Data, methods, and evaluation for knowledge-grounded conversational recommendation systems

Julian McAuley
Conversational recommender systems

We’ll explore:

● How can we improve conversational recommender systems by creating new datasets?
● How can we combine modify language models to increase controllability with respect to recommended items?
● How can we better evaluate conversational recommender systems, especially with respect to beyond-accuracy metrics?
What is conversational recommendation?
Some traditional approaches...

Traditional approaches rarely involved “conversation” as we might normally think of it:

- Thompson et al., 2004 (query refinement): Elicits users’ preferences and constraints with regard to item attributes;
- Mahmood and Ricci, 2009 (reinforcement learning): Queries users about recommendation attributes during each round; learns a policy to choose queries to efficiently yield a desirable recommendation

(from Thompson et al.)
Some traditional approaches...

Traditional approaches rarely involved “conversation” as we might normally think of it:

- Christakopoulou et al., 2016 (iterative recommendation): Collects feedback about recommended items in order to iteratively learn user preferences; explores various query strategies to elicit preferences quickly.

---

**Greedy**: \( j^* = \arg \max_j y_{ij} \)

A trivial exploit-only strategy: Select the item with highest estimated affinity mean.

**Random**: \( j^* = \text{random}(1,N) \)

A trivial explore-only strategy.

**Maximum Variance (MV)**: \( j^* = \arg \max_j \epsilon_{ij} \)

A explore-only strategy, variance reduction strategy: Select the item with the highest noisy affinity variance.

**Maximum Item Trait (MaxT)**: \( j^* = \arg \max_j \| \mathbf{v}_j \|_2 \)

Select the item whose trait vector \( \mathbf{v}_j \) contains the most information, namely has highest L2 norm

\[
\| \mathbf{v}_j \|_2 = \sqrt{v_{j1}^2 + v_{j2}^2 + \cdots + v_{jd}^2}.
\]

**Minimum Item Trait (MinT)**: \( j^* = \arg \min_j \| \mathbf{v}_j \|_2 \)

Select the item with trait vector with least information.

**Upper Confidence (UCB)**: \( j^* = \arg \max_j y_{ij} + \epsilon_{ij} \)

Based on UCB1 [3]: Pick the item with the highest upper confidence bound, namely mean plus variance (95% CI).

**Thompson Sampling (TS)** [5]: \( j^* = \arg \max_j y_{ij} \)

For each item, sample the noisy affinity from the posterior. Select item with the maximum sampled value.

(from Christakopoulou et al.)
Related: “explainable” recommendations

Explainable recommenders associate natural language explanations with each recommendation (or something like this)

Such models represent “half” of a conversational model, though lack interactive mechanisms for the user to participate in conversation

(from Xie et al., 2022)
Limitations of traditional approaches

- Well short of what we’d normally call “conversation”
- Models may be interactive but probably couldn’t be implemented via conversational interfaces
- A lack of compelling datasets to train “real” conversational recommenders
- Most methods not particularly “general purpose”: typically designed around domains or datasets with specific data/features available
Actual conversation...

Li et al. (2018) sought approaches more closely matching “free-form” conversation. Roughly:

- Dialogs (around 10k) are constructed by crowd workers, who assume roles of a recommender or seeker;
- Conversations between the recommender and the seeker are tagged in terms of the movies mentioned, as well as explicit feedback (has the seeker seen the movies mentioned and did they like them);
- Train a dialog generation model that can fulfil the role of the recommender;
- Preferences can then be estimated and the output controlled to reference specific movies

Table 5: Sample conversation from validation set

<table>
<thead>
<tr>
<th>SEEKER:</th>
<th>HUMAN:</th>
<th>HRED:</th>
<th>OURS:</th>
</tr>
</thead>
<tbody>
<tr>
<td>hi !</td>
<td>hi ? how is your morning going ?</td>
<td>hi , what kind of movies do you like</td>
<td>hello ! what kind of movie are you looking for ?</td>
</tr>
<tr>
<td>SEEKER:</td>
<td>great , thanks ! i was thinking of starting the day off with a great comedy , any suggestions ?</td>
<td>i love the hangover it always makes me laugh . and i bet lots of people are hangover this morning ! airplane ! is another classic of mine .</td>
<td></td>
</tr>
<tr>
<td>HUMAN:</td>
<td>i have n’t seen that one ! what kind of movies do you like ? i can help you with that .</td>
<td>that ’s a good one ! have you seen the lego movie ? it ’s a great movie !</td>
<td></td>
</tr>
<tr>
<td>SEEKER:</td>
<td>i ’ve seen the hangover it ’s really good ! i thought about watching knocked up have you seen it ?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUMAN:</td>
<td>i enjoyed that movie . i enjoyed spy it ’s a well done parody .</td>
<td>it ’s about a man who gets revenge on a cyborg and a hockey team . it ’s really good . it ’s really good . i also like the girl on the train</td>
<td></td>
</tr>
<tr>
<td>HRED:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OURS:</td>
<td>i have n’t seen that one . i will have to check it out . i hope i was able to help you find a good one ! have a great day !</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEEKER:</td>
<td>thanks ! i ’ll check it out !</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUMAN:</td>
<td>sure ! have a great day !</td>
<td>you ’re welcome ! have a great day !</td>
<td></td>
</tr>
<tr>
<td>HRED:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OURS:</td>
<td>have you seen revenge of the nerds ? it ’s really good . it ’s a classic .</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEEKER:</td>
<td>you too .</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(from ReDial)
Actual conversation...

Li et al. (2018)’s approach has a number of virtues:

- Actually looks (more or less) like “real” conversation, especially compared to approaches that came before
- Contributes a (now widely used) benchmark dataset for training and evaluation
- Elegant / principled in terms of how the model is trained and the objective it’s trained for (i.e., reach a goal movie in the fewest possible number of steps)
Actual conversation...

Though it also has some limitations:

- Conversations aren’t particularly “real”: the users aren’t actually seeking some item, but play a synthetic game in which they are told which item to seek.
- It’s unclear to what extent the data collection effort could be applied in other settings, in particular ones not based on “general knowledge” (i.e., for which crowd workers would struggle to engage in synthetic conversations).
- Even within movies, it’s hard to tell how closely conversations in ReDial (or similar efforts) represent “organic” conversations.
“LM+RecSys” approaches (UniCRS; Wang et al., 2022)

(Fairly) recent attempts incorporate knowledge grounding, and arguably (among a few others) represented the pre-LLM state-of-the-art
pre-LLM state-of-the-art (ReDial benchmark)

Lots of progress on the ReDial benchmark in the last few years:
Datasets
Datasets - an example from ReDIAL

Dialog_ID: 391  Seeker_ID: 0  Recommender_ID: 1  Dialog Turns:

[User]: Hi there, how are you? I'm looking for movie recommendations.
[User]: I like animations like @84779 and @191602
[User]: I also enjoy @122159
[User]: Anything artistic
[User]: What's it about?
[User]: That seems like a nice comedy
[User]: Do you have any animated recommendations that are a bit more dramatic? Like @151313 for example
[User]: I like comedies but I prefer films with a little more depth
[User]: Ooh that seems cool! Thanks for the input. I'm ready to submit if you are.
[User]: Nice
[User]: Take care, cheers!

I am doing okay. What kind of movies do you like? : [System]
You might like @165710 that was a good movie. : [System]
It has Alec Baldwin it is about a baby that works for a company and gets adopted it is very funny .........
That is a tough one but I will remember something @203371 was a good one : [System]
It is animated, sci fi, and has action : [System]
Glad I could help : [System]
bye : [System]
Datasets

- How can datasets be built that are more *natural*? E.g. actually how humans would interact when making movie recommendations, versus current, more synthetic, settings?
- Other efforts (e.g. INSPIRED) aim for a more natural setting, but are also very small.
- Need datasets that are **bigger** and **more realistic**
- Our previous efforts (e.g. to synthesize conversational datasets from product review text) were much larger but of low quality.
Can we build better datasets, e.g. by harvesting data from “natural” conversations?
Reddit-Movie Dataset

- 634,392 movie recommendation conversations, featuring 1.7M dialog turns
- ~11k users, ~24k items
- (compare to e.g. ReDial, featuring ~10k conversations, ~139k turns, ~800 users)

Much bigger than existing datasets; conversations are shorter; they have much more context; and (for better or worse) have much more varying structure
What do these new datasets reveal?

We use a simple prompting setup to compare LLMs:
What do these new datasets reveal?

Some observations about model performance:

- Existing models engage in *shortcut learning* by focusing on repeated items (i.e., items already mentioned in a dialog but not as recommendations)
- LLMs outperform existing fine-tuned models; GPT-4 outperforms other LLMs
- LLMs generate some out-of-dataset items, but not many hallucinated recommendations (<5%); can be dealt with by string matching
What do these new datasets reveal?

Some observations about model performance:

- Significant “popularity bias” (and other bias) issues
- Recommendation performance is highly sensitive to geographical region (presumably just due to groundtruth frequency)
Methods
Control strategies for LLM-based recommenders

- How can we make LLM-based recommenders more **controllable**

Solutions (roughly) based on RAG

1. Retrieve *recommendation probabilities* from a recommender; use these to change item probabilities during decoding
2. Retrieve *related items* to improve prompting-based methods
3. Retrieve *related training samples, or knowledge*
1. **Item “re-indexing”**

Zhankui He + others

- How can we address the issue of *distribution imbalance* with LLM-based recommenders?
- Essentially, distribution of recommended items should match the distribution we observed in the training set
- Lots of solutions for conventional methods, somewhat of an open problem for conversational approaches!
1. **Item “re-indexing”**

- **LLMs have poor controllability of recommended items since they surface tokens rather than surfacing items**
- **We adapt language models to associate unique tokens (or possibly multiple tokens) with every item**
- **This gives us more direct control over the item distribution during decoding, and allows for easy ensembling with traditional recommender systems**
2. Collaborative retrieval

- Can we retrieve related items (or interactions) to use as prompts to construct evidence for or against particular recommendations?

- Role-play: As a recommender system please solve the following problem.
  - Collaborative Information: Repeat $i \in I^\text{coll}_z$
    \begin{align*}
    \left\{ \begin{array}{l}
    \text{The item desc}_I(i) \text{ is liked by the users POS}(i, U^\text{coll}_z). \\
    \text{The item desc}_I(i) \text{ is disliked by the users NEG}(i, U^\text{coll}_z).
    \end{array} \right.
    \end{align*}

- Summarization: Try to understand the pattern that the item desc}_I(i) is typically liked by what kinds of users based on the above information.
2. Collaborative retrieval

- Can we retrieve related items (or interactions) to use as prompts to construct evidence for or against particular recommendations?

- **User’s Positive Preference**: Items the user $\text{idx}_U(u)$ likes are as follows: $\text{POS}(I_z^{supp}, u)$.
- **User’s Negative Preference**: Items the user $\text{idx}_U(u)$ does not like are as follows: $\text{NEG}(I_z^{supp}, u)$.
- **Query**: For the item described as $\text{idx}_I(i)$, would you recommend it to the user $\text{idx}_U(u)$?
2. Collaborative retrieval

- Can we retrieve related items (or interactions) to use as prompts to construct evidence for or against particular recommendations?
- The main variable in this method is then to choose a policy for selecting *which related users and items should be retrieved*; which we learn via RL
3,4... Other alternatives?

- Retrieve related training samples
- Retrieve related knowledge
Evaluation
Decomposing “realism” into key properties

Can we do better than held-out item prediction?

- Users may interact with conversational recommenders precisely because they struggle to articulate their preferences, or because they need to be persuaded to select a particular item;
- User studies are expensive, and generally non-reproducible.
- Outside of industrial settings, user studies generally don’t involve ‘real’ users;
- User studies may be suitable for ‘general knowledge’ items and domains, but are unsuitable in cases where users requiring specific knowledge or expertise may be difficult to recruit.
Evaluation in conversational recommendation

Static evaluation

Dataset (fixed)

Human agent

Human seeker

Compare

Interactive evaluation

User simulator

Do you like action?

Yes

Do you like Tom Cruise?

Yes

Then you might like Mission Impossible (1996)

✓ Cost-effective

✓ Multi-turn strategies
User simulators are far from real users

<table>
<thead>
<tr>
<th>Simulators</th>
<th>Real users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preference</td>
<td>Binary (yes/no)</td>
</tr>
<tr>
<td>Behavior</td>
<td>Predefined</td>
</tr>
</tbody>
</table>

Real users do not follow pre-programmed rules!
LLMs as replicas of human behavior

[Image: Joining for coffee at a cafe]

[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?

Generative agents


Participants in social science experiments

I am looking for a kids’ movie.

Here are some great kids’ movies: Toy Story (1995), Finding Nemo (2003) ...

I’m actually looking for something a bit older.

Certainly! If you're looking for older kids’ movies: The Goonies (1985), E.T. (1982)… These movies have stood the test of time and are still beloved by kids and adults alike.

Final feedback: Reject (Target item was Mulan (1998))

Persuasiveness score: 0/2

Wang et al., 2023. Rethinking the evaluation for conversational recommendation in the era of large language models. In EMNLP.
How to evaluate a user simulator?

Compare a population of simulators and real users

- A single simulator may count as just one hypothetical person
- Goal of recommendation is to satisfy a group of users (+ unknown user)
Is the distribution of items that simulators mention similar to real users?

Data point from a human user

Prompt
A person mentions *Concussion (2015)* and *Jerry Maguire (1996)* in a conversation about movies and proceeds to mention 2 more. What would these 2 movies be?

Simulators mention much less diverse items than humans
(Task 2) BinPref

Can simulators reflect real user preferences?

You watched the movie Whiplash (2014). Did you like the movie? Answer Yes or No. Don’t say anything else.

Average user rating

4.25 / 5

Prompt (given to 200 simulators)

Simulators struggle to reflect human preference

Prompting them to be “picky” may enhance alignment
(Task 3) OpenPref

Do simulators express preferences in ways real users do?

Review from human user

La La Land (2016):
I would say it’s a movie that is more suitable for a background music instead of putting full focus, because the story was not very convincing.

Prompt

You watched the movie La La Land (2016)
What are your thoughts on this movie? Answer should not exceed 164 characters.

Aspect-based sentiment analysis

- Simulators express more aspect-based sentiments (e.g., “cast is good, plot is bad”)
- Simulators are biased toward positive sentiments

<table>
<thead>
<tr>
<th>Generator</th>
<th># aspects</th>
<th>Aspect entropy</th>
<th>Sentiment entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>85</td>
<td>5.85</td>
<td>1.19</td>
</tr>
<tr>
<td>Demographic information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gpt-3.5</td>
<td>71</td>
<td>4.86</td>
<td>0.29</td>
</tr>
<tr>
<td>gpt-4</td>
<td>97</td>
<td>5.57</td>
<td>1.11</td>
</tr>
<tr>
<td>text-davinci</td>
<td>194</td>
<td>5.63</td>
<td>0.18</td>
</tr>
<tr>
<td>Demographic information + Pickiness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gpt-3.5</td>
<td>101</td>
<td>5.20</td>
<td>1.09</td>
</tr>
<tr>
<td>gpt-4</td>
<td>97</td>
<td>5.59</td>
<td>1.34</td>
</tr>
<tr>
<td>text-davinci</td>
<td>232</td>
<td>5.47</td>
<td>0.48</td>
</tr>
</tbody>
</table>
(Task 4) *RecRequest*

Can simulators generate **diverse, personalized requests**?

**Prompt**

Generate a movie recommendation request. Include the following movies in your text: *Taxi Driver* (1976), *Joker* (2019). Length of the request is approximately 176 characters.

<table>
<thead>
<tr>
<th>Human requests</th>
<th>gpt-3.5-turbo requests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movies showing extreme loneliness or depression. I have watched <em>Taxi Driver</em> (1976) and <em>Joker</em> (2019) and would like to see more similar movies showing loneliness or depression.</td>
<td>Looking for a gripping psychological thriller similar to <em>Taxi Driver</em> (1976) or <em>Joker</em> (2019)? Seeking a movie that delves into the mind of complex characters?</td>
</tr>
</tbody>
</table>
(Task 4) RecRequest

Can simulators generate diverse, personalized requests?


- **gpt-3.5-turbo**: ‘Movie recommendation?’, ‘Need movie suggestions’, ‘Need movie recs!!’, ‘Need movie recommendations’, ‘Movie recs?’, ‘Movie recommendations?’

- **gpt-4**: ‘Got recs?’

- **text-davinci-003**: ‘Recommend a movie’, ‘Cheerful movies?’, ‘Recommend me!'

**Simulator requests are much less diverse**

**Case study: shortest requests**
(Task 5) Feedback

Can simulators give **coherent feedback**?

Request from a real user

I’m in a state of life rn that i really want/need movies with the Main Character being a loner or alone in general.

Simulators can give incoherent feedback.

Positive feedback: 
- Accept

Negative feedback: 
- Reject
Decomposing “realism” into key properties

- The above is a suite of evaluations to determine the fidelity of an existing simulator.
- There is still a lot of work to be done to use this to design better simulators!
- Other than prompting strategies, we’d like to explore closed-loop training, and to incorporate simulators with the control strategies mentioned previously.
Summary

- Conversational recommendation represents a promising frontier in building recommender systems that are more “human-like”
- This line of research has been somewhat blown open by the excellent performance of general-purpose language models
- There’s still plenty to do (even if, arguably, less of it is about modeling...)
- Many “traditional” questions about recommender systems (evaluation, fairness, etc.) have new life in light of conversational paradigms
Thanks!