You Sound Like Someone Who Watches Drama Movies: Towards Predicting Movie Preferences from Conversational Interactions

Sergey Volokhin[*], **Joyce Ho**[*], **Oleg Rokhlenko**[+], and **Eugene Agichtein**[*+]

[*]Emory University, GA USA
[+]Amazon, Seattle, WA, USA

{svolokh, joyce.c.ho}@emory.edu, {olegro, eugeneag}@amazon.com

Abstract

The increasing popularity of voice-based personal assistants provides new opportunities for conversational recommendation. One particularly interesting area is movie recommendation, which can benefit from an open-ended interaction with the user, through a natural conversation. We explore one promising direction for conversational recommendation: mapping a conversational user, for whom there is limited or no data available, to most similar external reviewers, whose preferences are known, by representing the conversation as a user's interest vector, and adapting collaborative filtering techniques to estimate the current user's preferences for new movies. We call our proposed method *ConvExtr* (Conversational Collaborative Filtering using External Data), which 1) infers a user's sentiment towards an entity from the conversation context, and 2) transforms the ratings of "similar" external reviewers to predict the current user's preferences. We implement these steps by adapting contextual sentiment prediction techniques, and domain adaptation, respectively. To evaluate our method, we develop and make available a finely annotated dataset of movie recommendation conversations, which we call *MovieSent* . Our results demonstrate that *ConvExtr* can improve the accuracy of predicting users' ratings for new movies by exploiting conversation content and external data.

1 Introduction and Background

With the increasing popularity of voice-assistants, there has been a lot of research on making well-established user experiences, like recommendations, conversational (Allan et al., 2012; Culpepper et al., 2018; Radlinski et al., 2019). One such area is movie recommendation.

Movie recommendation in general has been an actively researched area in conversational recommendation systems (Bennett et al., 2007; Khatri et al., 2018) and has been explored with a variety of approaches, most popular of which include collaborative filtering (CF) (Katarya and Verma, 2017; He et al., 2019), content-based filtering (Elahi et al., 2017), incorporating user reviews (Zhao et al., 2017; Dubey et al., 2018). Several attempts at both conversational recommendation (Christakopoulou et al., 2016; Sun and Zhang, 2018; Torbati et al., 2021), and specifically conversational movie recommendations (Dalton et al., 2018) have been made.

However, establishing the new user's preferences through a conversation, in order to make an effective recommendation, remains an open question, and there exists little conversational data for such a task. In our initial exploration, we examine if current well-researched approaches and algorithms can be adapted and used to help solve this problem and to establish a baseline of such approaches for future improvement and reference. We explore a new approach to conversational recommendation by incorporating preferences of other, external users with established preferences, via shared discussed entities, and the user's sentiment towards them, which also addresses the resulting "cold start" problem. In this setting, users do not ask for recommendations directly, but rather have a more natural conversation with a Wizard, and receive recommendations based on this discussion. Previous approaches such as the "hierarchy of recommendation goals" (Kang et al., 2017) and "narrative-driven recommendation" (Bogers and Koolen, 2017; Eberhard et al., 2019) are not applicable under these conditions.

Instead, we propose a novel *knowledge-aware conversational recommendation* approach, which combines conversational context with external knowledge, such as movie reviews, to predict users' ratings of unseen movies. To this end, we extended and refined an existing conversational dataset (Radlinski et al., 2019) to make it amenable to perform experiments in conversational recom-

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3091–3096
June 6–11, 2021.

mendation. A short snippet of one of the available conversations can be seen in the middle left box of Figure 1. The original dataset was created via an MTurk experiment, where two workers (one playing the Wizard and another playing the User) were asked to discuss several movies. The Wizard is "coached" to ask the most informative questions, prompting the User to express their opinions towards a mentioned movie. We extended this dataset to include entity annotations linked to RottenTomatoes[1], and fine-grained user sentiment labels (details in Section 3.1). We are making this dataset (*MovieSent*) publicly available. Next, Section 2 states the problem formally and describes our approach; Section 3 provides details of the models evaluated, and construction of the *MovieSent* dataset; Section 4 outlines results and states the conclusions of our work, Section 5 discusses the future work.

2 *ConvExtr* : Conversational Recommendation with External Data

2.1 Problem Statement and model overview

Given a prefix of k turns of conversation and mentions of m movies, we aim to predict the rating for the next movie $m+1$ to be mentioned in the conversation. In this paper's experimental setting, the value of m is set to 2, which approximates the average number of movies mentioned in a conversation with a voice assistant, but could be extended. In this setting, we would estimate the user's preferences based on the first two movies mentioned in the conversation, and predict their rating for a 3rd (yet unseen) movie. This tradeoff between the length of preference elicitation and the accuracy of recommendation could be explored in future work.

Our approach is illustrated in Figure 1. Given a conversation about movies, we estimate user sentiment towards the mentioned movies and use it as input to a CF model to predict the rating for an unseen movie. The CF model uses a large set of external critics' ratings and reviews, which should include critics similar to the current user. The final model uses 3 main inputs: (1) CF predictions for the unseen movie; (2) similarity between the conversation user and critics; (3) similarity between the conversation and the movies' metadata.

Specifically, we first construct the conversation representation and use pre-trained BERT to embed it using the sequence-encoder functionality of the model(Xiao, 2018), which gives us one vector per conversation. Then we infer the fine-grained user sentiment for the movies discussed in the conversation using a Random Forest model, trained on the labeled dataset *MovieSent* (described in 3.1). Since this is not the focus of our work, we evaluated the prediction performance on a development set against manually annotated sentiment labels, resulting in RMSE of 0.88 (mean over 10 tries, with std 0.06), which was sufficient for the current work.

The next step expands on a CF model (described in Section 2.2), constructed from an external reviews corpus, and predicts the score for the unseen movie. To make this prediction, we identify reviewers similar to the current conversation user, via the similarity of their reviews to the current conversation text. We then calculate BERT based sentence embeddings for all reviews of those critics and represent each critic as a centroid of their review vectors. Finally, we use the similarity between the conversation- and critics' representations to transform the critics' scores to predict the conversational users' ratings.

2.2 Collaborative Filtering

Collaborating filtering (CF) has been shown to be an effective approach for recommendation. We experimented with item-based CF algorithms, including variants of K-Nearest Neighbors (KNN)-based algorithms with Mean Squared Difference, Cosine, and Pearson similarity metrics, as well as Singular Value Decomposition (SVD) and SVD++ algorithms, available as part of the *surprise* [2] Python package. We report the results using the SVD++ model for Collaborative Filtering (Vozalis and Margaritis, 2007), which exhibited the best performance in development experiments.[3] To use CF in our setting, after inferring the user's sentiment towards the mentioned movies, the sentiment scores are converted to ratings and provided to the CF model to estimate the user's sentiment towards a new movie.

Domain Adaptation: From Critic Ratings to User Preferences: Our research indicates that Critics, who are paid professionals, significantly differ from Conversational Users. Therefore we need other features to be able to adapt the score from Critics space to Users space. To achieve this, we computed the dot product and the earth-movers dis-

[1] www.rottentomatoes.com/

[2] https://surprise.readthedocs.io/en/stable/

[3] other models are omitted due to space limitations

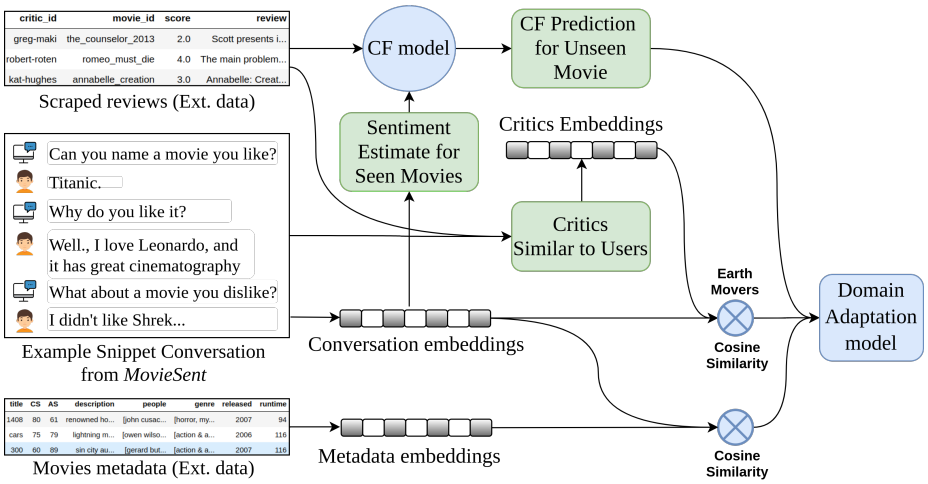

Figure 1: Overview of the *ConvExtr* system for conversational elicitation and prediction of movie preferences.

tance between conversation vector and weighted critics vector as features. Other features are created by using either raw movie metadata (year of release, runtime, number of critics' and users' reviews, RottenTomatoes average critics score, RottenTomatoes average users score, all used as numbers) or dot product between vectors of BERT sequence embedded movie features (title, description, actors, genre) and conversation vector, which gives us 10 metadata-based features. We hypothesize that including these features would let the model learn the difference between Critics and Users, and map or merge the scores from one source to the other.

We experimented with different model implementations, including pure CF without domain adaptation and a GBRT[4] model trained to translate the critics' preferences to user scores.

3 Experimental Setup

Baselines: First, we establish natural baselines:

- *AverageCritics*: Critics score from RT, which is the popularity proxy in a cold start problem
- *AverageAudience*: Audience score from RT, another popularity proxy, which potentially is closer to Conversational Users, than Critics

Evaluation Metrics: To evaluate the sentiment prediction performance and the overall system performance, we use the standard RMSE[5] and MAE[6] metrics.

[4]Gradient Boosted Regression Tree
[5]Root Mean Squared Error
[6]Mean Average Error

3.1 *MovieSent* : Sentiment Elicitation Dataset

The conversational Movie Sentiment Elicitation Dataset (*MovieSent*) that we created, is an extension to the dataset released in reference(Radlinski et al., 2019), which consists of Preference Elicitation conversations between "coached" crowd workers, playing the roles of Wizards and Apprentices. However, the movies mentioned in the dataset were not linked to a unique identifier, which required additional manual annotation to benefit from external knowledge. Hence, we manually labeled all movies in the dataset with their RottenTomatoes ID. Then, we asked human annotators to label each user response with a sentiment score on [-3; +3] scale, as well as a "None" score. The labeling was done by 8 independent judges with a 20% overlap (at most 2 people labeled the same sample). Inter-rater reliability for judges agreement on the labels was calculated using Cohen's kappa (Cohen, 1960) for binary labels, which is standard for this task, and it was 0.90 on 238 samples, indicating substantial inter-rater agreement. Reliability for the numerical sentiment was measured using a weighted Kappa (Cohen, 1968), it was 0.77 on 163 samples. An example can be found in Figure 2.

Reviews dataset construction: As mentioned in Section 1, most of the existing movie rating datasets are not suitable for our task, therefore we had to create a new dataset, to be released publicly. The basis of our CF system was Critics' ratings from an external source, specifically, a popular website RottenTomatoes. To construct the corpus, for each movie in *MovieSent* , we retrieved unique IDs for Critics who left reviews on that movie's page. We then retrieved all the reviews those critics have

	Conv ID	Utt number	Wizard Utterance	User Utterance	Entity	Judge Label
0	CCPE-0126d	3	Perfect! Now, what would be one of your favorite movies?	I love Mr. and Mrs. Smith. That's a great one.	Mr. & Mrs. Smith	3
1	CCPE-0ef1f	22	I can see why you wouldn't be interested in those, have you seen the shape of water	I started watching that, but I just couldn't get into it enough to finish it.	The Shape of Water	-2
2	CCPE-0d49b	9	Have you seen Bridesmaids?	Nope.	Bridesmaids	None

Figure 2: Example of utterances labeled with sentiment scores

Table 1: Statistics of the reviews dataset, and *MovieSent* annotated conversational movie sentiment dataset.

Reviews dataset		*MovieSent*	
Reviews	715,766	Conversations	489
Critics	3,664	Sentiment labels	2,488
Med.Reviews / Critic	34	Unique entities	712
Unique Movies	42,423		

ever left for any movie and normalized the numerical ratings to a discrete scale from 1 to 5. We used the resulting sparse matrix of critics-to-movies rating scores as input to the CF algorithm, described in Section 2.2. The statistics of the created datasets are reported in Table 1.

Conversation Representation for User Sentiment Inference: While not the main focus of our work, our method requires estimating the user's sentiment towards the mentioned movies. We experimented with different representations, finally picking the concatenation of the previous Wizard utterance and the current user utterance, resulting in RMSE of 0.96 and MAE of 0.72 of the predicted sentiment against human annotations. We use this sentiment prediction setup for all experiments.

3.2 Experimental procedure

To conduct an informative evaluation of our methods, we restrict the set of conversations in *MovieSent* to include only those, which had at least 3 movies with IDs mentioned in separate utterances, each of which had reviews in the corpus described above. The resulting conversational dataset contained 238 conversations out of initial 489. All experiments were conducted using 5-fold cross-validation, with 48 conversations on average in each split.

4 Results and Discussion

Results for all discussed models are reported in Table 2. Our method (last row) uses a natural conversation to both estimate the user's sentiment for a movie, and to retrieve relevant Critics to estimate the rating of this User for an unseen movie. Both baselines performed similarly, with RMSE of around 1.3. The improvements that our models were able to achieve were significant, with the best model, using GBRT, achieving RMSE of 0.96 and MAE of 0.72 (+25% improvement on both metrics). Finally, to gain intuition on the performance of *ConvExtr* , we simulated the best performance possible with CF (using SVD++ model), if the conversational user were one of the critics in the Reviews dataset. The resulting predictions can be considered an upper bound difficult to reach, as conversational preference elicitation can at best approximate the true user's preferences.

Table 2: Main results: RMSE and MAE errors (lower better) for predicting user preferences (best in bold), significance from AvgAudience baseline marked with "*")

Model	RMSE	MAE
Baseline methods:		
5 AverageCritics	1.34	0.99
AverageAudience	1.24	0.95
ConvExtr (our method):		
KNN (no adapt.)	1.20	0.94
SVD (no adapt.)	1.18*	0.95
SVD++ (no adapt.)	1.14*	0.92
GBRT	**1.09***	**0.84**
Best possible:	0.84	0.64

We explored the problem of Conversational Movie Recommendation, to take advantage of the vast amounts of external user-generated content on the Web, such as movie reviews, to improve the recommendation quality. As a first critical step in that direction, we focused on estimating a user's preference for an unseen movie based on estimated sentiment towards other movies mentioned in the conversation. Specifically, we applied sentiment analysis models to infer a User's sentiment towards movies mentioned in a conversation. These sentiment scores were used as a proxy for a user's explicit ratings which could be used by traditional

Collaborative Filtering algorithms, applied to extensive external data of movie reviews and ratings. Our second insight was that the actual conversation content could provide additional benefit in representing the user's interests, for improved recommendation quality.

Our results demonstrated that incorporating conversation content to select a more similar group of users for Collaborative Filtering, improves the recommendation performance, compared to using the inferred ratings alone. To advance research in this area, and for full transparency and reproducibility, the labeled conversation dataset and the code to retrieve the external review data are available on GitHub[7].

Together, our contributed dataset and experiments, and the resulting insights, offer a promising direction for improving conversational recommendation systems through augmentation with external data.

5 Future work

This work was only a first step, with many potential areas for further improvement. The baseline comparison could be improved: adapting and using more sophisticated methods, like Neural Matrix Factorization(He et al., 2017), or Wide and Deep Learning models(Cheng et al., 2016) would make for a better baseline, and it might also improve the performance of our approach as well, as the CF score has one of the highest feature importance in our model. Another direction is trying content-based approaches, which were not used in the current paper due to scarcity of data for each conversational user.

We also observed that our model was biased against predicting lower ratings since the conversations tend to focus on movies that a user liked. In future work we plan to explore correcting for this positive bias and other extensions to predicting user sentiment from a conversation more robustly. While our initial attempts to represent the conversation content improved the prediction accuracy, a fruitful direction of research is improving the representation of the conversation content for recommendations (i.e., for retrieving similar reviewers).

In Section 2.1 we mention k-turn conversations and mention of m movies. So a natural direction for research would be analysis of recommendation accuracy depending on the length of the conversation (number of turns), on the number of mentioned movies (does more movies equal better quality?), and on the ratio of sentiment bearing statements in a conversation (how many are factual/neutral?).

As additional conversational data becomes available, our approach could be extended to include other sources of user preferences such as Twitter/Reddit-based conversations, and actual past conversations of other users with a recommender bot.

[7] https://github.com/sergey-volokhin/conversational-movies

Acknowledgements

This work was partially supported by a grant from Amazon Alexa towards the study of conversational search and recommendation.

References

James Allan, Bruce Croft, Alistair Moffat, and Mark Sanderson. 2012. Frontiers, challenges, and opportunities for information retrieval: Report from swirl 2012 the second strategic workshop on information retrieval in lorne. In *ACM SIGIR Forum*, volume 46, pages 2–32. ACM.

James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *KDD*, volume 2007, page 35. New York, NY, USA.

Toine Bogers and Marijn Koolen. 2017. Defining and supporting narrative-driven recommendation. In *RecSys*, RecSys '17, pages 238–242, New York, NY, USA. ACM.

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *SIGKDD*, KDD '16, pages 815–824, New York, NY, USA. ACM.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jason Cohen. 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological bulletin*, 70 4:213–20.

J Shane Culpepper, Fernando Diaz, and Mark D Smucker. 2018. Research frontiers in information retrieval: Report from the third strategic workshop on information retrieval in lorne (swirl 2018). In *ACM SIGIR Forum*, volume 52, pages 34–90. ACM.

Jeffrey Dalton, Victor Ajayi, and Richard Main. 2018. Vote goat: Conversational movie recommendation. In *SIGIR*, SIGIR '18, pages 1285–1288, New York, NY, USA. ACM.

Abhishek Dubey, Ayush Gupta, Nitish Raturi, and Pranshu Saxena. 2018. Item-based collaborative filtering using sentiment analysis of user reviews. In *ICACCT*, pages 77–87. Springer.

Lukas Eberhard, Simon Walk, Lisa Posch, and Denis Helic. 2019. Evaluating narrative-driven movie recommendations on reddit. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 1–11, New York, NY, USA. ACM.

Mehdi Elahi, Yashar Deldjoo, Farshad Bakhshandegan Moghaddam, Leonardo Cella, Stefano Cereda, and Paolo Cremonesi. 2017. Exploring the semantic gap for movie recommendations. In *RecSys*, pages 326–330. ACM.

Ming He, Bo Wang, and Xiangkun Du. 2019. Hi2rec: Exploring knowledge in heterogeneous information for movie recommendation. *IEEE Access*, 7:30276–30284.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.

Jie Kang, Kyle Condiff, Shuo Chang, Joseph A. Konstan, Loren Terveen, and F. Maxwell Harper. 2017. Understanding how people use natural language to ask for recommendations. In *RecSys*, RecSys '17, pages 229–237, New York, NY, USA. ACM.

Rahul Katarya and Om Prakash Verma. 2017. An effective collaborative movie recommender system with cuckoo search. *Egyptian Informatics Journal*, 18(2):105–112.

Chandra Khatri, Behnam Hedayatnia, and Anu Venkatesh et al. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize.

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *SIGdial*.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *SIGIR*, SIGIR '18, pages 235–244, New York, NY, USA. ACM.

Ghazaleh Haratinezhad Torbati, Andrew Yates, and Gerhard Weikum. 2021. You get what you chat: Using conversations to personalize search-based recommendations. In *43rd European Conference on IR Research*. Springer.

M.G. Vozalis and K.G. Margaritis. 2007. Using svd and demographic data for the enhancement of generalized collaborative filtering. *Information Sciences*, 177(15):3017 – 3037.

Han Xiao. 2018. bert-as-service. `https://github.com/hanxiao/bert-as-service`.

Zhou Zhao, Qifan Yang, Hanqing Lu, Tim Weninger, Deng Cai, Xiaofei He, and Yueting Zhuang. 2017. Social-aware movie recommendation via multimodal network learning. *IEEE Transactions on Multimedia*, 20(2):430–440.