Revisiting Retrieval-based Approaches for Conversational Recommender Systems

Ahtsham Manzoor*, Dietmar Jannach

University of Klagenfurt, Universitätsstraße 65-67, Klagenfurt am Wörthersee, 9020, Austria

Abstract

Conversational recommender systems (CRS) interact with users in natural language to support them in their decision-making process. Recently, an increased interest in developing novel approaches to CRS can be observed. Mainly, current research efforts rely on neural models trained on recorded recommendation dialogs between humans, thereby implementing an end-to-end learning process. Given a user utterance in an ongoing dialog, the trained models *generate* suitable responses. An alternative to generation-based approaches is to *retrieve* responses from the recorded dialog data and adapt them to the given dialog context. Such retrieval approaches have proven to be effective in various NLP tasks, but have received limited attention for CRS so far. In our ongoing research, we re-assess the potential value of retrieval-based approaches and compare their performance with recent generation-based approaches. Our results point to various limitations of current neural models and indicate that retrieval-based approaches can be an effective complement to today's generation-based techniques.

Keywords

Conversational recommendation, retrieval and ranking, language generation, evaluation

1. Introduction

Conversational recommender systems (CRS) are software agents that converse with humans in natural language. The main goal of such agents is to recommend items of interest to the users and help them in making their decisions. In recent years, we observe that CRS obtained increased attention, mainly due to the spread of voice assistants like Alexa and Siri, and due to advancements in the area of natural language processing (NLP) and machine learning (ML) in general; see [1] for a recent survey on conversational recommenders.

Current approaches to building CRS [2, 3, 4, 5, 6] mainly adapt an end-to-end learning paradigm. One promise of such end-to-end learning approaches is to avoid the knowledge engineering bottlenecks of traditional critiquing-based or constraint-based CRS [7, 8, 9]. Specifically, these systems rely on ML models that are trained on large corpora of recommendation dialogs recorded with the help of paired humans, e.g., the *ReDial* dataset [2]. Various algorithmic approaches were proposed in recent years in which the system relies on models that were trained on such datasets to *generate* a response given a user utterance in an ongoing dialog.

While today's systems are often capable of returning meaningful responses, they may still fail in many cases. In the online material[1] provided by the authors of the first neural model built on the ReDial dataset [2], for instance, the system once responds to a user as follows: *"what kind of movies do you like ? what kind of movies do you like ?"*. A further inspection of more examples revealed that there are various situations where the proposed system had difficulties to generate suitable responses. Therefore, the question arises to what extent current systems may actually be usable in practice. Moreover, since the performance evaluation of recent neural models is mostly based on offline experiments and computational metrics it remains unclear how users would *perceive* the quality of the responses of such systems.

While generation-based approaches can have different advantages, e.g., that they should be capable to respond to previously unseen dialog situations, we find that today's systems often encounter difficulties to correctly interpret the user's current intent and that they sometimes also return responses that are too short or too general [10]. An alternative to recent NLP-enabled generation-based systems are *retrieval-based* approaches. In such approaches, the idea is to retrieve appropriate responses from the underlying dataset and adapt them to the context of the dialog if needed. Retrieval-based approaches have been applied in various NLP tasks such as question-answering (Q&A) systems, machine translation, and open-domain dialog systems [11, 12, 13, 14, 15]. One main advantage of retrieval-based systems comes from the fact that such approaches do not require long and expensive training of models. Moreover, the returned utterances are mostly complete and semantically correct as they were originally made by humans, see also [16] for a comparison of generation and retrieval-based methods.

In this paper, we summarize our recent findings on the state of the art in conversational recommendation published in [17, 18, 19, 20]. Our main observations reported in these works include *(a)* that today's generation-based models may have difficulties to respond in a meaningful way in a substantial number of cases (30-40 %); *(b)* that some of these systems almost exclusively "generate" sentences that appear in the training data in the exact same form; and *(c)* that *retrieval-based* approaches may represent a promising alternative or complement to generation-based approaches. Differently from most current works, our analyses and experiments are mainly based on user-centric evaluation approaches, given the known limitations of offline experiments and certain linguistic metrics when evaluating dialog systems [21, 22].

2. Assessing Generation-based and Retrieval-based CRS

Analysis of generation-based CRS. We started our journey with the analysis of the system proposed by Li et al. [2], which we refer to as DeepCRS from here on. DeepCRS which was built on the ReDial dataset was also released in the same paper. This dataset contains over 10,000 recommendation dialogs that were obtained with the help of crowdworkers. Later on, Chen et al. [3] proposed another system based on the ReDial dataset, named KBRD, and they reported improved performance over DeepCRS in different dimensions.

As a first step in our research, we used the code of these two systems (DeepCRS and KBRD) to generate system responses for 70 dialog situations. We then relied on human judges who assessed the returned system responses in terms of (1) the quality of the generated responses

---

[1]https://proceedings.neurips.cc/paper/2018/hash/800de15c79c8d840f4e78d3af937d4d4-Abstract.html

on a binary scale, (2) the quality of the *specific recommendations* made in the dialogs. Moreover, we automatically measured the originality of the returned responses, i.e., if they can also be found in the training data in identical or similar form.

The results of our analysis exhibited that 31 % of the responses by DEEPCRS and 42 % of the responses by KBRD were not considered meaningful by the human judges, see [17, 18] for details of these analyses. Also, about 40 % (DEEPCRS) and 45 % (KBRD) of the recommendations of the two systems were not found to be suitable. Typical problems include responses that do not match the dialog situation, broken sentences, or duplicated responses or recommendations in the same dialog. Moreover, it turned out that in almost all cases both neural approaches actually did not generate *new* sentences, but returned sentences that appeared in the exact same or very similar form in the training data. Interestingly, the manual analysis with human judges indicated that DEEPCRS was actually the better system when using our specific evaluation procedure, which stands in contrast to the findings reported in [3].

What about rule-based approaches?   Many of today's online chatbots, e.g., ones that use Google's DialogFlow[2], are based on pre-defined text templates and an inference mechanism that determines the user's current intent and then selects a suitable template as a response. Our next research question was how far we can get with a simple, manually engineered rule-based systems and how such a system would fare when compared to DEEPCRS and KBRD. To that purpose, we analyzed the ReDial dataset to identify the most frequent user intents. We then developed a simple rule-based pattern-matching CRS—in the spirit of the famous ELIZA [23] system—which guesses the intent of a given user utterance and then selects one of several suitable pre-curated sentences as a response. We compared our rule-based system with DEEPCRS and KBRD through a user study, where study participants (N=58) had to assess the quality of the responses, using a 5-point scale, by the different systems for a given dialog situation, see [18] for details. Specifically, each participant was tasked to asses 10 dialog situations, one of them was considered as an attention check, thereby in total 522 situations were assessed. The study showed that our simple rule-based system, which mainly consists of a few dozen if-statements, on average led to better quality perceptions than the recent neural models. However, there were still many situations in which all three compared systems failed.

Towards retrieval-based conversational recommendation.   Since we found that the neural models mainly returned existing utterances from the underlying dataset, and we do not consider engineered rule-based systems to be the future of CRS, we designed a *retrieval*-based approach. The first version of this approach and its evaluation are discussed in [19]. Given the last user utterance as an input, the system first determines similar user utterances in the dataset. The responses to these similar user utterances in the dataset are considered as candidates to return to the user, and we designed a small set of heuristics to select one of these candidates. In case the chosen response actually includes a recommendation, i.e., it is not some other form of utterance like a greeting, a specific recommendation module is used to determine a suitable item recommendation, which is then integrated into the selected response text. Again, we evaluated our system through a user study, with (N=60) subjects, like the one used to evaluate

---

[2]https://dialogflow.cloud.google.com/

our rule-based system. The study, which again included DEEPCRS and KBRD as baselines, showed that the retrieval-based system on average led to the best quality perception among the compared systems. As a side result, we also found that KBRD performed better than DEEPCRS in the user study, which supports the findings reported in [3].

Inspired by these promising findings, we further improved our retrieval-based systems, see [20]. Specifically, we found that our system has difficulties to respond to very short user utterances. Therefore, we adapted the approach in a way that it considers not only the last user utterance as a context but several of the most recent utterances in the dialog. Technically, we retrieve several *sets of candidates*, and we then use an outlier pruning (or: clustering) technique to identify the most plausible candidates, which in our case are candidates that are similar to each other. In a final step, the remaining candidates undergo a final selection process, where *perplexity* is used as a linguistic metric to rank the candidates. Before the response is returned to the user, again suitable item recommendations are injected using the recommender module.

The evaluation was done through a user study with (N=90) subjects. This time we included KBRD and the more recent KGSF system [4] as baselines in the experiment. While KGSF outperformed KBRD, our improved retrieval-based system led to even better results than KGSF. In [20], we also report the outcomes of a number of additional analyses. We for example compare the performance of the systems at different dialog stages, we discuss difficult situations (intents) and typical failure points, and we specifically analyze in which ways users express their preferences in the ReDial dataset. Furthermore, we explore the relationships between the perceived meaningfulness of a response and certain (linguistic) characteristics of the response.

3. Discussion

Substantial progress was made recently for CRS that support natural-language interactions. Still, our work shows that building effective CRS remains a "grand AI challenge" [24]. Our research indicates that retrieval-based approaches can be an alternative or complement to today's dominating approaches based on language generation. Given that both types of approaches have their advantages, future works may therefore more often consider hybrid systems that combine sentence retrieval and sentence generation in one system. Positive experiences with hybrid solutions were observed for related problem settings, e.g., for the *AliMe chat*, a single-turn Q&A system [25] or *Microsoft's XiaoIce*, a popular social chatbot system [26].

Our research also highlights that today's efforts to build CRS is hampered by the datasets that are available for learning. Dialogs in the ReDial dataset, for example, in many cases mention individual movies, e.g., when users indicate their preferences, but not preferences in terms of genres. In addition, there are almost no situations where the recommendation seeker asks for an explanation, which makes it almost impossible for the CRS to learn how to explain. Better datasets are therefore required, also ones that allow for more social interactions, e.g., [27].

Finally, our works shed light on potential issues of today's predominant evaluation practices. Researchers often seem to overly rely on computational metrics. While some reported human-evaluations, which are typically not described in detail. With our research works we lay out one possible way of *how to evaluate CRS* with humans in the loop. More work however remains to be done to ensure consistent progress in this challenging research area.

References

[1] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Computing Surveys 54 (2021) 1–36.

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

[3] 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.

[4] K. Zhou, W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, J. Yu, Improving conversational recommender systems via knowledge graph based semantic fusion, in: KDD '20, 2020, pp. 1006–1014.

[5] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, J.-R. Wen, Towards topic-guided conversational recommender system, in: ICCL '20, 2020, pp. 4128–4139.

[6] J. Zhou, B. Wang, R. He, Y. Hou, CRFR: Improving conversational recommender systems via flexible fragments reasoning on knowledge graphs, in: EMNLP '21, 2021, pp. 4324–4334.

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

[8] A. Felfernig, G. Friedrich, D. Jannach, M. Zanker, Constraint-based recommender systems, in: Recommender Systems Handbook, Springer, 2015, pp. 161–190.

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

[10] Y. Song, C.-T. Li, J.-Y. Nie, M. Zhang, D. Zhao, R. Yan, An ensemble of retrieval-based and generation-based human-computer conversation systems, in: IJCAI '18, 2018, pp. 4382–4388.

[11] S. Riezler, A. Vasserman, I. Tsochantaridis, V. O. Mittal, Y. Liu, Statistical machine translation for query expansion in answer retrieval, in: ACL '07, 2007, pp. 464–471.

[12] W. Sakata, T. Shibata, R. Tanaka, S. Kurohashi, FAQ retrieval using query-question similarity and BERT-based query-answer relevance, in: SIGIR '19, 2019, pp. 1113–1116.

[13] G. Bonetta, R. Cancelliere, D. Liu, P. Vozila, Retrieval-augmented Transformer-XL for close-domain dialog generation, in: FLAIRS '21, 2021.

[14] A. Bartl, G. Spanakis, A retrieval-based dialogue system utilizing utterance and context embeddings, in: ICMLA '17, 2017, pp. 1120–1125.

[15] H. Sugiyama, T. Meguro, R. Higashinaka, Y. Minami, Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures, in: SIGDIAL '13, 2013, pp. 334–338.

[16] L. Yang, J. Hu, M. Qiu, C. Qu, J. Gao, W. B. Croft, X. Liu, Y. Shen, J. Liu, A hybrid retrieval-generation neural conversation model, in: CIKM '19, 2019, pp. 1341–1350.

[17] D. Jannach, A. Manzoor, End-to-end learning for conversational recommendation: A long way to go?, in: IntRS Workshop at RecSys '20, Online, 2020.

[18] A. Manzoor, D. Jannach, Conversational recommendation based on end-to-end learning: How far are we?, Computers in Human Behavior Reports (2021) 100139.

[19] A. Manzoor, D. Jannach, Generation-based vs. retrieval-based conversational recommendation: A user-centric comparison, in: RecSys '21, 2021.

[20] A. Manzoor, D. Jannach, Towards retrieval-based conversational recommendation, Infor-

mation Systems (2022) 102083.

[21] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, J. Pineau, How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation, in: EMNLP '16, 2016, pp. 2122–2132.

[22] D. Jannach, C. Bauer, Escaping the mcnamara fallacy: Towards more impactful recommender systems research, AI Magazine 41 (2020) 79–95.

[23] J. Weizenbaum, ELIZA – Computer program for the study of natural language communication between man and machine, Communications. ACM 9 (1966) 36–45.

[24] D. Jannach, L. Chen, Conversational Recommendation: A Grand AI Challenge, AI Magazine 43 (2022).

[25] M. Qiu, F.-L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, W. Chu, AliMe chat: A sequence to sequence and rerank based chatbot engine, in: ACL'17, 2017, pp. 498–503.

[26] L. Zhou, J. Gao, D. Li, H.-Y. Shum, The design and implementation of XiaoIce, an empathetic social chatbot, Computational Linguistics 46 (2020) 53–93.

[27] S. A. Hayati, D. Kang, Q. Zhu, W. Shi, Z. Yu, INSPIRED: Toward sociable recommendation dialog systems, in: EMNLP '20, 2020.