for most movies. The creation of richer datasets also include more social interactions like [35] and which can be important for effective communication [36] seems needed. Another opportunity for future work in this area could lie in the integration of general pre-trained language models like BERT or GPT-3. In [37], the authors recently analyzed what BERT knows about movies and items in other domains, and found that BERT is able to provide content-related information, e.g., about genres, in many cases.

A second implication of the analyses in this paper relates to how we evaluate CRS. Since any CRS is a software system designed to interact with users, its evaluation cannot be limited to offline experiments, except for particular subtasks like entity recognition. When using linguistic measures for offline evaluation, it furthermore has to be ensured that these computational measures are actually suitable proxies for human quality perceptions.

When involving humans, relative comparisons by human annotators seem to be not sufficient to assess the usability of today's CRS in practice. Also, average statistics, as often reported in papers, do not tell us enough about in which situation a CRS succeeds and where improvements are still required. Generally, scientific research in the area should inform about what has been achieved, but also about where research gaps exist.

5.2. Research Limitations

One first potential threat to the validity in studies like ours lies in the reliability of the study participants. Since we applied different measures to ensure that we do not include unreliable respondents, we are however confident that this risk is relatively low. In the end, we removed about half of the respondents due to our strict quality assurance measures. Note that most of the participants were frequent movie watchers, as illustrated in Table 4. For the human annotations in *Study-1a*, *Study-2b*, we involved two independently working judges for each task, and we can in general report high to very high annotator agreement.

Regarding Study-1a, one potential bias could come from the fact that the judges were able to see the human-recommender response along with the system-generated response they had to evaluate. Therefore, it might be possible that some judges compared the system response with the response of the human recommender and used the human response as a basis for their evaluation. However, we believe that this risk is modest, in particular as we observed high agreement between the two independent annotators.

In our studies, we have so far included two learning-based systems, and the question of course remains to what extent the findings of our study generalize to other approaches. We are however confident that the two systems are at least in some ways good representatives for current works on end-to-end learning. Both papers were published recently and were presented at high-quality conferences. Our study can be easily expanded to include additional approaches, as long as the code is made available by the authors and the models can be applied to the dialogues in the ReDial dataset. At the time being, we however have no knowledge about similar systems.

Our own rbCRS system is technically very simple and based on a few dozen if-statements, optimized for the types of dialogues we observed in the ReDial dataset. For dialogue datasets that contain richer conversations, it might be much more difficult to come up with a rule-based system. Moreover, our system, which is manually engineered for the given problem, will most probably not work well in other domains and maybe not even for other datasets of the same domain. In this context, note that we therefore in no way argue that writing a bunch of if-statements should be the future of building CRS or that static rule-based approaches should be considered as an alternative to *learning-based* approaches. We developed the system mainly to be able to assess how difficult it is to reach similar or better performance levels as the investigated learning-based systems.

6. Summary

The future of CRS lies in the increased use of learning-based systems, and in recent years an increased research interest in these types of systems were observed. In this paper, we have analyzed two recent end-to-end learning systems in terms of the quality of their system responses. In our first study, we found that these systems still fail in a significant number of cases to respond to the user in a meaningful way. In a second study, we were interested to understand how much explicit knowledge has to be encoded to obtain quality perceptions that are at least as good as those by the learning-based systems. The main conclusion of our studies are that *(i)* future systems should rely more often on a hybrid architecture that combines learning from data with explicit knowledge, and *(ii)* that user-centric evaluations are essential to ensure progress in conversational recommendation.

References

[1] D. Jannach, A. Manzoor, End-to-end learning for conversational recommendation: A long way to go?, in: IntRS Workshop at ACM RecSys 2020, Online, 2020.

[2] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Comput. Surv. 54 (5).

[3] R. D. Burke, K. J. Hammond, B. Yound, The FindMe approach to assisted browsing, IEEE Expert 12 (4) (1997) 32–40.

[4] R. Burke, The Wasabi Personal Shopper: a case-based recommender system, in: AAAI/IAAI, 1999, pp. 844–849.

[5] F. Ricci, Q. N. Nguyen, Acquiring and revising preferences in a critique-based mobile recommender system, Intelligent Systems 22 (3) (2007) 22–29.

[6] K. McCarthy, J. Reilly, L. McGinty, B. Smyth, On the dynamic generation of compound critiques in conversational recommender systems, in: AH '04, 2004, pp. 176–184.

[7] L. Chen, P. Pu, Preference-based organization interfaces: aiding user critiques in recommender systems, in: UM '07, 2007, pp. 77–86.

[8] D. Jannach, ADVISOR SUITE – A knowledge-based sales advisory system, in: ECAI '04, 2004, pp. 720–724.

[9] M. Göker, C. Thompson, The adaptive place advisor: A conversational recommendation system, in: Proceedings of the 8th German Workshop on Case Based Reasoning, 2000, pp. 187–198.

[10] K. Christakopoulou, F. Radlinski, K. Hofmann, Towards conversational recommender systems, in: KDD '16, 2016, pp. 815–824.

[11] T. Mahmood, F. Ricci, Improving recommender systems with adaptive conversational strategies, in: HT '09, 2009, pp. 73–82.

[12] R. Li, S. E. Kahou, H. Schulz, V. Michalski, L. Charlin, C. Pal, Towards deep conversational recommendations, in: NIPS '18, 2018, pp. 9725–9735.

[13] Q. Chen, J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, J. Tang, Towards knowledge-based recommender dialog system, in: EMNLP-IJCNLP '19, 2019, pp. 1803–1813.

[14] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, J.-Y. Nie, A hierarchical recurrent encoder-decoder for generative context-aware query suggestion, in: CIKM '15, 2015, p. 553–562.

[15] S. Subramanian, A. Trischler, Y. Bengio, C. J. Pal, Learning general purpose distributed sentence representations via large scale multi-task learning, in: ICLR '18, 2018.

[16] S. Sedhain, A. K. Menon, S. Sanner, L. Xie, Autorec: Autoencoders meet collaborative filtering, in: TheWebConf '15, 2015, pp. 111–112.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS '17, 2017, pp. 5998–6008.

[18] M. Ott, S. Edunov, D. Grangier, M. Auli, Scaling neural machine translation, in: WMT '18, 2018, pp. 1–9.

[19] P. J. Liu, M. Saleh, E. Pot, B. Goodrich, R. Sepassi, L. Kaiser, N. Shazeer, Generating wikipedia by summarizing long sequences, in: ICLR '18, 2018.

[20] Q. Chen, J. Lin, Y. Zhang, H. Yang, J. Zhou, J. Tang, Towards knowledge-based personalized product description generation in e-commerce, in: KDD '19, 2019, pp. 3040–3050.

[21] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: EMNLP, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392.

[22] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. W. Cohen, R. Salakhutdinov, C. D. Manning, HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering, in: EMNLP, 2018.

[23] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer, et al., DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia, Semantic web 6 (2) (2015) 167–195.

[24] Y. Jin, W. Cai, L. Chen, N. N. Htun, K. Verbert, MusicBot: Evaluating critiquing-based music recommenders with conversational interaction, in: CIKM '19, 2019, pp. 951–960.

[25] P. Wärnestål, User evaluation of a conversational recommender system, in: IJCAI '05 Workshop on Knowledge and Reasoning in Practical Dialogue Systems, 2005.

[26] D. Kang, A. Balakrishnan, P. Shah, P. Crook, Y.-L. Boureau, J. Weston, Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue, in: EMNLP-IJCNLP '19, 2019, pp. 1951–1961.

[27] N. Tintarev, J. Masthoff, Designing and evaluating explanations for recommender systems, in: Recommender systems handbook, Springer, 2011, pp. 479–510.

[28] Y. Sun, Y. Zhang, Conversational recommender system, in: SIGIR '18, 2018, pp. 235–244.

[29] L. Liao, R. Takanobu, Y. Ma, X. Yang, M. Huang, T.-S. Chua, Deep conversational recommender in travel, ArXiv abs/1907.00710.

[30] M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, M. Galley, A knowledge-grounded neural conversation model, in: AAAI, 2018.

[31] L. Nie, W. Wang, R. Hong, M. Wang, Q. Tian, Multimodal dialog system: Generating responses via adaptive decoders, in: MM '19, 2019, pp. 1098–1106.

[32] C. Trattner, D. Jannach, Learning to recommend similar items from human judgements, User Modeling and User-Adapted Interaction 30 (2020) 1–49.

[33] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, J. A. Konstan, User perception of differences in recommender algorithms, in: RecSys '14, 2014, pp. 161–168.

[34] W. Cai, L. Chen, Predicting user intents and satisfaction with dialogue-based conversational recommendations, in: UMAP '20, 2020, p. 33–42.

[35] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, Z. Yu, Inspired: Toward sociable recommendation dialog systems, in: EMNLP, 2020.

[36] P. Thomas, M. Czerwinski, D. McDuff, N. Craswell, Theories of conversation for conversational IR, in: International Workshop on Conversational Approaches to Information Retrieval, 2020.

[37] G. Penha, C. Hauff, What Does BERT Know about Books, Movies and Music? Probing BERT for Conversational Recommendation, in: RecSys '20, 2020, p. 388–397.