Guideline Bias in Wizard-of-Oz Dialogues

Victor Petrén Bach Hansen,[1 2] **Anders Søgaard**[1]
[1]Department of Computer Science, University of Copenhagen, Denmark
[2]Topdanmark A/S, Denmark
{victor.petren,soegaard}@di.ku.dk

Abstract

NLP models struggle with generalization due to sampling and annotator bias. This paper focuses on a different kind of bias that has received very little attention: *guideline bias*, i.e., the bias introduced by how our annotator guidelines are formulated. We examine two recently introduced dialogue datasets, CCPE-M and Taskmaster-1, both collected by trained assistants in a Wizard-of-Oz set-up. For CCPE-M, we show how a simple lexical bias for the word *like* in the guidelines biases the data collection. This bias, in effect, leads to poor performance on data without this bias: a preference elicitation architecture based on BERT suffers a 5.3% absolute drop in performance, when *like* is replaced with a synonymous phrase, and a 13.2% drop in performance when evaluated on out-of-sample data. For Taskmaster-1, we show how the order in which instructions are presented, biases the data collection.

1 Introduction

Sample bias is a well-known problem in NLP – discussed from Marcus (1982) to Barrett et al. (2019) – and annotator bias has been discussed as far back as Ratnaparkhi (1996). This paper focuses on a different kind of bias that has received very little attention: *guideline bias*, i.e., the bias introduced by how our annotator guidelines are formulated.

Annotation guidelines are used to train annotators, and guidelines are therefore in some sense intended to and designed to prime annotators. What we will refer to in our discussion of guideline bias, is rather the unintended biases that result from how guidelines are formulated, and the examples used in those guidelines. If a treebank annotation guideline focuses overly on parasitic gap constructions, for example, inter-annotator agreement may be higher on those, and annotators may be biased to annotate similar phenomena by analogy with parasitic gaps.

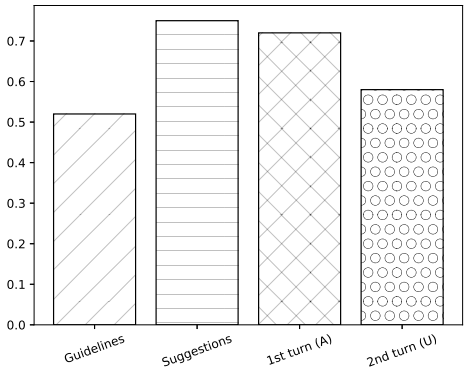

Figure 1: The percentage of sentences with the word *like* in the CCPE-M annotation guidelines (Guidelines), the suggested questions to ask users, in the guidelines (Suggestions), (c) the *actual* first turns by the assistants (1st turn), and (d) the actual replies by the users (2nd turn). In all cases, more than half of the sentences contain the word *like*.

We focus on two recently introduced datasets, the Coached Conversational Preference Elicitation corpus (CCPE-M) from Radlinski et al. (2019), related to the task of conversational recommendation (Christakopoulou et al., 2016; Li et al., 2018), and Taskmaster-1 (Byrne et al., 2019), which is a multi-purpose, multi-domain dialogue dataset. CCPE-M consists of conversations about movie preferences, and the part of Taskmaster-1, we focus on here, conversations about theatre ticket reservations. Both corpora were collected by having a team of assistants interact with users in a Wizard-of-Oz (WoZ) set-up, i.e. a human plays the role of a digital assistant which engages a user in a conversation about their movie preferences. The assistants were given a set of guidelines in advance, as part of their training, and it is these guidelines that induce biases. In CCPE-M, it is the overwhelming use of the verb *like* (see Figure 5) and its trickle-down effects, we focus on; in Taskmaster-1, the order of

Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, pages 8–14
August 5–6, 2021.

the instructions. In fact, the CCPE-M guidelines consist of 324 words, of which 20 (**6%**) are inflections or derivations of the lemma *like*: As shown in Figure 5 in the Appendix, more than **50%** of the sentences in the guidelines include forms of *like*! This very strong bias in the guidelines has a clear downstream effect on the assistants that are collecting the data. In their first dialogue turn, the assistants use the word *like* in **72%** of the dialogues. This again biases the users responding to the assistants in the WoZ set-up: In **58%** of their first turns, given that the assistant uses a form of the word *like*, they also use the verb *like*. We show that this bias leads to overly optimistic estimates of performance. Additionally, we also demonstrate how the guideline affects the user responses through a controlled priming experiment. For Taskmaster-1, we show a similar effect of the guidelines on the collected dialogues.

Contributions We introduce the notion of *guideline bias* and present a detailed analysis of guideline bias in two recently introduced dialogue corpora (CCPE-M and Taskmaster-1). Our main experiments focus on CCPE-M: We show how a simple bias toward the verb *like* easily leads us to overestimate performance in the wild by showing performance drops on semantically innocent perturbations of the test data, as well as on a new sample of movie preference elicitations that we collected from Reddit for the purpose of this paper. We also show that debiasing the data, improves performance. The CCPE-M provides a very clear example of *guideline bias*, but other examples can be found, e.g., in Taskmaster-1, which we discuss in §3. We discuss more examples in §4.

2 Bias in CCPE-M

We first examine the CCPE-M dataset of spoken dialogues about movie preferences. The dialogues in CCPE-M are generated in a Wizard-of-Oz set-up, where the assistants type their input, which is then translated into speech using text-to-speech technologies, at which point users respond by speech. The dialogues were transcribed and annotated by the authors of Radlinski et al. (2019).

Sentence classification We frame the CCPE-M movie preference detection problem as a sentence-level classification task. If a sentence contains a labeled span, we let this label percolate to the sentence level and be a label of the entire sentence. If a sentence contains multiple unique label spans the sentence is assigned the leftmost label. A sentence-level label should therefore be interpreted as saying *in this sentence, the user elicits a movie or genre preference*. Our resulting sentence classification dataset contains five different preference labels, including a *NONE* label. We shuffle the data at the dialogue-level and divide the dialogues into training/development/test splits using a 80/10/10 ratio, ensuring sentences from the same dialogue will not end up in both training and test data. As the assistants utterances rarely express any preferences, we only include the user utterances to balance the number of negative labels. See Table 2 for statistics regarding the label distribution.

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

Figure 2: Example of test sentence permutations.

Perturbations of test data In order to analyse the effects of guideline bias in the CCPE-M dataset, we introduce perturbations of the instances in the test set where *like* occurs, replacing *like* with a synonymous word, e.g. *love*, or paraphrase, e.g. *holds dearly*. We experiment with four different replacements for *like*: (i) *love*, (ii) *was incredibly affected by*, (iii) *have as my all time favorite movie* and (iv) *am out of this world passionate about*. See Figure 2 for an example sentence and its perturbed variants. The perturbations occasionally, but rarely, lead to grammatically incorrect input.[1] We emphasize that even though we increase the length of the sentence, the phrases we replace *like* with should signal an even stronger statement of preference, which models should be able to pick up on. Since our data consists of informal speech it includes adverbial uses of *like*; we only replace verb occurrences, relying on SpaCy's POS tagger.[2] We replace 219 instances of the verb *like* throughout the test set.

Perturbations of train data We also augment the training data to create a less biased resource.

[1] Our models are generally robust to such variation, and, as we will see in our experiments below, the perturbations are less harmful than collecting a new sample of evaluation data and evaluating your model on this sample.

[2] https://spacy.io/

Testing on (↓)/Training on (→)	CCPE-M		CCPE-M$_{thesaurus}$	
	BiLSTM	BERT	BiLSTM	BERT
CCPE-M	74.79	**79.07**	75.16	78.73
CCPE-M$_{love}$	74.39	78.82	75.43	**78.87**
CCPE-M$_{was\ incredibly\ affected\ by}$	70.32	75.03	73.36	**77.42**
CCPE-M$_{have\ as\ my\ all\ time\ favorite\ movie}$	70.75	74.37	67.85	**76.93**
CCPE-M$_{am\ out\ of\ this\ world\ passionate\ about}$	70.70	73.76	72.84	**78.24**
Reddit	44.55	65.86	46.48	**67.45**

Table 1: Comparison of in-sample F_1 performance, performance on the same data with *like* replaced with phrases with similar meaning, and performance on Reddit data. Results are reported for training models on biased CCPE-M as well as a debiased CCPE-M$_{thesaurus}$ which improves model performance in almost all cases.

Label	train	dev	test	Reddit
NONE	4508	535	545	60
MOVIE_OR_SERIES	2736	346	313	119
MOVIE_GENRE_OR_CATEGORY	1274	169	166	20
PERSON	66	6	9	11
SOMETHING_ELSE	21	0	0	1
total	8605	1056	1033	211

Table 2: CCPE-M and Reddit sentence-level statistics

Here we adopt a slightly different strategy, also to evaluate a model trained on the debiased training data to the above perturbed test data: We use six paraphrases of the verb *like* listed in a publicly available thesaurus,[3] none of which overlap with the words used to perturb the test data, and randomly replace verbal *like* with a probability of 20%. The paraphrases are sampled from a uniform distribution. A total of 401 instances are replaced in the training data using this approach. This is not intended as a solution to guideline bias, but in our experiments below, we show that a model trained on this simple, debiased dataset generalizes better to out of sample data, showing that the bias toward *like* was in fact one of the reasons that our baseline classifier performed poorly in this domain.

Reddit movie preference dataset In addition to the perturbed CCPE-M dataset, we also collect and annotate a challenge dataset from Reddit threads discussing movies for the purpose of preference elicitation. The comments are scraped from Reddit threads with titles such as '*Here's A Simple Question. What's Your Favorite Movie Genre And Why?*' or '*What's a movie that you love that everyone else hates?*' and mostly consist of top-level comments. These top-level comments typically respond directly the question posed by the thread, and explicitly state preferences. We also include some random samples from discussion trees that contain no preferences, to balance the label distribution slightly. In this data, we observe the word *like*, but less frequently: The verb *like* occurred in 15/211 examples. The data is annotated at the sentence level, as described previously, and we follow the methodology described by Radlinski et al. (2019) and identify anchor items such as names of movies or series, genres or categories and then label each sentence according to the preference statements describing said item, if any. The dataset contains roughly 100 comments, that when divided into individual sentences resulting in 211 datapoints. The statistics can be found in the final column of Table 2. We make the data publicly available.[4]

Results We evaluate the performance on two different models on the original and perturbed CCPE-M, as well as on our Reddit data: (i) a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) sentence classifier, trained only on CCPE-M, including the embeddings, and (ii) a fine-tuned BERT sentence classification model (Devlin et al., 2018). For (i), we use two BiLSTM layers ($d = 128$), randomly initialized embeddings ($d = 64$), and a dropout rate of 0.5. The model is trained for 45 epochs. For (ii), we use the base, uncased BERT model with the default parameters and finetune for 3 epochs. Model selection is conducted based on performance on the development set. Performance is measured using class-weighted F_1 score. We report results in Table 1 on the various perturbation test sets as well as the Reddit data, when (i) the models are trained on the unchanged CCPE-M data, and (ii) the models are trained on the debiased version CCPE-M$_{thesaurus}$.

[3] http://thesaurus.com. The paraphrases consists of: (1) *derive pleasure from*, (2) *get a kick out of*, (3) *appreciate*, (4) *take an interest in*, (5) *cherish*, (6) *find appealing*.

[4] https://github.com/vpetren/guideline_bias

On the original dataset, BERT performs slightly better than the BiLSTM architecture, but the differences are relatively small. Both BiLSTM and BERT suffer a drop in performance, when examples are perturbed and the word *like* is replaced with synonymous words or phrases. Note how longer substitutions result in a larger drop in performance, e.g. *love* vs. *am out of this world passionate about*. We see the drops follow the same pattern for both architectures, while BiLSTM seems a bit more sensitive to our test permutations. Both models do even worse on our newly collected Reddit data. Here, we clearly see the sensitivity of the BiLSTM architecture, which suffers a 30% absolute drop in F_1; but even BERT suffers a bit performance drop of more than 13%, when evaluated on a new sample of data. When training on CCPE-M$_{thesaurus}$, both models become more invariant to our perturbations,with up to 4.5 F_1 improvements for BERT model and 3 F_1 improvements for the BiLSTM, without any loss of performance on the original test set. We also observe improvements on our collected Reddit data, suggesting that *the initial drop in performance can be partially explained by guideline bias and not only domain differences*.

Controlled priming experiment To establish the priming effect of guidelines in a more controlled setting, we set up a small crowdsourced experiment. We asked turkers to respond to a hypothetical question about movie preferences. For example, turkers were asked to imagine they are in a situation in which they 'are asked what movies' they 'like', and that they like a specific movie, say *Harry Potter*. The turker may then respond: *I've always liked Harry Potter*. We collected 40 user responses for each of the priming verbs *like*, *love* and *prefer*, 120 total, and for each of the verbs used to prime the turkers, we compute a probability distribution over most of the verbs in the response vocabulary that are likely to be used to describe a general preference towards something. Figure 3 shows the results of the crowdsourced priming experiments. We can observe that when a specific priming word, such as *like*, is used, there is a significantly higher probability that the response from the user will contain that same word, illustrating that when keywords in guidelines are heavily overrepresented, the collected data will also reflect this bias.

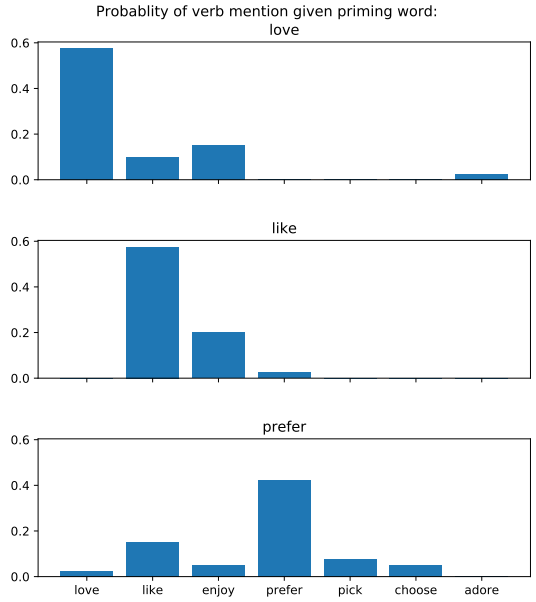

Figure 3: Probability that a verb that describes a preference towards a movie is mentioned, given a priming word by the annotator is mentioned.

3 Bias in Taskmaster-1

The order in which the goals of the conversation is described to annotators in the guidelines can also bias the order in which these goals are pursued in conversation. Taskmaster-1 contains conversations between a user and an agent where the user seeks to accomplish a *goal* by, e.g., booking tickets to a movie, which is the domain we focus on. When booking tickets to go see a movie, we can specify the movie title before the theatre, or vice versa, but models may not become robust to such variation if exposed to very biased examples.

Unlike CCPE-M, the Taskmaster-1 dataset was (wisely) collected using two different sets of guidelines to reduce bias, and we can therefore investigate the downstream effects of of the bias induced by the two sets of guidelines. To quantify the guideline bias, we compute the probability that a goal x_1 is mentioned before another one x_2 in an dialogue, given that x_1 precedes x_2 in the guidelines. We only consider dialogues where all goals are mentioned at least once, i.e., $\sim$ 900 in total; the conversations are then divided into two, based on the guideline that was used. Figure 4 shows the heat map of these relative probabilities. The guidelines have a clear influence on the final structure of the conversation, i.e. if the movie title (x_1) is mentioned before the city (x_2) in the guideline, there is

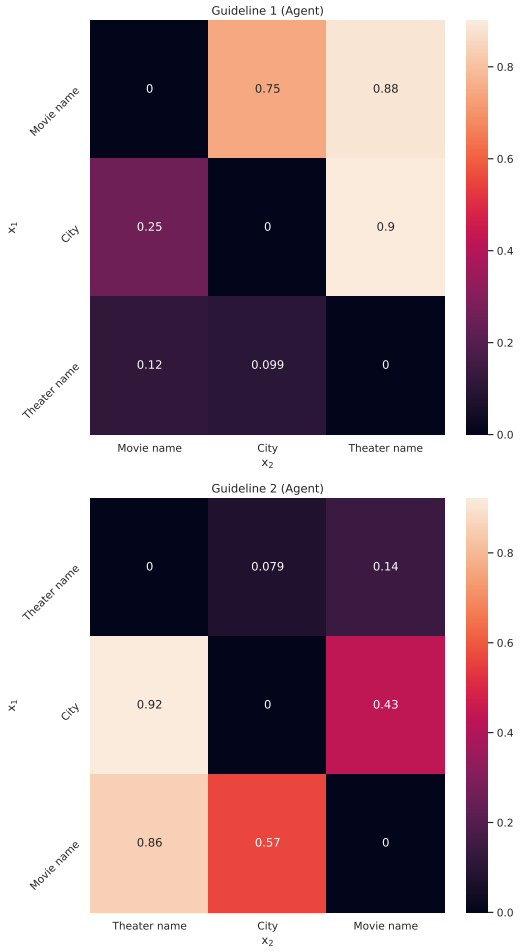

Figure 4: Probability that a guideline goal x_1 is mentioned before another one x_2 in an actual dialogue, given that x_1 comes before x_2 in the agent's guideline.

a high probability (0.75) that the same is true in the dialogues. If they are not, the probability is much lower (0.57).

4 Related Work

Plank et al. (2014) present an approach to correcting for adjudicator biases. Bender and Friedman (2018) raise the possibility of (demographic) bias in annotation guidelines, but do not provide a means for detecting such biases or show any existing datasets to be biased in this way. Amidei et al. (2018) also discuss the possibility, but in a footnote. Geva et al. (2019) investigates how crowdsourcing practices can introduce annotator biases in NLU datasets and therefore result in models overestimating confidence on samples from annotators that have contributed to both the training and test sets. Liu et al. (2018), on the other hand, discuss a case in which annotation guidelines are biased by being developed for a particular domain and not easily applicable to another. Cohn and Specia (2013) explores how models can learn from annotator bias in a somewhat opposite scenario from ours, e.g. when annotators deviate from annotation guidelines and inject their own bias into the data, and by using multi-task learning to train annotator specific models, they improve performance by leveraging annotation (dis)agreements. There are, to the best of our knowledge, relatively few examples of researchers identifying concrete guideline-related bias in benchmark datasets: Dickinson (2003) suggest that POS annotation in the English Penn Treebank is biased by the vagueness of the annotation guidelines in some respects. Friedrich et al. (2015) report a similar guideline-induced bias in the ACE datasets. Dandapat et al. (2009) discuss an interesting bias in a Bangla/Hindi POS-annotated corpus arising from a decision in the annotation guidelines to include two labels for when annotators were uncertain, but not specifying in detail how these labels were to be used. Goldberg and Elhadad (2010) define structural bias for dependency parsing and how it can be attributed to bias in individual datasets, among other factors, originating from their annotation schemes. Ibanez and Ohtani (2014) report a similar case, where ambiguity in how special categories were defined, led to bias in a corpus of Spanish learner errors.

5 Discussion & Conclusion

In this work, we examined *guideline bias* in two newly presented WoZ style dialogue corpora: We showed how a lexical bias for the word *like* in the annotation guidelines of CCPE-M, through a controlled priming experiment leads to a bias for this word in the dialogues, and that models trained on this corpus are sensitive to the absence of this verb. We provided a new test dataset for this task, collected from Reddit, and show how a debiased model performs better on this dataset, suggesting the 13% drop is in part the result of guideline bias. We showed a similar bias in Taskmaster-1.

Acknowledgements

This work was funded by the Innovation Fund Denmark and Topdanmark.

References

Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In *COLING*.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *EMNLP*.

Emily Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. In *TACL*.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. 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 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 815–824, New York, NY, USA. ACM.

Trevor Cohn and Lucia Specia. 2013. Modelling annotator bias with multi-task Gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32–42, Sofia, Bulgaria. Association for Computational Linguistics.

Sandipan Dandapat, Priyanka Biswas, Monojit Choudhury, and Kalika Bali. 2009. Complex linguistic annotation – no easy way out! a case from bangla and hindi pos labeling tasks. In *LAW*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Markus Dickinson. 2003. Detecting errors in part-of-speech annotation. In *EACL*.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. 2015. Annotating genericity: a survey, a scheme, and a corpus. In *LAW*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Yoav Goldberg and Michael Elhadad. 2010. Inspecting the structural biases of dependency parsing algorithms. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 234–242, Uppsala, Sweden. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Maria Del Pilar Valverde Ibanez and Akira Ohtani. 2014. Annotating article errors in spanish learner texts: design and evaluation of an annotation scheme. In *PACLIC*.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9725–9735. Curran Associates, Inc.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah Smith. 2018. Parsing tweets into universal dependencies. In *NAACL*.

Mitch Marcus. 1982. Building non-normative systems – the search for robustness. In *ACL*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *EACL*.

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

Adwait Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *EMNLP*.

A Appendices

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

Figure 5: CCPE-M Guidelines to Assistants