Conversational Recommendation as Retrieval: A Simple, Strong Baseline

Raghav Gupta, Renat Aksitov, Samrat Phatale, Simral Chaudhary, Harrison Lee, Abhinav Rastogi
Google Research
{raghavgupta, raksitov, samratph, simral, harrisonlee, abhirast}@google.com

Abstract

Conversational recommendation systems (CRS) aim to recommend suitable items to users through natural language conversation. However, most CRS approaches do not effectively utilize the signal provided by these conversations. They rely heavily on explicit external knowledge e.g., knowledge graphs to augment the models' understanding of the items and attributes, which is quite hard to scale. To alleviate this, we propose an alternative information retrieval (IR)-styled approach to the CRS item recommendation task, where we represent conversations as queries and items as documents to be retrieved. We expand the document representation used for retrieval with conversations from the training set. With a simple BM25-based retriever, we show that our task formulation compares favorably with much more complex baselines using complex external knowledge on a popular CRS benchmark. We demonstrate further improvements using user-centric modeling and data augmentation to counter the cold start problem for CRSs.

1 Introduction

Recommendation systems have become ubiquitous in recent years given the explosion in massive item catalogues across applications. In general, a recommendation system learns user preference from historical user-item interactions, and then recommends items of user's preference. In contrast, CRSs directly extract user preferences from live dialog history to precisely address the users' needs. An example dialogue from the popular ReDial benchmark (Li et al., 2018) for CRSs is shown in Table 1: the CRS' task is to recommend items (in this case, movies) based on the user's indicated preference.

Generally, a CRS integrates two modules: a **dialogue module** which generates natural language responses to interact with users, and a **recommendation module** which recommends desirable items to users using the dialog context and external knowledge. We focus on the latter module in this work: we posit that once the correct item to recommend is identified, newer pretrained language models (PLMs) can easily generate fluent agent responses.

Role	Message
User	Hello! I am looking for some movies.
Agent	What kinds of movie do you like? I like animated movies such as Frozen (2013).
Rec. item	Frozen (2013)
User	I do not like animated films. I would love to see a movie like Pretty Woman (1990) starring Julia Roberts. Know any that are similar?
Agent	Pretty Woman (1990) was a good one. If you are in it for Julia Roberts you can try Runaway Bride (1999).
Rec. item	Runaway Bride (1999)

Table 1: An example dialogue from ReDial. The items to recommend are in blue, with their inferred attributes in red. The ground truth recommended items for agent utterances are also shown.

It is notable that the conversational context provides sufficient signal to make good recommendations (Yang et al., 2021). E.g., in Table 1, attributes about the items to recommend (e.g., genre and cast, in red) provide potentially sufficient information to the model to recommend relevant items.

Most approaches to CRS rely heavily on external knowledge sources, such as knowledge graphs (KGs) and reviews (Lu et al., 2021). Such approaches require specific sub-modules to encode information from these sources like graph neural networks (Kipf and Welling, 2016), which are hard to scale with catalog additions. Existing approaches require either re-training the entire system when the KG structure changes (Dettmers et al., 2018) or adding complex architectures on top to adapt (Wu et al., 2022). Newer approaches utilize PLMs (Radford et al.; Lewis et al., 2020), but they often encode item information in model parameters, making it hard to scale to new items without retraining.

Looking for a fast, more scalable approach, we re-formulate the item recommendation task for

Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023), pages 155–160

CRSs as an information retrieval (IR) task, with recommendation-seeking conversations as queries and items to recommend as documents. The document content for retrieval is constructed using plain text metadata for the item paired with conversations where the said item is recommended, in order to enhance semantic overlap between the queries which are themselves conversations.

We apply a standard non-parametric retrieval baseline - BM25 - to this task and show that the resulting model is fast and extensible without requiring complex external knowledge or architectures, while presenting improvements over more complex item recommendation baselines. Our contributions are summarized as follows:

- We present an alternate formulation of the CRS recommendation task as a retrieval task.
- We apply BM25 to this task, resulting in a simple, strong model with little training time and reduced reliance on external knowledge.
- We further improve the model using user-centric modeling, show that the model is extensible to new items without retraining, and demonstrate a simple data augmentation method that alleviates the cold start problem for CRSs.

2 Related Work

Conversational recommendation systems constitute an emerging research area, helped by datasets like REDIAL (Li et al., 2018), TG-REDIAL (Zhou et al., 2020b), INSPIRED (Hayati et al., 2020), DuRecDial (Liu et al., 2020, 2021), and CPCD (Chaganty et al., 2023). We next describe the recommender module architectures of CRS baselines.

ReDial (Li et al., 2018) uses an autoencoder to generate recommendations. CRSs commonly use knowledge graphs (KGs) for better understanding of the item catalog: DBpedia (Auer et al., 2007) is a popular choice of KG. KBRD (Chen et al., 2019) uses item-oriented KGs, while KGSF (Zhou et al., 2020a) further incorporates a word-based KG (Speer et al., 2017). CR-Walker (Ma et al., 2021) performs tree-structured reasoning on the KG, CRFR (Zhou et al., 2021) does reinforcement learning and multi-hop reasoning on the KG. UniCRS (Wang et al., 2022) uses knowledge-added prompt tuning with and KG & a fixed PLM. Some methods also incorporate user information: COLA (Lin et al., 2022) uses collaborative filtering to build a user-item graph, and (Li et al., 2022) aims to find lookalike users for user-aware predictions.

Eschewing KGs, MESE (Yang et al., 2022) trains an item encoder to convert flat item metadata to embeddings then used by a PLM, and TSCR (Zou et al., 2022) trains a transformer with a Cloze task modified for recommendations. Most above approaches, however, either rely on complex models with KGs and/or need to be retrained for new items, which is very frequent in present-day item catalogs.

3 Model

We formally define the item recommendation task, followed by our retrieval framework, details of the BM25 retrieval model used, and finally our user-aware recommendation method on top of BM25.

3.1 Conversational Item Recommendation

A CRS allows the user to retrieve relevant items from an item catalog $V = \{v_1, v_2 \cdots v_N\}$ through dialog. In a conversation, let a be an agent response containing an item(s) from V recommended to the user. Let $d_t = \{u_1, u_2, \cdots u_t\}$ be the t turns of the conversation context preceding a, where each turn can be spoken by the user or the agent.

We model the recommendation task as masked item prediction, similar to Zou et al. (2022). For each agent response a where an item $v_i \in V$ is recommended, we mask the mention of v_i in a i.e. replace it with the special token `[REC]`, yielding the masked agent response a'. We now create training examples with input $q = d_t \oplus a'$ and ground truth v_i ($\oplus$ denotes string concatenation).

We define Q^{train} and Q^{test} as the set of all conversational contexts $q = d_t \oplus a'$ with an item to predict, from the training and test sets respectively. For each item v_i, we also define $\boldsymbol{Q^{train}_{v_i}} \subset Q^{train}$ as the set of all conversational contexts in Q^{train} where v_i is the ground truth item to recommend.

3.2 Item Recommendation as Retrieval

Information retrieval (IR) systems are aimed at recommending documents to users based on the relevance of the document's content to the user query. We reformulate masked item prediction as a retrieval task with Q^{train} or Q^{test} as the set of queries to calculate relevance to, and V as the set of items/documents to recommend from.

To match a query $q \in Q^{test}$ to a document/item $v_i \in V$, we define the document's content using two sources: **metadata** in plaintext about item v_i, and $\boldsymbol{Q^{train}_{v_i}}$ i.e. all conversational contexts from the training set where v_i is the recommended item,

concatenated together, similar to document expansion (Nogueira et al., 2019). Our motivation for adding $Q_{v_i}^{train}$ to the document representation is that it is easier to match queries (which are conversations) to conversations instead of plain metadata since conversations can be sparse in meaningful keywords. For an item v_i we create a document as:

$$Doc(v_i) = Metadata(v_i) \oplus Q_{v_i} \quad (1)$$

For test set prediction, we can now apply retrieval to recommend the most relevant document $Doc(v_i), v_i \in V$, for each test set query $q \in Q^{test}$.

3.3 Retrieval Model: BM25

BM25 (Robertson et al., 2009) is a commonly used sparse, bag-of-words ranking function. It produces a similarity score for a given document, doc and a query, q, by matching keywords efficiently with an inverted index of the set of documents. Briefly, for each keyword in each document, we compute and store their term frequencies (TF) and inverse document frequencies (IDF) in an index. For an input query, we compute a match score for each query keyword with each document using a function of TF and IDF, and sum this score over all keywords in the query. This yields a similarity score for the query with each document, which is used to rank the documents for relevance to the query.

3.4 User Selection

Our IR formulation also gives us a simple way to incorporate user information for item recommendation. Let $U = \{u_1, u_2 \dots u_J\}$ be the set of all users in the dataset. Each conversation context in Q^{train} be associated with a user $u_j \in U$. We use a simple algorithm for user-aware recommendations:

- For each user $u \in U$, we obtain the set of items they like based on conversations in Q^{train}, and also construct a unique BM25 index for each user u_j using only conversations associated with u_j.
- For a test set query $q \in Q^{test}$, we identify movies liked by the seeker in the current q, and use it to find the M most similar users in the training set.
- We now compute and add up similarity scores for the query with all documents based on the per-user BM25 indices for these M selected users.
- Finally, we linearly combine these user-specific similarity scores per document with the similarity scores from the BM25 index in Section 3.3, and use these combined scores to rank all documents.

Model	R@1	R@10	R@50
ReDial (Li et al., 2018)	2.3	12.9	28.7
KBRD* (Chen et al., 2019)	3.0	16.4	33.8
KGSF* (Zhou et al., 2020a)	3.9	18.3	37.8
CR-Walker* (Ma et al., 2021)	4.0	18.7	37.6
CRFR* (Zhou et al., 2021)	4.0	20.2	39.9
COLA* (Lin et al., 2022)	4.8	22.1	42.6
UniCRS* (Wang et al., 2022)	5.1	22.4	42.8
MESE† (Yang et al., 2021)	5.6	25.6	45.5
TSCR* (Zou et al., 2022)	7.2	25.7	44.7
BM25 w/o Metadata	4.8	19.5	37.4
BM25†	5.2	20.5	38.5
BM25 + User Selection†	5.3	21.1	38.7

Table 2: Item recommendation results on the ReDial benchmark. Our BM25-based models outperform many baselines despite being much, lighter and not using complex KGs. * denotes models using DBPedia KG, † denotes models using plaintext IMDb metadata.

4 Experiments

4.1 Dataset and Evaluation

ReDial (Li et al., 2018) is a popular benchmark of annotated dialogues where a seeker requests movie suggestions from an agent. Figure 1 shows an example. It contains 956 users, 51,699 movie mentions, 10,006 dialogues, and 182,150 utterances.

For evaluation, we reuse Recall@k (or R@k) as our evaluation metric for ReDial from prior work. It evaluates whether the target human-recommended item appears in the top-k items produced by the recommendation system. We compare against baselines introduced in Section 2.

4.2 Training

For movie recommendations, we extract metadata from *IMDb.com* to populate $Metadata(v_i)$ for movies $v_i \in V$, which includes the movie's brief plot and names of the director and actors.

Parameters k_1 and b for BM25 are set to 1.6 and 0.7 respectively. For user selection, we select the $K = 5$ most similar users, and linearly combine the user-specific BM25 scores with the overall BM25 scores with a coefficient of 0.05 on the former. Constructing the BM25 index on the ReDial training set and inference on the test set took ~5 minutes on a CPU (+10 minutes for the user selection method). Alongside BM25 with and without user selection, we also experiment with a BM25 variant without metadata i.e. using only past conversation contexts as the document content for a movie/item.

5 Results

Table 2 shows $R@\{1, 10, 50\}$ on ReDial for the baselines and our models. Our BM25-based models perform strongly, outperforming many baselines which use complex KGs and/or complex model architectures e.g., tree-structured reasoning and reinforcement learning. Improvement is most visible on $R@1$ and less so on $R@50$. Our fairest comparison is with **MESE**, which uses the exact same data (text metadata + dialogues): our best model achieves 95% of its $R@1$ and 85% of its $R@50$ with a faster and simpler model. Note that all baselines except TSCR are jointly optimized for the item recommendation and response generation tasks, therefore their recommendation-only performance can potentially be better than reported.

A surprising result is **BM25 w/o Metadata** doing better than many baselines, without using any external knowledge whatsoever, in contrast to all other baselines except **ReDial**. This indicates that prior conversations indeed contain sufficient signal for good conversational item recommendation.

Our simple **user selection** raises recall by 1-3% across thresholds, with more potential gains from better user-centric modeling (Li et al., 2022).

6 Cold Start and Data Augmentation

Conversational recommenders often suffer from the **cold start problem**: it is difficult for a new item i.e. not seen during training, to be recommended, since not much is known about it beyond metadata.

Our model is not immune to this problem. The red lines in Figure 1 show $R@10$ values for the BM25 model for different sets of movies in ReDial based on how many times they are seen in the training set: the model never or rarely recommends movies with 10 or fewer occurrences in training.

To counteract this, we perform **data augmentation** using few-shot prompting (Liu et al., 2023). In particular, we randomly select 6 conversations from ReDial's training set, use them to prompt a PaLM 2-L model (Anil et al., 2023), and generate up to 20 dialogues per movie. We do this only for movies seen 10 or fewer times during training, since the model does the worst on these.

Figure 1's blue curve shows notably improved $R@10$ for the movies for which data was augmented, without hurting $R@10$ for more frequent movies. Overall $R@10$ also improves by ~8% using just ≤ 20 artificial dialogues per movie. Further

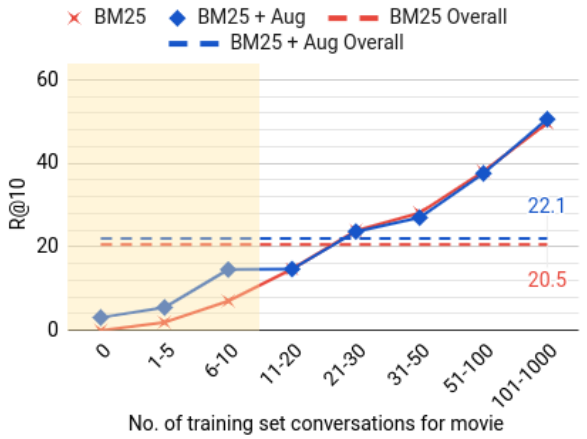

Figure 1: Impact of data augmentation on $R@10$. The shaded area represents the set of movies for which data augmentation was performed.

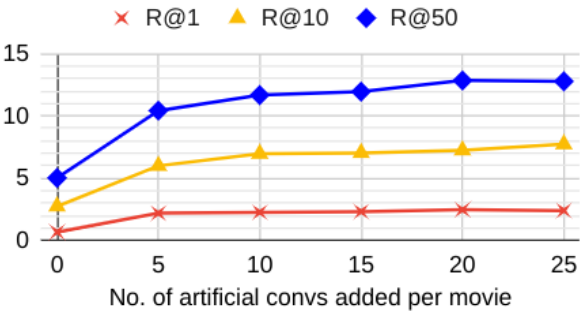

Figure 2: Recall for the BM25 model with varying amounts of augmented conversations.

combining augmentation with user selection lifts ***R*@1** to **5.9**, ***R*@10** to **22.3**, and ***R*@50** to **40.7**.

Figure 2 plots recall for BM25 model with the number of artificial dialogues added for low-frequency movies. Based on this plot, we opted to generate at most 20 conversations per movie.

7 Conclusion

We present a retrieval-based formulation of the item recommendation task, used to build CRSs, by modeling conversations as queries and items as documents. We augment the item representation with conversations recommending that item; the retrieval task then reduces to matching conversations to conversations. Using BM25-based retrieval with this task results in a model that is very fast and inexpensive to train (~5 min on CPU) while being flexible to add-ons like user selection. We also show that new items can be easily added without retraining the model, and that simple data augmentation with as few as 20 conversations counters the cold start problem for new items: fewer than most neural network finetuning methods would need.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*, pages 722–735. Springer.

Arun Tejasvi Chaganty, Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, and Filip Radlinski. 2023. Beyond single items: Exploring user preferences in item sets with the conversational playlist curation dataset. *ArXiv*, abs/2303.06791.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813, Hong Kong, China. Association for Computational Linguistics.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.

Shuokai Li, Ruobing Xie, Yongchun Zhu, Xiang Ao, Fuzhen Zhuang, and Qing He. 2022. User-centric conversational recommendation with multi-aspect user modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 223–233.

Dongding Lin, Jian Wang, and Wenjie Li. 2022. Cola: Improving conversational recommender systems by collaborative augmentation. *arXiv preprint arXiv:2212.07767*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Durecdial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049.

Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. RevCore: Review-augmented conversational recommendation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1161–1173, Online. Association for Computational Linguistics.

Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 2021. Cr-walker: Tree-structured graph reasoning and dialog acts for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1839–1851.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1929–1937.

Tianxing Wu, Arijit Khan, Melvin Yong, Guilin Qi, and Meng Wang. 2022. Efficiently embedding dynamic knowledge graphs. *Knowledge-Based Systems*, page 109124.

Bowen Yang, Cong Han, Yu Li, Lei Zuo, and Zhou Yu. 2021. Improving conversational recommendation systems' quality with context-aware item meta information. *arXiv preprint arXiv:2112.08140*.

Bowen Yang, Cong Han, Yu Li, Lei Zuo, and Zhou Yu. 2022. Improving conversational recommendation systems' quality with context-aware item meta-information. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 38–48, Seattle, United States. Association for Computational Linguistics.

Jinfeng Zhou, Bo Wang, Ruifang He, and Yuexian Hou. 2021. Crfr: Improving conversational recommender systems via flexible fragments reasoning on knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4324–4334.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020a. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1006–1014.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020b. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain, December 8-11, 2020*.

Jie Zou, Evangelos Kanoulas, Pengjie Ren, Zhaochun Ren, Aixin Sun, and Cheng Long. 2022. Improving conversational recommender systems via transformer-based sequential modelling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2319–2324.