Generation-based vs. Retrieval-based Conversational Recommendation: A User-Centric Comparison

AHTSHAM MANZOOR and DIETMAR JANNACH, University of Klagenfurt, Austria

In the past few years we observed a renewed interest in conversational recommender systems (CRS) that interact with users in natural language. Most recent research efforts use neural models trained on recorded recommendation dialogs between humans, supporting an end-to-end learning process. Given the user's utterances in a dialog, these systems aim to *generate* appropriate responses in natural language based on the learned models. An alternative to such language generation approaches is to *retrieve* and possibly adapt suitable sentences from the recorded dialogs. Approaches of this latter type are explored only to a lesser extent in the current literature.

In this work, we revisit the potential value of retrieval-based approaches to conversational recommendation. To that purpose, we compare two recent deep learning models for response generation with a retrieval-based method that determines a set of response candidates using a nearest-neighbor technique and heuristically reranks them. We adopt a user-centric evaluation approach, where study participants (N=60) rated the responses of the three compared systems. We could reproduce the claimed improvement of one of the deep learning methods over the other. However, the retrieval-based system outperformed both language generation based approaches in terms of the perceived quality of the system responses. Overall, our study suggests that retrieval-based approaches should be considered as an alternative or complement to modern language generation-based approaches.

CCS Concepts: • **Information systems → Recommender systems**; • **Human-centered computing → Human computer interaction (HCI)**.

Additional Key Words and Phrases: Conversational Recommendation, Language Generation, Retrieval, End-to-end Learning, Evaluation

ACM Reference Format:
Ahtsham Manzoor and Dietmar Jannach. 2021. Generation-based vs. Retrieval-based Conversational Recommendation: A User-Centric Comparison. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21), September 27-October 1, 2021, Amsterdam, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3460231.3475942

1 INTRODUCTION

The promise of modern-day conversational recommender systems (CRS) is to be able to support a human-like interactive dialog with users in natural language. In recent years, we observed an increased research interest in such systems, which is fueled by the increased spread of voice-controlled devices and advances in natural language processing and machine learning in general; see [8, 12] for recent surveys on the topic.

Many recent works in the area follow an end-to-end learning approach, where machine learning models are trained on a larger set of recorded recommendation dialogs between users. The *ReDial* dataset is a prominent example of such a dataset from the movie domain, which was used for the development of a CRS, for example, in [4] and [18]. Technically, most of these recent CRS are based on a *language generation* approach, probably due to the popularity of deep neural models in general, where the system uses the learned model to generate sentences in natural language in response to

Manuscript submitted to ACM

utterances by the user. Technically, different deep neural network architectures are commonly used for the language generation task, including recurrent neural networks (RNN) and sequence-to-sequence encoder-decoder models. These networks are often combined with additional components and external knowledge bases for the recommendation task, i.e., for determining the items to be recommended.

An alternative to language generation approaches are *retrieval*-based approaches. Here, the main idea is to retrieve and possibly adapt existing utterances from the dataset of recorded dialogs. Retrieval-based approaches are commonly used in question-answering (Q&A) systems, where the system tries to identify the best response to a given question within a Q&A database [28–30, 36]. However, in the context of CRS, retrieval-based approaches are currently receiving less attention, maybe due to the current popularity of (large) language models in general. One potential advantage of retrieval-based approaches over language generation-based ones might lie in the fact that they return utterances that were previously made by humans, which means that they are usually grammatically correct and in themselves semantically meaningful.[1] Also, retrieval-based methods do not require the potentially costly training of complex language models. On the other hand, retrieval-based methods might lack in creativity, i.e., the resulting system may have problems to appropriately react to previously unseen situations; see also [36] for a comparison of generation-based and retrieval-based approaches.

In the current literature, e.g., in [4] and [18], researchers commonly only compare generation based models against each other. It therefore remains unclear how retrieval-based approaches would fare in such a comparison. With this work, our goal is to shed light on this question and to also reproduce previous claimed improvements. To answer this question, we conducted an online user study (N=60), where participants rated the quality of the responses by three different CRS for a number of given dialog situations. Specifically, we included two neural language generation approaches—one that we call DeepCRS [18] and the other called KBRD [4]—and a comparably simple retrieval-based system in this comparison. We have chosen DeepCRS and KBRD in our comparison for two reasons: *(i)* both DeepCRS and KBRD were developed and evaluated in the context of the ReDial dataset, and *(ii)* they are two of the most recent neural approaches to CRS that provide the code and material to reproduce the results of their original experiments. We developed an own retrieval-based system for the purpose of this study. It consists of a nearest-neighbor approach to find response candidates and a small set of heuristics to rerank the retrieved candidates. Like all compared approaches, our system also includes a specific modules for the generation of the recommendations.

Through the study we could reproduce the findings from [4] that responses by the KBRD system are on average perceived to be of better quality than those returned by DeepCRS. However, both complex neural methods were outperformed by the basic retrieval-based system to a statistically significant extent. On an absolute scale, the average scores regarding response quality for all compared systems lie between 3 and 4 on a five-point scale (with 5 as a maximum). This indicates that there is still substantial room for improvement. More detailed analyses furthermore show that all systems fail in the majority of cases to respond to questions about item-metadata. For example, one failure situation was observed when the seeker said "*I havent seen 'You Don't Mess with the Zohan (2008)' what it about? (sic)*", and one of the systems responded with "*it is about a group of friends who find their lives in the world of the. (sic)*" In the given response, there are actually two problems. First, the response is broken, ending with an incomplete sentence. Second, the description of the movie is wrong as it is not primarily about a group of friends. Moreover, as observed in previous research in [11], we could reproduce the phenomenon in our study that the generation-based systems barely generate new sentences, i.e., almost all returned responses appeared in the same or almost identical form in the

[1] In our experiments we found that the examined language generation approaches sometimes create ungrammatical or incomplete sentences.

training data. Overall, we see huge potential for approaches that combine generation and retrieval approaches and which furthermore rely on additional components to be able to answer questions about item-metadata, either based on explicitly encoded knowledge bases or through open-domain language models such as BERT or GPT-3.

The paper is organized as follows. In Section 2, we provide details on the compared neural systems and briefly review evaluation approaches for CRS in general. Next, in Section 3, we give an overview of the developed retrieval-based system. The evaluation setup is described in Section 4 and results are given in Section 5. The paper ends with a discussion of implications and an outlook on future directions.

2 PREVIOUS WORK

Language Generation Approaches: DeepCRS and KBRD. Several CRS approaches based on language generation models were proposed in recent years, and various neural network components were used in these systems, including RNNs, CNNs, GANs, encoders-decoders pairs, etc. [4, 10, 14, 15, 18, 36].

In this work, we focus on the DeepCRS and KBRD systems as recent examples of works, which were both evaluated on the same dataset (ReDial). Basically, this dataset was created in the context of the development of DeepCRS with the help of crowdworkers who were tasked to participate in a conversation, where one person had the role of recommendation *seeker* and one was the (human) *recommender*. Overall, the dataset, which uses movie recommendations as an application domain, consist of over 10,000 dialogs.

The *DeepCRS* system, published at NeurIPS '18, uses an architecture that involves a hierarchical recurrent encoder derived from HRED [31] and a switching decoder based on [32]. For each observed seeker utterance, the system checks if a movie was mentioned and, if this is the case, uses an RNN to infer the sentiment (preference) towards the movie. The user's preferences are then handed over to an auto-encoder based recommendation module which is pre-trained on MovieLens data. The recommendations are finally used by the output decoder to generate the response to the user.

KBRD, presented at EMNLP-IJCNLP '19, is also a neural end-to-end learning system. Technically, it is based on a sequence-to-sequence encoder-decoder Transformer framework [33]. The Transformer was preferred over HRED in this architecture due to its often better performance for different NLP tasks [5, 20, 22, 27, 33, 37]. The recommendation module of KBRD moreover relies on external information about movies, and it involves a knowledge graph based on DBpedia data [17]. Thereby, the system is able to consider movies and their features that are observed in the dialog history in the recommendation process.

Retrieval-based Approaches. Various retrieval-based approaches have been proposed for a number of NLP tasks, e.g., machine translation or question-answering (Q&A), see, e.g., [1, 7, 28–30]. In particular, Q&A style systems gained importance in recent years due to the widespread use of digital assistants like Apple's Siri, but they often do not support multi-turn dialogs required by a CRS. Nonetheless, several of these systems are interesting from a technical perspective.

The *AliMe* chatbot system [26], for example, is a Q&A system that combines retrieval and generation elements, where a sequence-to-sequence model is used to rerank a previously retrieved set of candidate responses. Similarly, a retrieval-based Q&A system based on an inference network model is proposed in [2], which implicitly supports ranking of the candidate responses in a Q&A database using the query structure and annotations in addition to keywords. Apart from single-turn Q&A systems, a few hybrid approaches were also applied for multi-turn dialog systems. For example, [36] presents a hybrid approach where the retrieval component returns a set of candidates from an existing dataset that contains queries and responses. Another practical example of such a hybrid approach is Microsoft's *XiaoIce*, a popular social chatbot system [38], which determines response candidates with two query-response databases.

Overall, while such hybrid approaches were found to be useful for Q&A scenarios and general dialog systems, we are not aware of works that aim to combine the outputs of language generation and retrieval modules in a similar way for the problem of conversational recommendation. Moreover, we did not identify any natural language based CRS that is mainly based on the retrieval of past responses like the system that we propose in this work and that we use in our study.

Evaluation of Conversational Recommenders. Evaluating a CRS is often considered to be challenging, as the evaluation process typically requires more than an assessment if the made recommendations are relevant for the user, using, for example, metrics like Precision and Recall. Some specific quality dimensions like dialog *efficiency*, i.e., how quickly the system finds a recommendation that matches the user's interest, are sometimes assessed with simulation experiments. Some researchers also resort to linguistic measures to asses the quality of the generated system responses, which however can have major limitations [19]. Ultimately, studies involving humans in the loop in almost all cases are required to assess the quality of a CRS in a more holistic manner. Results of various user studies are reported in the literature, e.g., in [3, 9, 13, 16, 21, 23, 25, 34], where the studies focus on various quality aspects, e.g., perceived effort or overall usefulness of the system.

The DeepCRS [18] and KBRD [4] systems investigated in this paper were evaluated in various dimensions, using both offline experiments, e.g., regarding recommendation accuracy or the accuracy of sentiment classification, and studies that involve human evaluators. The human evaluation of the DeepCRS system was based on collecting feedback from several annotators regarding the *relative* quality of the utterances generated by the DeepCRS and the HRED baseline system. The authors found that DeepCRS performed better than HRED based on the feedback by the annotators. Regarding KBRD, the evaluation of the system was also based on different evaluation approaches. In the human evaluation part of the original studies, the authors of KBRD asked human judges to assign scores on a scale from 1-3 to the responses generated by the DeepCRS and the KBRD system. The quality criterion for this assessment was *consistency* with the utterances in the dialog so far. Their study showed that the KBRD on average received better scores than the DeepCRS system.

In this work, we not only aim to reproduce these findings regarding the superiority of KBRD over DeepCRS in terms of the user-perceived quality perception of responses including recommendations, but also want to understand how a basic retrieval-based system would fare in such a comparison. As done in the evaluation of the KBRD system [4], we ask study participants to provide absolute scores for the responses by the different systems. In our case, we however involve a larger set of independent human evaluators in an online study. Note that in the evaluation of DeepCRS and KBRD only 10 judges were involved, and little is known about how they were recruited or which specific instructions they received.

3 A BASIC RETRIEVAL-BASED CONVERSATIONAL RECOMMENDER SYSTEM (RB-CRS)

The RB-CRS system used in our comparison has two main modules. The *retrieval* module is responsible for retrieving candidate responses from the "training" dataset, given the current dialog situation, which includes the last utterance of the recommendation *seeker*. Furthermore, it selects one of the candidates as a response using heuristics. In case the selected response includes a recommendation—i.e., it is not, for example, a greeting response—the *recommendation module* is then responsible for determining a suitable item to recommend in the given situation. The item name (i.e., a

movie title) is then integrated into the selected response in which we replaced specific movie mentions with placeholders in a preprocessing step. To ensure reproducibility, we share all code and material used in our study online[2].

Retrieval and Ranking Approach. In the first step of the candidate retrieval process we take the last seeker utterance in the current dialog (for which we seek a response) and determine the n most similar seeker utterances in the ReDial dataset. In our experiment, we used $n = 5$ as it led to a quite diverse set of responses and good results in a pretest.[3] To compute the similarity scores, we use a TF-IDF encoding of the utterances to which we previously applied common preprocessing steps like conversion to lowercase, replacing movie titles with placeholders, removing stop words and special characters etc.[4]

Given the n most similar seeker utterances, we consider the immediately following recommender responses in the dataset as response candidates. In order to avoid that too short and too long sentences are considered, we ensured that all returned candidates have a length between 2 and 12 words. While too short responses might not contain enough relevant information, long responses often address more than one question or intent by a seeker. Note that the mean length of recommender responses in the ReDial dataset is 9.68 (SD=7.75), which informed the choice of our boundaries.

To select a response among the $n = 5$ candidates, we first check if there are at least two[5] responses that mention a movie title. In this case, we assume that answering with a movie recommendation is the best option and we return the shortest of these candidates. Our intuition for this choice is that for a recommendation response, more concise answers are preferable. If, on the other hand, almost all (at least four) response candidates do *not* mention a movie, we return the response candidate without a movie mention which had the highest similarity value in the previous retrieval step.

Recommendation Approach. Our system supports two possible user *intents*, once it has determined that the response should be a recommendation: *(i)* the user seeks recommendations based on their stated movie preferences, or *(ii)* the user looks for recommendations of a certain genre, e.g., 'horror'. We distinguish between the two intents by scanning the dialog so far for movie and genre mentions. Roughly speaking, if the seeker mentioned some movies in the dialog, we assume a positive preference towards these movies and assume the first intent. Otherwise, if one of about 30 predefined genre keywords—these can be derived from datasets such as MovieLens—is found, we assume that genre-based movie recommendations are preferable.

Technically, the method that relies on stated movie preferences—used for intent *(i)* above—retrieves a set of movies that are most similar in the latent factor space to the last mentioned movie in the dialog.[6] To that purpose, we first apply matrix factorization on the underlying MovieLens-25M dataset and then use cosine similarity to measure the distance between item embeddings. Technically, we used the *TruncatedSVD* method from Python's *sklearn* library, with the number of latent factors set to 20. From the most similar movies in the latent space, we then retain those that also have a genre overlap with the last mentioned movie. The purely genre-based method used for intent *(ii)*, on the other hand, matches the last identified genre in the dialog so far with the genres of the movies in the catalog. Clearly, for both methods more elaborate schemes are possible. In both cases, we finally apply popularity filters to avoid too obscure

[2]https://github.com/ahtsham58/RB-CRS
[3]We only considered the very last seeker utterance for retrieval; in our future works, we plan to investigate if considering more context in the retrieval phase is beneficial as proposed in [35].
[4]We also experimented with Sentence-BERT [28] as a semantic encoding, but this led to worse results.
[5]This and other thresholds were determined through pretests, in which we used another sample of randomly selected dialogs as test set to fine-tune our approach.
[6]There are a few situations where the mentioned movie is not part of the MovieLens dataset. In this case, we use a genre-based approach described below, based on movie metadata in the ReDial dataset.

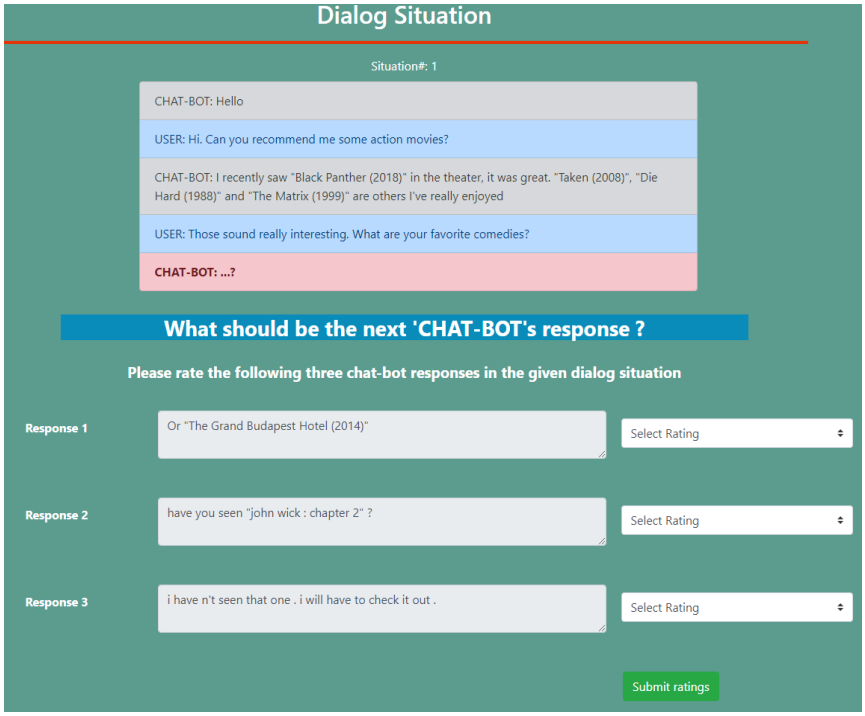

Fig. 1. Response Rating User Interface

recommendations, which can occur in purely similarity-based approaches, see also [6]. Note that in case a response contains both a movie mention and a genre keyword, we apply the movie-based recommendation approach.

	DeepCRS	KBRD	RB-CRS
Avg. score	3.13	3.46	**3.71**
Std. deviation	1.49	1.45	**1.32**

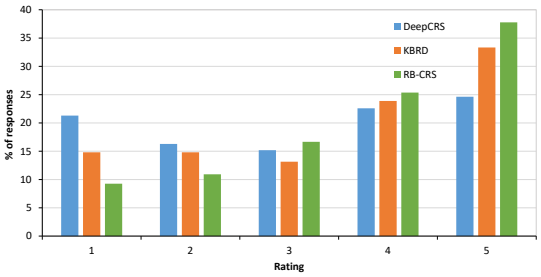

Fig. 2. Average Ratings and Rating Distribution

In the final step before we return the system response, we parameterize the chosen response candidate to contain the title of the recommended movie and we also replace relevant metadata in the response. The metadata replacement is based on a small set of heuristics. If, for example, the chosen response sentence is "Maybe you should check out [movie-ID]. It's extremely scary", we replace any genre mentions (here: "scary") with the correct value for the recommended

movie. Similar replacements are also done for mentions of actors, using the underlying MovieLens dataset. In that respect, our system is similar to KBRD, which also relies on external meta-data in the response generation process.

4 EVALUATION METHODOLOGY

We created a web application for the purpose of our study, where study participants were tasked, after informed consent, to assess to what extent the given responses by three different chatbots implemented through DeepCRS[7], KBRD[8], and RB-CRS, are *meaningful*. For DeepCRS and KBRD we reused the original code provided by the authors to generate the responses given the seeker utterances. We trained the models by ourselves using the same hyperparameters and dataset splits for training, validation, and test as in the original works; assuming the hyperparameters were optimized by the authors for this dataset. For our RB-CRS system, we applied the same splitting approach.

This evaluation approach differs from the original ones for different reasons. First, in our work, we are interested in *absolute* assessment and not *relative* ones as done in [18]. Second, differently from [4], we use a 5-point scale instead of a 3-point scale to obtain more fine-grained feedback. Third, we asked participants to assess the meaningfulness of the responses instead of the linguistic consistency as in [4], because *(i)* linguistic consistency may be too narrow as a criterion as it does not cover recommendation quality and *(ii)* assessing consistency requires a certain level linguistic expertise [4], which we do not generally assume in our study.

Participants were provided with instructions on how to assess the meaningfulness of a response, e.g., in terms of a logical dialog continuation or a recommendation, using a scale labeled from "Entirely meaningless" to "Perfectly meaningful". Specifically, in case a response includes a recommendation, participants were told to assess the quality of the complete response, including the suitability of the provided recommendation. The user interface regarding a dialog situation and responses from three systems is shown in Figure 1.

We randomly sampled 70 dialogs from the ReDial dataset for the experiment. For each participant, we randomly selected 10 *dialog situations* that start at the beginning of the dialog and end with a seeker utterance which was randomly determined. The responses from the three different systems, i.e., DeepCRS, KBRD, and RB-CRS, therefore correspond to possible continuations of this situations that the study participants had to assess. The order of the presented responses was also randomized. To be able to check that participants did the task with care, we implemented an *attention check* in one of the 10 dialog situations.

For an independent evaluation, we recruited 87 participants both through Amazon Mechanical Turk and through personal contacts. After removing 27 non-attentive participants, we ended up with N=60 participants and 540 dialog situations with 3 response assessments each. On average participants needed 11.2 minutes for the task. The collected demographic data reveal that the majority of the participants were fluent in English and frequent movie watchers. This indicates that the participants involved in this study on average had the right background with suitable skills to accomplish the task. Detailed statistics about the participant demographics can be found in the online material.

5 RESULTS

Figure 2 shows the average scores to the responses by the three systems and the distribution of the ratings. Regarding the comparison of DeepCRS and KBRD, our findings are in line with those made with the help of human evaluators in [4], i.e., that KBRD is the favorable system on average. More importantly, however, we found that a retrieval-based system based on nearest neighbors on average leads to even better results than recent and complex neural language generation

[7] https://github.com/RaymondLi0/conversational-recommendations
[8] https://github.com/THUDM/KBRD

approaches. An ANOVA analysis revealed that the differences between the means of the three groups are statistically significant ($p < 0.01$). A Tukey HSD test furthermore shows that the differences in all pairwise comparisons between the three systems are significant as well, with $p < 0.01$ in all cases. Overall, the results suggest that retrieval-based methods can be a promising alternative or complement to recent language generation approaches for CRS, in particular as the examined generation-based CRS do not actually generate many novel sentences, as observed in [11].

Given these observations, we made some additional (preliminary) investigations regarding the question in which situations the systems fail, an analysis which is mostly missing in previous works. A manual inspection of the responses show that all systems have difficulties in answering seeker questions regarding movie metadata, e.g., if a movie features a certain actor or falls into a certain genre, or when asked for an explanation. We identified 25 *unique* situations in our study where the dialog ended with such a specific seeker question. Based on the ratings for such responses, we found that DeepCRS worked best here, but also received very low ratings in more than two thirds of the cases. Our RB-CRS system also exhibited relatively poor performance in such situations. However, we observed an interesting pattern here. In the described situations, RB-CRS often received a score of 3 in case it reacted with a recommendation response. KBRD, in contrast, mostly returned non-suitable non-recommendation responses in the same situations. This might indicate that users may be somehow more tolerant when they are presented with (unexpected) recommendations instead of another type of non-perfect responses. More research is however required to better understand such phenomena.

6 CONCLUSION

In our work, we could reproduce that the more recent KBRD system for conversational recommendation in natural language can be favorable over the DeepCRS system in terms of user perceptions. Moreover, our study highlights the potential value of relying on retrieval-based components when building a CRS. Finally, our work provides a study design that may serve as a blueprint for future human-centric evaluations of such systems. Generally, we see our work as a first step in that direction, and many improvements are possible in terms of the algorithmic solution. Furthermore, our study emphasizes that datasets like ReDial can have their limitations that are due to the way how they were created. Dialogs happen mostly on the level of movie instances and conversations about meta-data barely happen. This generally limits what both generation-based and retrieval-based methods can achieve and calls for the integration of additional information sources and/or general-purpose language models such as BERT in such systems, see also [24].

REFERENCES

[1] Lisa Ballesteros and W Bruce Croft. 1997. Phrasal translation and query expansion techniques for cross-language information retrieval. In *ACM SIGIR Forum*, Vol. 31. 84–91.

[2] Matthew W Bilotti, Paul Ogilvie, Jamie Callan, and Eric Nyberg. 2007. Structured retrieval for question answering. In *SIGIR '07*. 351–358.

[3] Li Chen and Pearl Pu. 2006. Evaluating critiquing-based recommender agents. In *AAAI '06*. 157–162.

[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] Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards knowledge-based personalized product description generation in e-commerce. In *KDD '19*. 3040–3050.

[6] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User Perception of Differences in Recommender Algorithms. In *RecSys '14*. 161–168.

[7] Bilel Elayeb, Wiem Ben Romdhane, and Narjes Bellamine Ben Saoud. 2018. Towards a new possibilistic query translation tool for cross-language information retrieval. *Multimedia Tools and Applications* 77, 2 (2018), 2423–2465.

[8] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and Challenges in Conversational Recommender Systems: A Survey. arXiv:2101.09459

[9] Peter Grasch, Alexander Felfernig, and Florian Reinfrank. 2013. Recomment: Towards critiquing-based recommendation with speech interaction. In *RecSys '13*. 157–164.

[10] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward Sociable Recommendation Dialog Systems. In *EMNLP '20*.

[11] Dietmar Jannach and Ahtsham Manzoor. 2020. End-to-End Learning for Conversational Recommendation: A Long Way to Go?. In *IntRS Workshop at ACM RecSys 2020*. Online.

[12] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *Comput. Surveys* 54 (2021), 1–26. Issue 5.

[13] Yucheng Jin, Wanling Cai, Li Chen, Nyi Nyi Htun, and Katrien Verbert. 2019. MusicBot: Evaluating critiquing-based music recommenders with conversational interaction. In *CIKM '19*. 951–960.

[14] Chaitanya K. Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in Goal-Oriented Dialog. In *NeurIPS '17 Workshop on Conversational AI*.

[15] 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.

[16] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4 (2012), 441–504.

[17] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2 (2015), 167–195.

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

[19] 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.

[20] Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. In *ICLR '18*.

[21] Tariq Mahmood and Francesco Ricci. 2009. Improving recommender systems with adaptive conversational strategies. In *RecSys '09*. 73–82.

[22] Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling Neural Machine Translation. In *WMT '18*. 1–9.

[23] Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A model of social explanations for a conversational movie recommendation system. In *HAI '19*. 135–143.

[24] Gustavo Penha and Claudia Hauff. 2020. What Does BERT Know about Books, Movies and Music? Probing BERT for Conversational Recommendation. In *RecSys '20*. 388–397.

[25] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *RecSys '11*. 157–164.

[26] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. 2017. Alime chat: A sequence to sequence and rerank based chatbot engine. In *ACL'17*. 498–503.

[27] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP '16*. 2383–2392.

[28] Nils Reimers, Iryna Gurevych, Nils Reimers, Iryna Gurevych, Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP '19*.

[29] Stefan Riezler, Alexander Vasserman, Ioannis Tsochantaridis, Vibhu O Mittal, and Yi Liu. 2007. Statistical machine translation for query expansion in answer retrieval. In *ACL '07*. 464–471.

[30] Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. 2019. FAQ retrieval using query-question similarity and BERT-based query-answer relevance. In *SIGIR '19*. 1113–1116.

[31] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *CIKM '15*. 553–562.

[32] Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. In *ICLR '18*.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS '17*. 5998–6008.

[34] Pontus Wärnestål. 2005. User Evaluation of a Conversational Recommender System. In *IJCAI '05 Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

[35] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *SIGIR '16*. 55–64.

[36] Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. In *CIKM '19*. 1341–1350.

[37] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *EMNLP '18*.

[38] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.