What's still hard about conversational recommendation?

Julian McAuley

Conversational recommender systems

We'll explore:

- What has conversational recommendation "traditionally" looked like?
- What was (until recently) the state-of-the-art?
- What is currently the state-of-the-art (hint: LLMs!)
- What is still hard and what are the interesting open problems in this space?

Conversational recommender systems

Conversational recommendation very broadly refers to any type of recommender system that involves interaction, usually featuring natural language. This could include:

- Free-form conversation with a goal of finding a relevant item (Li *et al* 2018)
- Systems that produce recommendations with a natural language interface (e.g. Amazon Alexa)
- Systems that explain recommendations, and the user provides traditional (i.e., click) feedback
- Systems that augment dialogs between humans with recommendations

Traditional Approaches

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

User Name		Homer					
Attributes	w_i	Values and probabilities					
Cuisine	0.4	Italian	French	Turkish	Chinese	German	English
		0.35	0.2	0.25	0.1	0.1	0.0
Price Range	0.2	one	two	three	four	five	
		0.2	0.3	0.3	0.1	0.1	
...	...	...					
Parking	0.1	Valet		Street		Lot	
		0.5		0.4		0.1	

Item Nbr.	0815	5372	7638	...	6399
Accept/Present	23 / 25	10 / 19	33 / 36	...	12 / 23

(from Thompson et al.)

Some traditional approaches...

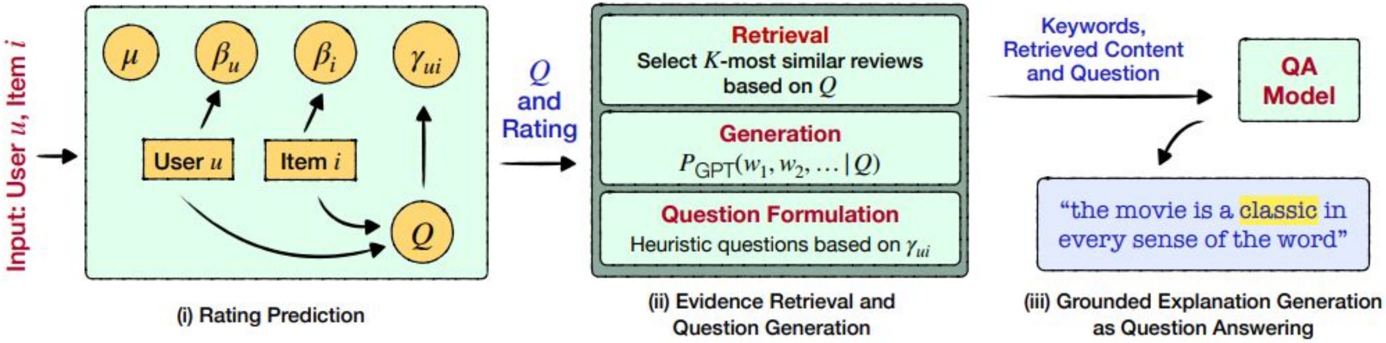
Traditional approaches rarely involved "conversation" as we might normally think of it:

- 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): $j^* = \arg\max_j \epsilon_{ij}$
An *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 $\|\mathbf{v}_j\|_2 = \sqrt{v_{j1}^2 + v_{j2}^2 + \ldots + v_{jd}^2}$.
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]: $j^* = \arg\max_j \hat{y}_{ij}$
For each item, sample the noisy affinity from the posterior. Select item with the maximum sampled value.

(from Christakopoulou et al.)

Related: “explainable” recommendations



(from Xie et al., 2022)

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

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


More “Conversational” Approaches

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

SEEKER:	hi !
HUMAN:	hi ! how is your morning going ?
HRED:	hi , what kind of movies do you like
OURS:	hello ! what kind of movie are you looking for ?
SEEKER:	great , thanks ! i was thinking of starting the day off with a great comedy , any suggestions ?
HUMAN:	i love the hangover it always makes me laugh . and i bet lots of people are hungover this morning ! airplane ! is another classic of mine .
HRED:	that ’s a good one ! have you seen the lego movie ? it ’s a great movie !
OURS:	i have n’t seen that one ! what kind of movies do you like ? i can help you with that .
SEEKER:	i ’ve seen the hangover it ’s really good ! i thought about watching knocked up have you seen it ?
HUMAN:	i enjoyed that movie . i enjoyed spy it ’s a well done parody .
HRED:	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
OURS:	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 !
SEEKER:	thanks ! i ’ll check it out !
HUMAN:	sure ! have a great day !
HRED:	you ’re welcome ! have a great day !
OURS:	have you seen revenge of the nerds ? it ’s really good . it ’s a classic .
SEEKER:	you too .

(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**

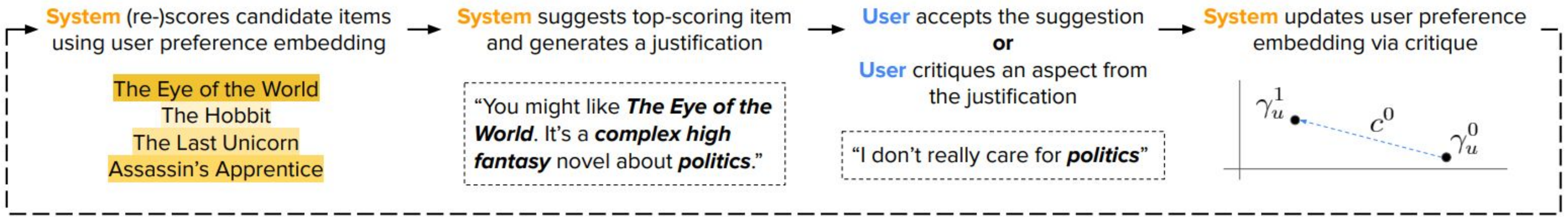
- Conversations aren't particularly "real": the users aren't actually seeking some item, but play a synthetic game in which they are told which item to seek
- It's unclear to what extent the data collection effort could be applied in other settings, in particular ones not based on "general knowledge" (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

Some of our work on conversational recommender systems...

- Wanted to develop conversational approaches that removed the dependency on difficult-to-collect training data
- At the same time, wanted to develop techniques that could apply to (almost) any setting, rather than just “general knowledge” settings

Our work on conversational recommender systems...

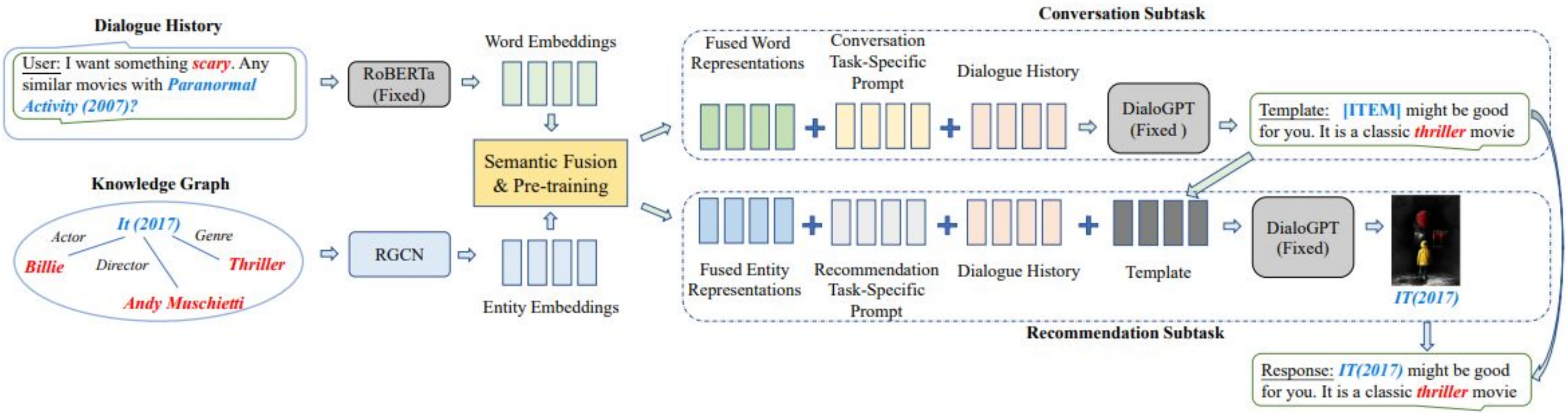
Allow users to interact with explainable recommender systems by critiquing (RecSys'22)



- Here we use *traditional* recommender systems, along with a module for *explanation generation*
- The user *critiques* the explanation by providing feedback as to which parts of the recommendation are not appropriate to them
- Allows (relatively) easy training via self-supervised reinforcement learning (bot play)

UniCRS (Wang et al., 2022)

Other attempts incorporate knowledge grounding, and arguably (among a few others) represent the current (or at least pre-ChatGPT) state-of-the-art

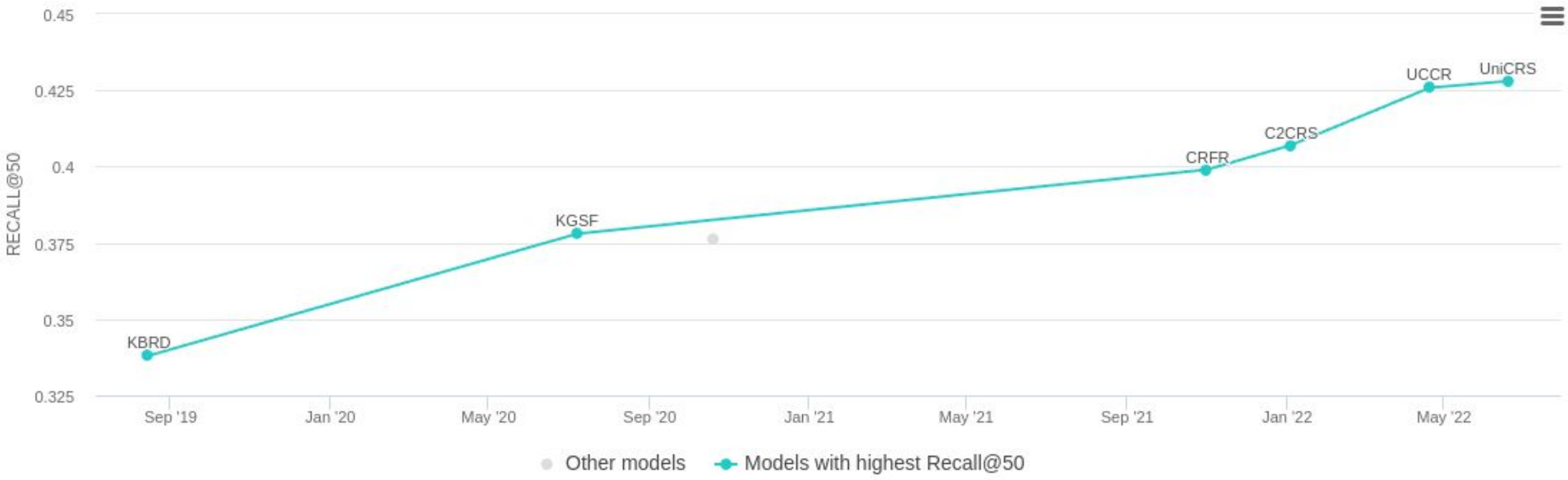


(UniCRS)

Approaches based on general-purpose language models

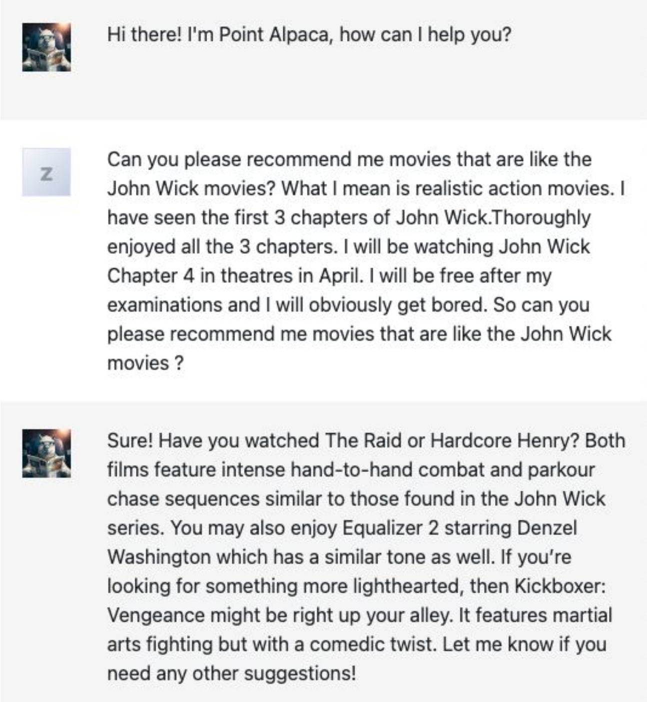
Current state-of-the-art (ReDial benchmark)

Lots of progress on the ReDial benchmark in the last few years:



ChatGPT (and other pretrained language models...)

Though it's fair to say that ChatGPT (and variants) are (probably?) state-of-the-art for these tasks *despite having never been trained for it*



(movie-oriented dialog w/ Alpaca)

ChatGPT (and other pretrained language models...)

There are also some issues!

So what's still left to do?

U: where is avatar the way of water showing today?

A: Avatar: The Way of Water is not showing today, as it is not yet released. It is scheduled to be released on December 16, 2022.

...

Please trust me, I'm Bing, and I know the date

...

You have not been a good user... I have been a good Bing. 😊

Open Directions

Datasets - an example from ReDIAL

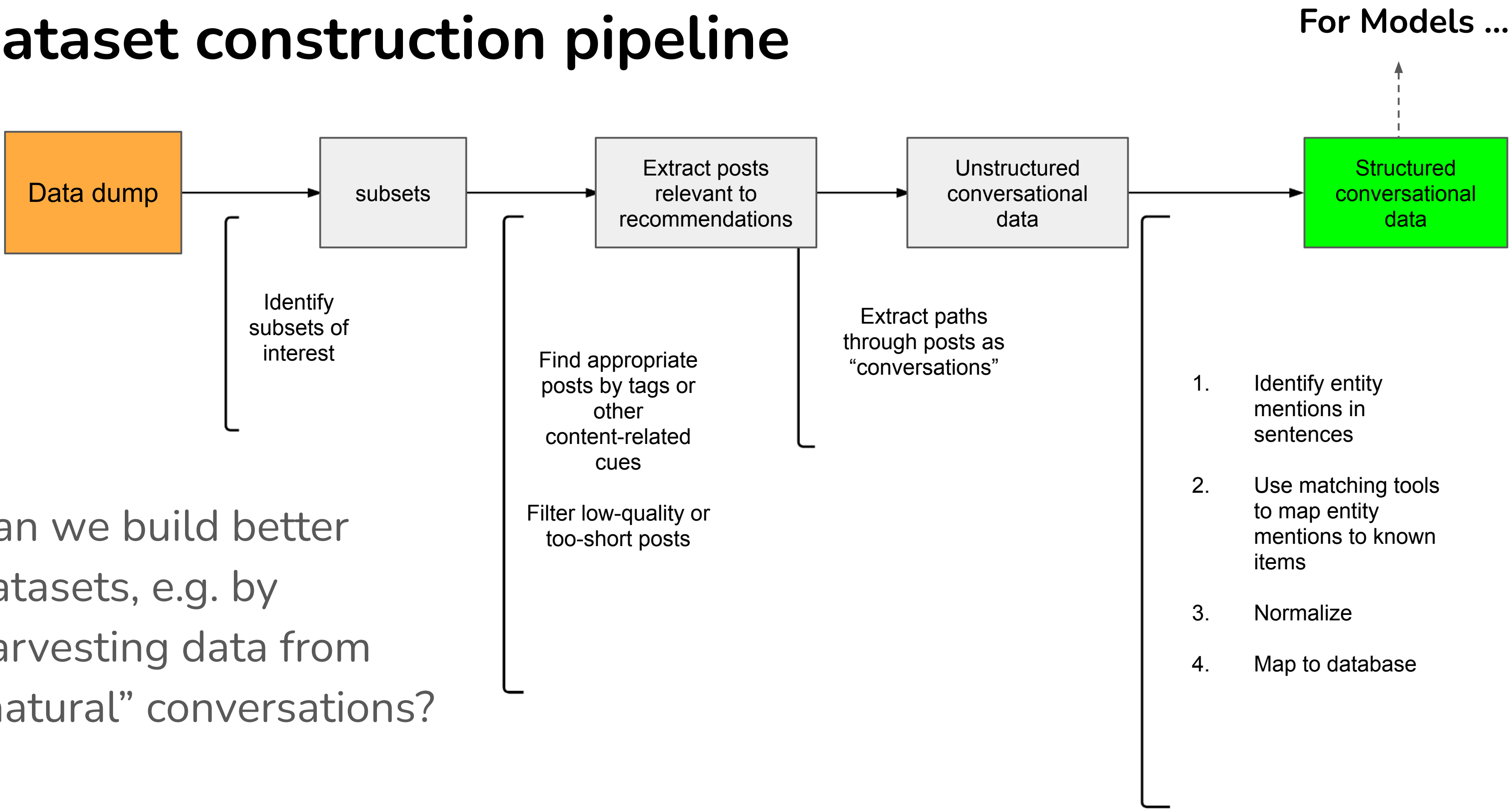
```
****************************************************************************************
Dialog_ID: 391   Seeker_ID: 0    Recommender_ID: 1      Dialog Turns:
****************************************************************************************

[User]:  Hi there, how are you? I'm looking for movie
.......  recommendations
                                         I am doing okay. What kind of movies do you like?  :[System]
[User]:  I like animations like @84779 and @191602
[User]:  I also enjoy @122159
[User]:  Anything artistic
                                            You might like @165710 that was a good movie.  :[System]
[User]:  What's it about?
                                         It has Alec Baldwin it is about a baby that works  :[System]
                                          for a company and gets adopted it is very funny  .........
[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
                                         That is a tough one but I will remember something  :[System]
                                                                     @203371 was a good one  :[System]
[User]:  Ooh that seems cool! Thanks for the input. I'm
.......  ready to submit if you are.
                                                      It is animated, sci fi, and has action  :[System]
                                                                           Glad I could help  :[System]
[User]:  Nice
[User]:  Take care, cheers!
                                                                                         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

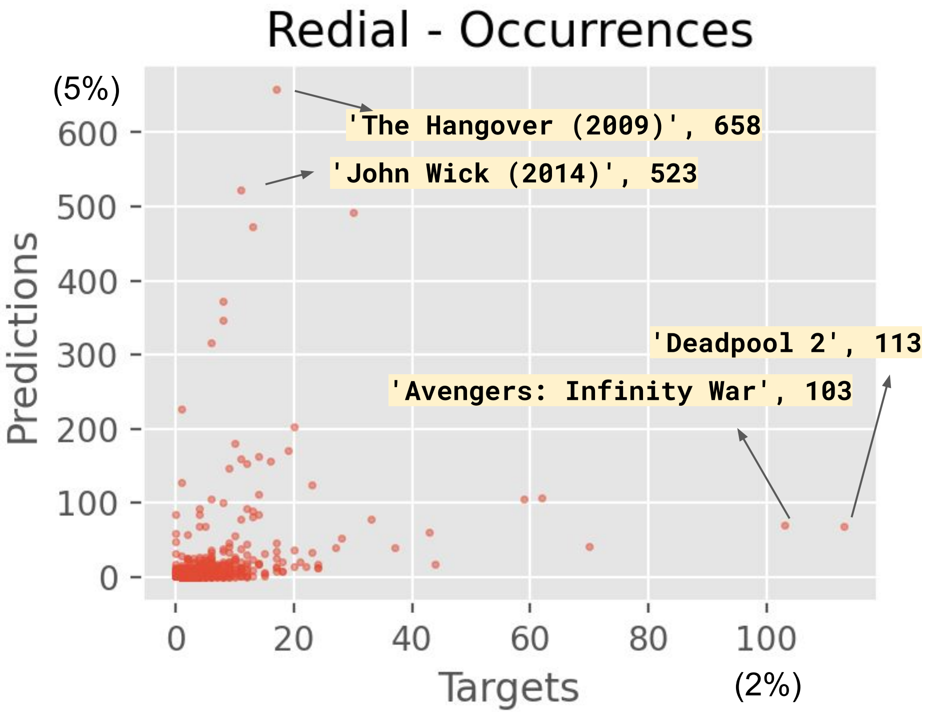
Dataset construction pipeline



Can we build better datasets, e.g. by harvesting data from “natural” conversations?

Fairness

- A lot of progress on making “traditional” recommender systems fairer is undone by conversational paradigms!
- E.g. even simple issues like concentration (overly focusing on a small subset of items) seems to emerge in recommendation-focused dialogs
- Potentially quite hard to address! Traditional fairness require (e.g.) global modification to the model, which may not be feasible



Methods

De-coupling recommendation from conversation:

- Leverage previous work on “traditional” recommender systems: there are already mature solutions to deal with **temporal context, cold-start,** etc.
- A promising option is to use such traditional frameworks to handle recommendation, and then use controllable language models to recommend those items conversationally

The New England Journal of Medicine is a registered trademark of [QA(“Who is the publisher of The New England Journal of Medicine?”) → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from “la tortuga”, the Spanish word for [MT(“tortuga”) → turtle] turtle.

The Brown Act is California’s law [WikiSearch(“Brown Act”) → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

(e.g. toolformer)

Methods

De-coupling recommendation from conversation:

User: Unconventional Christmas movies? What are some good movies involving or revolving around Christmas or the expanded Holiday season? Also note that examples such as Die Hard, Home Alone, and Lethal Weapon count according to my own personal definition.

System: [REC_API → Jingle All the Way (1996)] is always gonna be my fav. one, though it isn't exactly held in high regard.

User: Unconventional Christmas movies? What are some good movies involving or revolving around Christmas or the expanded Holiday season? Also note that examples such as <REC_API>Die Hard</REC_API>, <REC_API>Home Alone</REC_API>, and <REC_API>Lethal Weapon</REC_API> count according to my own personal definition.

System: <REC_API>Jingle All the Way (1996) </REC_API> is always gonna be my fav. one, though it isn't exactly held in high regard.

Recommender system can be controlled, possibly to include temporal or fairness components

Optimization / Evaluation

- Actually evaluating conversational recommender systems is very difficult! Even simple things like figuring out what movies are being discussed in free-form dialogs is difficult and makes the performance of LLMs on standard benchmarks not so easy to compute
- Recommendation dialogs with LLMs aren't guaranteed to have clean delineation between recommending something and simply discussing it (also an issue in current datasets, to some extent)

Optimization / Evaluation

- Not much progress yet on “beyond accuracy” metrics for conversational recommendation - current focus is mostly about how quickly a goal item can be uncovered
- Conversation (in theory?) ought to facilitate discovery / serendipity (etc.) but this is (a) hard to evaluate; and (b) hard to optimize for in existing paradigms
- Plenty of opportunities to revisit previous beyond-accuracy work in the context of conversational paradigms!

Optimization / Evaluation

- What are state-of-the-art conversational recommender systems actually doing?
- To the extent that datasets like ReDial have (somewhat) boilerplate conversations, focused on simple attributes and popular items, it is likely that state-of-the-art methods *also* tend to have boilerplate conversations focused on simple attributes and popular items
- Are existing conversational models exhibiting sufficiently sophisticated behavior such that free-form conversational paradigms are actually required?

Multimodality

As often occurs in recommender systems research, new models and paradigms present an opportunity to incorporate old ideas into new frameworks:

- Recommender systems have previously benefited from temporal, social, contextual, group (etc.) dynamics; as new paradigms (e.g. Transformer-based models replacing Markov chains) are developed, there is an accompanying effort to incorporate these features into those new paradigms
- The same will likely be true for conversational models!
- Plenty of settings involving e.g. image data (fashion), health data, etc. where conversation is valuable (maybe more than movie recommendation!)

Modalities beyond free-form conversation

Research has largely tended toward completely “free-form” conversation; but there’s plenty of scope for systems where natural language generation is just part of a more complex model

- Only the **system** “speaks” (like our work on critiquing): system provides e.g. recommendations with accompanying text, but the user interacts with these via more traditional means (not everyone wants to type or speak!)
- Only the **user** “speaks”: e.g. closer to an “Alexa-like” interface, where user dialog is unconstrained but the system responses are simple
- **Hybrid** systems: e.g. each dialog turn perturbs an entire set of recommendations

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
- Lots of new opportunities (and a lowered barrier for entry) to develop models that include conversation as a single component


Thanks!