Guideline Bias in Wizard-of-Oz Dialogues

Benchmarking: Past, Present and Future (ACL-IJCNLP 2021)

Victor Petrén Bach Hansen & Anders Søgaard

Topdanmark

UNIVERSITY OF COPENHAGEN

Annotator guidelines for dialogue

Task-oriented dialogue datasets are usually collected using the Wizard-of-Oz method

Crowd-sourcers play the role of the user/system to simulate a conversation between a human and automated agent.

Both user and agent will follow a set of guidelines to achieve their respective goals.

Annotator guidelines for dialogue

Annotation guidelines are intended to train the annotators.

Focus of this work is on the unintended bias that arises from how the guidelines are formulated.

This is what we refer to as *guideline bias*.

Guideline bias examples

We examine bias that stem from annotator guidelines in two WoZ datasets:

1. Coached Conversational Preference Elicitation (CCPE-M)
   a. How can a lexical bias in the guideline influence annotators?
   b. What is the downstream effect?

2. Taskmaster-1
   a. How does the order of guideline questions influence the conversation?

EX1: Guideline Bias in CCPE-M

First example is motivated by a lexical bias, namely the overwhelming use of the verb “like”:

- More than 50% of sentences contain inflection of “*like*”.
- 72% of agents use “*like*” in their first dialogue turn.
- 52% of users respond with a form of “*like*”.

This can contribute to optimistic estimates of performance.

General Instructions The goal of this type of dialogue is for you to get the users to explain their movie preferences: The KIND of movies they like and dislike and WHY. We really want to end up finding out WHY they like what they like movie AND why the DON’T like what they don’t like. We want them to take lots of turns to explain these things to you.

Important We want users to discuss likes and dislikes for kinds of movies rather than just about specific movies. (But we trigger these more general preferences based on remembering certain titles.) You may bring up particular movie titles in order to get them thinking about why they like or dislike that kind of thing. Do not bring up particular directors, actors, or genres. For each session do the following steps:

1. Start with a normal introduction: Hello. I’d like to discuss your movie preferences.
2. Ask them what kind of movies they like and why they generally like that kind of movie.
3. Ask them for a particular movie name they liked.
4. Ask them what about that KIND of movie they liked. (get a couple of reasons at least – let them go on if they choose)
5. Ask them to name a particular movie they did not like.
6. Ask them what about that movie they did not like. (get a couple of reasons at least or let them go on if they choose)
7. Now choose a movies using the movie generator link below. Ask them if they liked that movie (if they haven’t seen it: (a) ask if they have heard of it. If so, ask if they would see it (b) then choose another that they have seen to ask about). Once you find a movie from the list they have seen, ask them why they liked or disliked that kind of movie (get a couple of reasons).
8. Finally, end the conversation gracefully

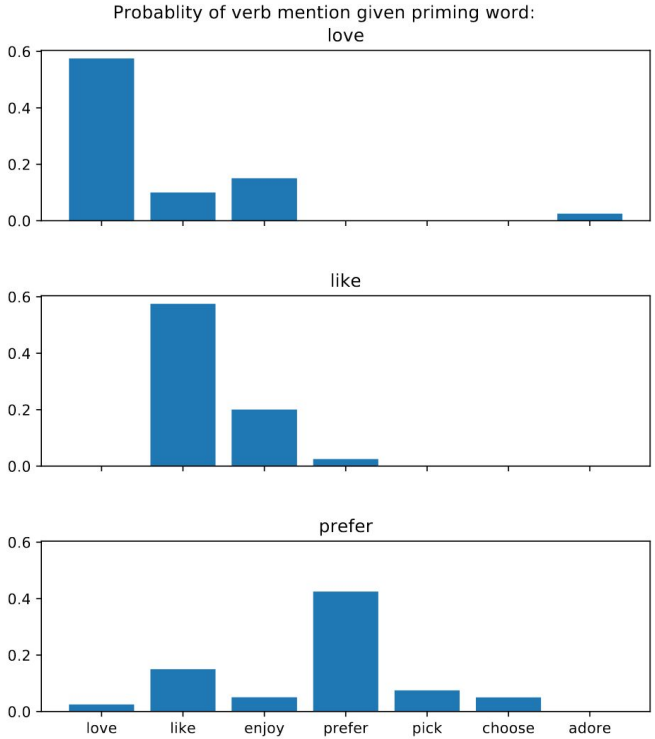
EX1: Guideline Bias in CCPE-M

To study the priming effect, we do a small crowdsourced experiment.

We ask different turkers what movies they *like/love/prefer.*

Observe verb mentions in the responses given.

Turkers overwhelmingly respond with the same verb the question was asked with.



EX1: Guideline Bias in CCPE-M

We frame CCPE-M movie preference detection problem as a sentence classification task.

ASST	Can you tell me what types of movies you enjoy watching?
USER	Mostly action adventure drama fantasy, documentary
ASST	And what about those kinds of movies do you like?
USER	Usually I like movies like that that consist with those genres.
USER	I just like the storytelling, how deep some movies can get into, the characters, and pretty much like the creativity.
ASST	Alright, whats one of your favorite movies?
USER	Mr. Nobody
ASST	And what was it about that movie that made it enjoyable?
USER	Because it dealt with basically all of life, making choices, and that each choice has a different consequence.
USER	So, think before you do.
ASST	I see.

Colors indicate preference towards MOVIES, GENRES or NONE

EX1: Guideline Bias in CCPE-M

We examine the downstream effects by replacing instances of (adverbial) "*like*" in the test set with synonymous words or phrases.

Original:

- I [like] Terminator 2

Perturbed:

- I [love] Terminator 2
- I [was incredibly affected by] Terminator 2
- I [have as my all time favorite movie] Terminator 2
- I [am out of this world passionate about] Terminator 2

EX1: Guideline Bias in CCPE-M

In addition, we collect a small out-of-domain test set.

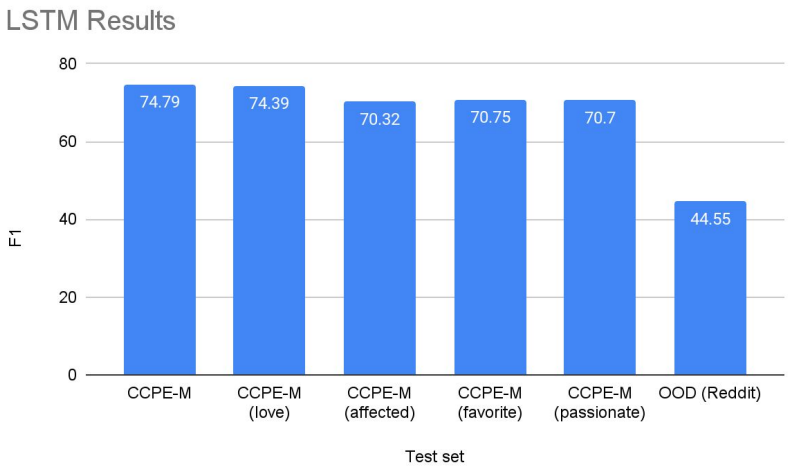
We scrape comments from Reddit threads that discuss movie preferences.

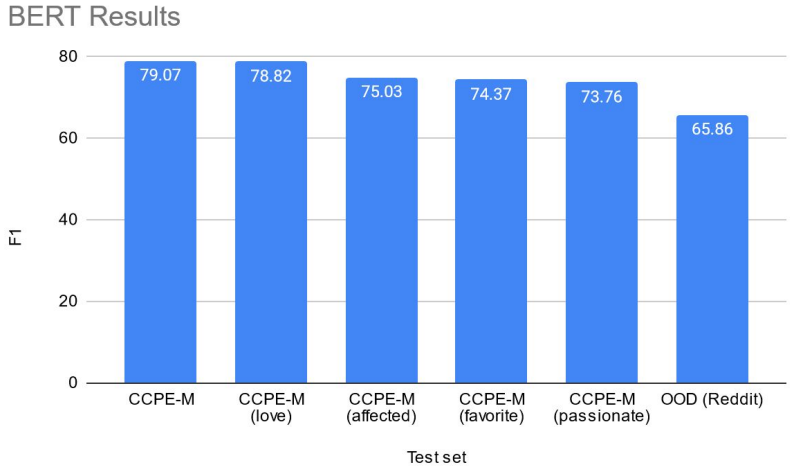
In general, the occurrence of "*like*" is much less frequent.

We annotate the data according to the instructions by the CCPE-M authors.

EX1: Guideline Bias in CCPE-M

Results on perturbed and OOD data





EX1: Guideline Bias in CCPE-M

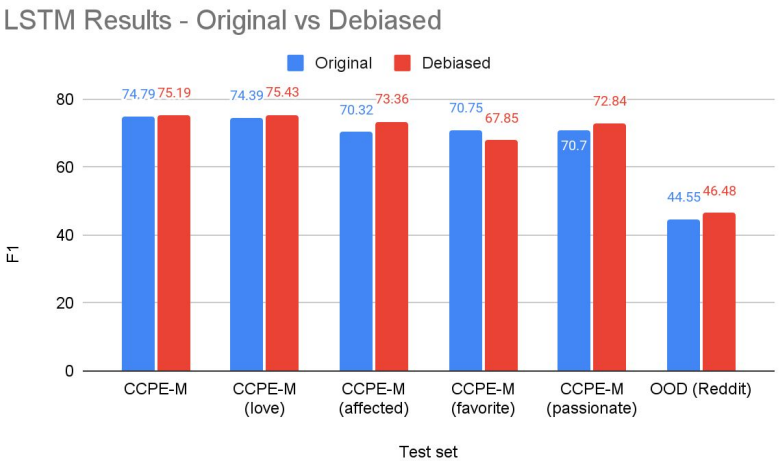
We also experiment with debiasing the training data by augmenting instances of "*like*".

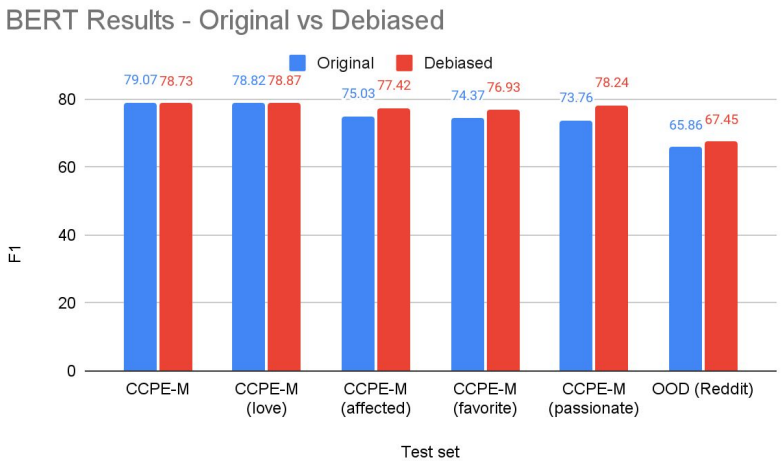
Go through train data and replace 20% of forms of *"like"* with one of the following synonymous words/phrases extracted from a thesaurus:

1. *derive pleasure from*
2. *get a kick out of*
3. *appreciate*
4. *take an interest in*
5. *cherish*
6. *find appealing*

EX1: Guideline Bias in CCPE-M

Results on perturbed and OOD data with debiased train data





EX2: Bias in Taskmaster-1

Order of goals of conversations in guidelines can also bias the conversation.

We examine this phenomena in the Taskmaster-1 dataset.

The goal of the conversation (movie ticket booking) is to lock in a series of requirements, ie:

- Movie title
- Theatre name
- City

EX2: Bias in Taskmaster-1

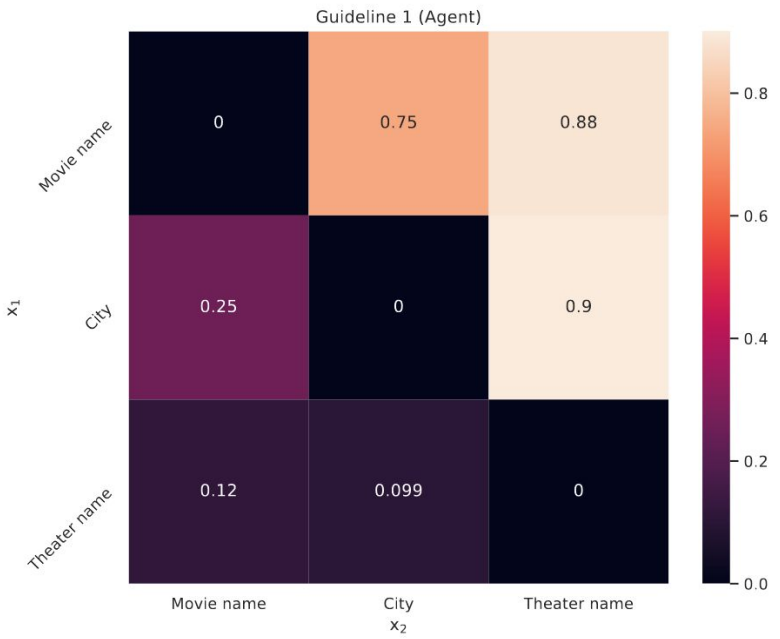
We quantify the guideline bias by computing the probability that a conversation goal is mentioned before another, given the guideline also does so.

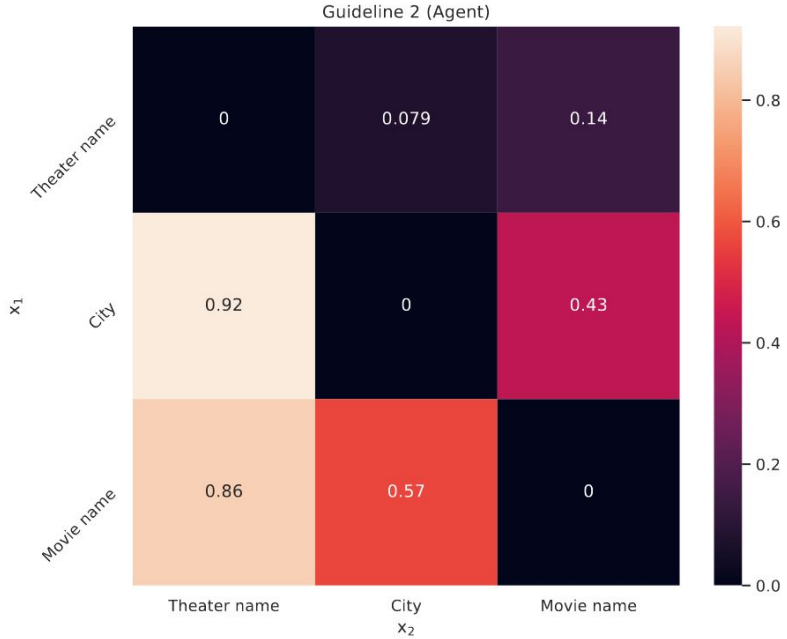
E.g. if *move title* is mentioned before *theatre name* in the guideline, does the same hold for the downstream conversation?

Conversations are collected using two different guidelines, which means we can compare the two.

EX2: Bias in Taskmaster-1

Probability heatmaps





In Conclusion

- We examined *guideline bias* in two newly presented WoZ style dialogue corpora
- We show how a lexical bias for the word “like” in the guidelines can:
  - Lead to a bias for this word in the dialogues
  - Influence model performance when there is an absence of this word
- Can be somewhat mitigated with a data augmentation strategy
- Analyze how the order of guideline goals affect the structure of the conversation