End-to-End Learning for Conversational Recommendation: A Long Way to Go?

Dietmar Jannach
University of Klagenfurt, Austria
dietmar.jannach@aau.at

Ahtsham Manzoor
University of Klagenfurt, Austria
ahtsham.manzoor@aau.at

ABSTRACT

Conversational Recommender Systems (CRS) have received increased interest in recent years due to advances in natural language processing and the wider use of voice-controlled smart assistants. One technical approach to build such systems is to learn, in an end-to-end way, from recorded dialogs between humans. Recent proposals rely on neural architectures for learning such models. These models are often evaluated both with the help of computational metrics and with the help of human annotators. In the latter case, the task of human judges may consist of assessing the utterances generated by the models, e.g., in terms of their consistency with previous dialog utterances.

However, such assessments may tell us not enough about the true usefulness of the resulting recommendation model, in particular when the judges only assess how good one model is compared to another. In this work, we therefore analyze the utterances generated by two recent end-to-end learning approaches for CRS on an absolute scale. Our initial analyses reveals that for each system about one third of the system utterances are not meaningful in the given context and would probably lead to a broken conversation. Furthermore, about less than two third of the recommendations were considered to be meaningful. Interestingly, none of the two systems "generated" utterances, as almost all system responses were already present in the training data. Overall, our works shows that *(i)* current approaches that are published at high-quality research outlets may have severe limitations regarding their usability in practice and *(ii)* our academic evaluation approaches for CRS should be reconsidered.

CCS CONCEPTS

• **Information systems → Recommender systems**;

KEYWORDS

Conversational Recommender Systems; Evaluation.

ACM Reference Format:
Dietmar Jannach and Ahtsham Manzoor. . End-to-End Learning for Conversational Recommendation:, A Long Way to Go?. In *IntRS Workshop at ACM RecSys '20, September 2020, Online.* ACM, New York, NY, USA, 5 pages.

IntRS@RecSys '20, September 2020, Online

1 INTRODUCTION

In Conversational Recommender Systems (CRS), a software agent interacts with users in a multi-turn dialog with the goal of supporting them in finding items that match their preferences [9]. From an interaction perspective, such systems therefore go beyond the one-shot recommendation paradigm of typical recommenders that can be found, e.g., on e-commerce sites, as they try to elicit the user's specific preferences in an interactive conversation. Such CRS have been in the focus of researchers for more than two decades now, starting with early critiquing approaches in the mid-1990s [7]. Since then, a number of alternative technical approaches were proposed, ranging from more elaborate critiquing techniques [3], over knowledge-based advisory systems [8], to learning-based systems that are based, e.g., on reinforcement learning techniques [5, 15].

In recent years, CRS have gained increased research interest again, mostly due to technological progress in the context of natural language processing (NLP) and the wide-spread use of voice-controlled smart assistants like, e.g., Apple's Siri, Amazon's Alexa or Google Home. Differently from many previous works that are based on substantial amounts of statically defined domain knowledge (e.g., about item properties or pre-defined dialog states and transitions), some recent approaches try to adopt an *end-to-end* learning approach [4, 11–13]. Informally speaking, the main task in such an approach is to learn a machine learning model given a set of recommendation dialogs that were held between humans. One promise of such solutions is that the amount of knowledge engineering can be kept low and that such a system should continuously improve when more dialogs become available.

The evaluation of the usefulness of such CRS in academic environments in general is, however, challenging. To assess the quality of a system, it is not only important to check if the recommendations are adequate in a given dialog situations, one has also to assess the quality of the dialog itself. Dialog quality could, for example, also relate to the question if a system is able to react to chit-chat utterances (phatic expressions) in an appropriate way.

In some recent works, researchers use a combination of objective and subjective measures to assess a CRS. Objective measures can, for example, be typical recommendation accuracy measures, but may also include linguistic measures like *perplexity* that capture the fluency of natural language [4]. Subjective evaluations sometimes include judgments of independent human evaluators. In [4], for example, the task of the evaluators was to rate the consistency of a system-generated utterance at a given dialog state on an absolute scale. In [12], a ranking of different alternatives was requested from the annotators. Such a form of human evaluation however has some limitations. In case of relative comparisons, we do not know if any of the compared systems are useful at all. In case of an absolute scale, the KBRD system from [4], for example, reached a consistency

score of about 2 on a scale from one to three. This, however, cannot inform us fully about how useful such systems are in practice. In a deployed application, users might, for example, not tolerate too many conversation breakdowns, where the system is incapable of responding in a reasonable way.

In this work, we therefore analyze quality aspects of two state-of-the-art end-to-end learning systems ([4, 12]) in a complementary way. Specifically, we ask human evaluators to assess, using a binary scale, if a given response by the system appears meaningful to them or not. Examples of non-meaningful utterances would be a repetition of what was previously said or a system utterance that does not match the context of the dialog. In addition, we asked the evaluators to judge every item recommendation in a subjective way.

Our analysis shows that about one third of the utterances generated by the investigated systems are considered to be *not* meaningful in the given dialog context. Moreover, in both of the systems, about more than one third of the recommendations did not suit the assumed preferences of the recommendation seeker. These observations raise questions both regarding the practical usefulness of the proposed systems and the way we evaluate CRS in academia. A main implication of our analyses is that more realistic ways of evaluating CRS are needed. In particular, such evaluation approaches should help us understand if a *(i)* proposed end-to-end learning system reaches a quality level, in terms of generating plausible responses, that is actually acceptable for users and if it is *(ii)* able to avoid bad recommendations that can be detrimental to the system's quality perception and use [2].

2 ANALYZED APPROACHES

Technical Approaches – DeepCRS and KBRD. We analyzed two recent approaches from the literature. The first one, which we denote as *DeepCRS*, was published at NeurIPS 2018 [12]. Its architecture consists of four sub-components, which accomplish different tasks such as sentence encoding, next-utterance prediction, sentiment classification and recommendation. Technically, the architecture is inspired by the hierarchical HRED architecture from [16] and based on RNNs and an autoencoder for the recommendation task.

The second approach is called *KBRD* [4] (Knowledge-based Recommender Dialog System) and was published at EMNLP-IJCNLP '19. The system's components include a Transformer-based sequence-to-sequence module as a dialog system, a knowledge graph that captures dialog-external knowledge about the domain (movies), and a switching network that connects the dialog and knowledge module.

Underlying Data – The ReDial Dataset. Both approaches, DeepCRS and KBRD, are trained on the ReDial[1] dataset. This dataset was collected by the authors of DeepCRS with the help of crowdworkers and consists of more than 10,000 conversations between a recommendation *seeker* and a *recommender*. The crowdworkers were given specific instructions regarding the conversation: For example, each participant had to take one of two roles, seeker or recommender. The seeker had to specify which movies she likes, and the recommender's task was to make recommendations based on the assumed interests of the seeker. The conversations were collected through a web-based interface, where the crowdworkers typed their utterances in natural language. At least four movie mentions were required per dialog, and each dialog had to have at least ten utterances. The resulting conversations were then further enriched. Movie mentions were tagged with movie names and release years. Furthermore, different labels were assigned to the movies, e.g., whether or not a dialog seeker has seen or liked it.

Original Evaluation. The DeepCRS system was evaluated in different dimensions, including the quality of sentiment classification, recommendation quality, and overall dialog assessment. Accuracy was measured in terms of the Cohen's kappa coefficient [6] given the like/dislike labels in the ReDial dataset, which however has a very skewed distribution with over 90% like statements. The more interesting part was the human evaluation. Here, ten participants of a user study were given ten dialogs from the ReDial dataset that contained 56 system-generated utterances. The task of the participant was to rank each utterance compared to (a) the true utterance in the original dialog and (b) the utterance generated by the HRED model [16], which was used as a baseline. The results of the ranking exercise showed that the human recommendations were most often considered to be the best ones and that the proposed DeepCRS model was better than HRED. An example of different responses in a given situation is shown in Table 1.

Table 1: Example of Conversation with Alternative Recommendation Utterances used for Evaluation [12].

...	
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 .
HRED:	i haven't seen that one yet but i heard it was good. i haven't seen that one. have you seen the last house on the left ?
DeepCRS:	star wars : the force awakens is also a good one return of the jedi all good movies

The KBRD approach was evaluated in three dimensions. Two computational metrics, perplexity and *distinct n-gram* measure the fluency and diversity of the natural language. Recommendation quality was measured in terms of Recall. KBRD proved to be favorable over two baselines (DeepCRS and a Transformer model) in terms of all computational metrics. For the human evaluation, ten annotators with knowledge in linguistics were asked to assess the consistency of a generated utterance with the previous dialog history on a scale from 1 to 3. An average consistency rating of 1.99 was obtained, which was about 15% higher than the average rating for the DeepCRS baseline.

Overall, the (relative) ranking exercise for DeepCRS unfortunately does not tell us much about the absolute meaningfulness and usefulness of the system-generated utterances. For KBRD, the evaluation was done on an absolute scale. The average score was at about 2 on the 1-3 scale, i.e., in the middle. No details are, however,

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

provided regarding the distribution of the ratings. It is, for example, not clear if trivial system responses like *goodbye* to a seeker's *goodbye* were counted as consistent dialog continuations.

Given the difficulty of assessing the usefulness of the proposed systems from the reported studies, our goal was to assess the quality of the system responses through a complementary analysis.

3 ANALYSIS METHODOLOGY

Looking at example conversations published in the supplementary material of [12][2], we found that even in these hand-selected examples many system responses (labeled as OURS) were not meaningful. One main goal of our analysis was to quantify the extent of such problems. Our analysis procedure was as follows.

First, we randomly selected 70 dialogs from the ReDial test dataset. We then used the code provided by the authors of DeepCRS and KBRD to reproduce the systems and to generate system responses after each seeker utterance, given the dialog history up to that point. As a result, we obtained 70 dialogs, which not only contained the original seeker and human recommender utterances, but also the recommender sentences that were generated by the respective CRS.

In total, 758 system responses, 399 by DeepCRS and 359 by KBRD, were generated this way. We analyzed the responses both through manual and automated processes. For replicability, we share all study materials online[3]. The following main analyses were made:

(1) Creativeness or novelty of responses wrt. training data;
(2) Meaningfulness of system responses in the given context;
(3) Quality of the recommendations;

In the context of analysis (1), we were wondering how different the system-generated responses are from the training data. This is in particular relevant as the authors of KBRD measure perplexity and n-gram distance for the generated sentences. In our analysis, we therefore counted which fraction of the system responses was contained in an identical or almost identical form[4] in the training data. In case the generated sentences were mostly identical to sentences appearing in the training data, measuring perplexity and the n-gram distance of what are mostly genuine human utterances is not very informative.

To measure aspects (2) and (3), we relied on human annotators who marked each generated system response in the 70 dialogs as being meaningful or not. Furthermore, we asked them to label each utterance as being chit-chat or containing a recommendation. Two of the three annotators were PhD students at two universities with no background in conversational recommendation. One was evaluating the DeepCRS responses, the other annotated the KBRD responses. They were not informed about the background of their task. To obtain a second opinion and to avoid potential biases, both datasets were also manually labeled by one of the authors of this paper. The annotator agreement was generally very high (92.73% for DeepCRS and 93.89% for KBRD).

When instructing the external annotators, we did not provide specific instructions what "meaningful" means. Our analysis shows that typical non-meaningful sentences include situations where the system ignores the last user-utterance intent, repeats questions, abruptly ends the conversation, provides broken or incomplete response, or makes a bad recommendation. The judgment of what represents a bad recommendation is in many cases clear, e.g., when the system recommended a movie that the seeker has just mentioned, but to some extent it remains a subjective assessment.

[2]https://papers.nips.cc/paper/8180-towards-deep-conversational-recommendations
[3]https://drive.google.com/drive/folders/10gPOmaiFrZjIULIa3LsdmuyvJvnCV_Xq
[4]We considered sentences to be almost identical if the same set of words appeared in them with the same frequency, i.e., only the order was changed.

4 RESULTS

Analysis of Generated Sentences. Table 2 shows the characteristics of the utterances that are generated by DeepCRS and KBRD. Even though both algorithms were fed with the same dialogs to compete, the number of generated sentences for KBRD is lower as this method did not always return a response.

Regarding the novelty of the returned sentences, we, to some surprise, found that DeepCRS almost exclusively returns sentences that are found in identical form in the training data (except for the placeholder for movie names that are eventually replaced by the algorithm). KBRD also mostly returns sentences that are contained in the training data or are tiny modifications of such sentences. Five of the generated sentences were not in the training data. However, there were also 11 generated sentences that were broken.

Overall, it is surprising that both systems mostly return sentences they found in the training data, which resembles more a retrieval approach than a language generation problem. Measuring linguistic properties, e.g., perplexity on the sentence level, of what are genuinely human sentences, therefore is not too meaningful.

Analysis of Dialog and Recommendation Quality. In Table 3, we show the results of the labeling process by the annotators. The numbers in the table correspond to the rounded average of two annotators who, as mentioned above, have a very high agreement.

Table 2: Characteristics of Generated Sentences

	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: Analysis of Dialog and Recommendation Quality

	DeepCRS	KBRD
Number of dialogs	70	70
Generated sentences (overall)	399	359
Sentences labeled as meaningful	277 (69%)	209 (58%)
Sentences labeled as *not* meaningful	122 (31%)	150 (42%)
Dialogs 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. dialogs with no meaningful recs.	25 (36%)	20 (28%)
Nb. dialogs with no rec. made.	7 (10%)	6 (8.5%)

The results show that for both systems a substantial fraction of the generated responses—about 40%—were considered *not* meaningful by the annotators. As a result, there are only 5 (7%) dialogs for which there is not at least one issue. Overall, these findings raise the question if such high failure rates would be acceptable by users in practice?

A major fraction of the generated sentences (33% for DeepCRS, 24% for KBRD) were considered chit-chat. The analyzed systems were performing better in terms of generating such chit-chat messages than they were when generating other types of utterances. The percentage of meaningful chit-chat responses is 85% and 87% for DeepCRS and KBRD respectively. However, a larger fraction of these chit-chat exchanges consist of trivial responses to *'hello', 'hi', 'goodbye'* and *'thank you'* utterances by the recommendation seeker. Overall, the chit-chat messages account for 39% of all generated sentences that were marked as meaningful.

The analysis of the quality of the recommendations themselves is what we thought would be a more subjective part of our evaluation. The agreement between the annotators was, however, very high (93% for DeepCRS and 92% for KBRD). The annotators both relied on their own expertise in the movie domain and used external sources like movie databases such as IMDb to check the plausibility of the recommendations. Recommendations were typically considered not meaningful when the annotators could not establish any plausible link between seeker preferences and recommendations. An example is the system's recommendation of the movie "The Secret Life of Pets" after the seeker mentioned that s/he liked "Avengers - Infinity Wars". Quite interestingly, the subjective performance of KBRD is lower than for DeepCRS even though KBRD includes a knowledge graph, called DBpedia [5], that contains with information about movies and their relationships.

Overall, the results in Table 3 show that the perceived recommendation quality is modest. While for the DeepCRS and KBRD systems, less than two third of the movie recommendations were considered meaningful. However, DeepCRS produced not even a single recommendation in 7 (10%) dialogs and this was the case for 6 (8.5%) dialogs with KBRD.

Limitations of the ReDial Dataset. The existence of large-scale datasets containing human conversations is a key prerequisite for building a CRS based on end-to-end learning. The ReDial dataset is an important step in that direction. However, the dataset also has a number of limitations. This mostly has to do with the way it was created with the help of crowdworkers, which were given specific instructions about the minimum number of interactions and the minimum number of movie mentions.

As a result, many dialogs are not much longer than the minimum length, and the dialogs do not enter deeper discussions. The expression of preferences is very often based on movie mentions and only to a lesser extent based on preferences regarding certain features like genre or directions. The responses by the human recommenders are also mainly movie mentions. An explanation *why* a recommendation is a good match for the seeker's preferences are not very common. Developing an end-to-end system that is capable of providing explanations, which might be a helpful feature in any CRS, therefore might remain challenging.

[5] https://wiki.dbpedia.org/

The DeepCRS and KBRD systems did not provide explanations in the dialogs that we examined, except in cases where the system generated some sort of confirmatory utterances (*"it is a very good movie"*). Such utterances appeared in the training data and were correspondingly sometimes selected by the systems. The number of user intents that are actually supported by DeepCRS and KBRD are generally quite low, see [1, 9] for a list of possible intents in CRS. For the DeepCRS system, for example, the authors also explicitly state that with their recommendation mechanism they are unable to respond to a seeker who asks for "a good sci-fi movie" [12]. Not being able to support intents related to explanations or feature-based requests again raises questions about the practicability of the investigated approaches.

Finally, when examining the dialogs that we sampled, we observed that a few conversations were broken. This was for example the case when a crowdworker had not understood the instructions. In one case, for example, the seeker was not interested in a recommendation, but rather told the recommender that he would like to watch a certain movie. In our sample of 70 dialogs, we found 9 cases we considered to be broken. To what extent such noise in the data impacts the performance of end-to-end learning systems however requires more investigations in the future.

5 CONCLUSIONS AND IMPLICATIONS

We performed an alternative and independent evaluation of two recently published end-to-end learning approaches to building conversational recommender systems. A manual inspection of the responses of the two systems reveals that these systems in many cases fail to react in a meaningful way to user utterances. The quality of the recommendations in these dialogs also appears to be limited.

Our findings have important implications. First, current evaluation practices, at least those from the analyzed papers, seem to be not informative enough to judge the practical usefulness of such systems. Relying on *relative* subjective comparisons (as in [12]) cannot inform as about whether or not the better-ranked system is actually good. Absolute evaluations (as in [4]) indicated mediocre outcomes, but the aggregation of the human ratings into one single metric value prevents us from understanding how good the system works for specifics parts of the conversation (e.g., chit-chat, recommendation).

Measuring linguistic aspects like perplexity—or the BLEU score as done in some other works [10, 13]—might in principle be helpful, even though there are some concerns regarding the correspondence of BLEU scores with human perceptions [14]. However, our analysis revealed that the examined systems almost exclusively generate sentences that were already present in the training data in identical or almost identical form. These objective measures, when applied on the sentence level, would therefore mainly judge, e.g., the perplexity of the sentences by human recommenders. Therefore, a retrieval based approach might achieve the same performance or even better. This is an interesting future direction that we intend to explore.

As a result, our work calls for extended, alternative, and more realistic evaluation practices for CRS. In particular, in practical applications, certain guarantees regarding the quality of the system responses and recommendations might be required, which might be difficult to achieve with current end-to-end learning approaches.

REFERENCES

[1] Wanling Cai and Li Chen. 2020. Predicting User Intents and Satisfaction with Dialogue-Based Conversational Recommendations. In *UMAP '20*. 33–42.

[2] Patrick Y.K. Chau, Shuk Ying Ho, Kevin K.W. Ho, and Yihong Yao. 2013. Examining the effects of malfunctioning personalized services on online users' distrust and behaviors. *Decision Support Systems* 56 (2013), 180–191.

[3] Li Chen and Pearl Pu. 2012. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 125–150.

[4] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards Knowledge-Based Recommender Dialog System. In *EMNLP-IJCNLP '19*. 1803–1813.

[5] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *KDD '16*. 815–824.

[6] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[7] Kristian J. Hammond, Robin Burke, and Kathryn Schmitt. 1994. Case-Based Approach to Knowledge Navigation. In *AAAI '94*.

[8] Dietmar Jannach. 2004. ADVISOR SUITE – A Knowledge-based Sales Advisory System. In *ECAI '04*. 720–724.

[9] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li'e Chen. 2020. A Survey on Conversational Recommender Systems. *ArXiv* abs/2004.00646 (2020).

[10] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue. In *EMNLP-IJCNLP '19*. 1951–1961.

[11] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul A. Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a Communication Game: Self-Supervised Bot-Play for Goal-oriented Dialogue. *ArXiv* abs/1909.03922 (2019).

[12] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *NIPS '18*. 9725–9735.

[13] Lizi Liao, Ryuichi Takanobu, Yunshan Ma, Xun Yang, Minlie Huang, and Tat-Seng Chua. 2019. Deep Conversational Recommender in Travel. *ArXiv* abs/1907.00710 (2019).

[14] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP '16*. 2122–2132.

[15] Tariq Mahmood and Francesco Ricci. 2009. Improving Recommender Systems with Adaptive Conversational Strategies. In *HT '09*. 73–82.

[16] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI '16*. 3776–3783.