Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences

Filip Radlinski, Krisztian Balog, Bill Byrne and Karthik Krishnamoorthi
Google, Inc.
{filiprad, krisztianb, billb, krishnamoorthi}@google.com

Abstract

Conversational recommendation has recently attracted significant attention. As systems must understand users' preferences, training them has called for conversational corpora, typically derived from task-oriented conversations. We observe that such corpora often do not reflect how people naturally describe preferences.

We present a new approach to obtaining user preferences in dialogue: Coached Conversational Preference Elicitation. It allows collection of natural yet structured conversational preferences. Studying the dialogues in one domain, we present a brief quantitative analysis of how people describe movie preferences at scale. Demonstrating the methodology, we release the CCPE-M dataset to the community with over 500 movie preference dialogues expressing over 10,000 preferences.[1]

1 Introduction

Conversational information seeking has repeatedly been identified as a research direction of particular importance (Allan et al., 2012; Culpepper et al., 2018). From a practical perspective, it is a common task for personal digital assistants in many recommendation domains including movies, restaurants, and travel. However, today's systems are often limited in what they understand. We observe that in many cases, the actions allowed and the utterances understood reflect available metadata, such as movie genres or restaurant food categories, which may mirror uncertain assumptions of how users would choose to characterize their needs in an unconstrained setting. This can lead to conversational systems with unnatural or tedious dialog design.

Developing systems supporting natural interactions requires understanding how users would choose to express preferences to an idealized assistant. It has been noted that a lack of suitable conversational datasets limits such research (Joho et al., 2018). Thus we ask *what properties matter most to users? How do real people describe their preferences when encouraged to do so naturally in a conversational setting?*

We present a new robust approach for eliciting preferences, producing natural language that a conversational recommender should interpret, represent internally, and use in determining items to recommend. The semantic structure observed also provides new insights into how results could be described to users, to mirror their terminology.

We use a Wizard-of-Oz approach (WoZ): A human agent plays a digital assistant, and users are played by crowd-sourced workers. The human agent is given instructions specifically designed to elicit preferences, while keeping the conversation natural. We particularly focus on avoiding biases in prior approaches, yielding new insights into natural language processing challenges. Crucially, we argue that the focus should be *preference elicitation*, rather than standard *task completion*.

Although our approach is domain independent, we validate on movie preference elicitation, as it has received most past attention (Ricci et al., 2015). In particular, movies have high-quality metadata available (actors, directors, production dates, etc.), which is often used. We are able to ask which of these properties are actually normally mentioned by people, finding significant differences: Canonical attributes such as genre, leading actors and directors, paint an incomplete picture. Real users more often refer to less tangible

[1] Available at https://g.co/dataset/ccpe

and highly subjective aspects, like the plot style or attributes like violence. We argue that conversational recommender systems should take this into account when representing knowledge.

Our key contributions are three-fold. First, we present a new method for obtaining realistic conversational recommendation dialogues, addressing previous challenges in *quantitative* analysis of recommendation needs. Second, we release a dialog corpus that allows natural language understanding systems to assess how well they interpret user utterances in a conversational context, and promote their more closely mirroring natural dialogue. Finally, we present a brief analysis of user preferences in the movie domain.

2 Related work

2.1 Dialog Systems

Dialog systems are generally classified as goal-driven or non-goal-driven (Chen et al., 2017). The latter, commonly *chatbots*, mimic human responses in open domain dialogues, often powered by neural networks trained end-to-end on large corpora (Sordoni et al., 2015; Serban et al., 2016). Goal-driven (a.k.a. task-oriented) systems aim to assist users with specific tasks (e.g., select products). The architecture typically consists of natural language understanding, state tracking, dialogue policy, and language generation (Chen et al., 2018), each often implemented and optimized individually (Young et al., 2013). There is a growing interest in end-to-end trainable task-oriented systems (Bordes and Weston, 2016), yet most are restricted to narrow domains (Serban et al., 2018).

Commercial systems, like Google Assistant and Apple's SIRI, combine chat and task focus, supporting a hybrid of multi-domain task-oriented and open-domain chat. Yet user interaction is often relatively unnatural (Luger and Sellen, 2016). Combining task-based and chat modes of operation attracts active research (Akasaki and Kaji, 2017; Yan et al., 2017).

2.2 Conversational Recommendation

We focus on *conversational recommendation*, combining elements of chat, goal-oriented dialog, and question answering (Dodge et al., 2016; Li et al., 2018). Within the movie domain, a large body of prior work on models, test collections, and evaluation methodology exists (Ricci et al., 2015). Early work includes human-human movie recommendation, such as (Johansson, 2004), who focused on characterizing dialogue structure.

Dodge et al. (2016) develop a synthetic dataset with the purpose of training end-to-end neural dialog systems. Their Movie dataset combines question answering, recommendation, and general dialog. It is generated using a fixed set of simple templates, and mining a Reddit online forum.

Closest to our work is the REDIAL dataset (Li et al., 2018), containing human-to-human conversations about movies. Similar to our work, the dialogues are conducted on a crowdsourcing platform, where one participant is seeking recommendations which the other party provides. However, their main focus is on algorithmic aspects, and the conversations are driven by the explicit goal of making recommendations. As such, workers are required to mention at least four specific movies in each conversation. Our interest is more broadly targeted to understand how people naturally express preferences in a conversational setting.

Other relevant conversational recommendation work includes Sun and Zhang (2018), who capture long term user preferences in a deep reinforcement learning framework by asking the user for information about particular facets.

2.3 Data Collection Approaches

Conversational recommendation system training data can be obtained in many ways. Serban et al. (2018) provide a comprehensive overview, here we summarize the most relevant past approaches.

Implicit observations use logs from an existing system, e.g., for travel booking (Bennett and Rudnicky, 2002). It may be that the system is operated by humans (Hemphill et al., 1990). Such analysis is necessarily biased by current system policy, which drives user (re)actions. Past failures also influence logs, as they can create frustration (Kiseleva et al., 2016) after which users may avoid similar interactions.

Explicit preference observations are most commonly based on web review mining (Zhang et al., 2018) or mining online forums (Li et al., 2010; Dodge et al., 2016). Both suffer from population biases. More importantly, neither type of corpus necessarily represents what preferences would be expressed in a direct interaction with an intelligent assistant, nor how they would be stated.

Unstructured user studies produce more rigid yet smaller datasets. Participants express a need, which they refine through unstructured dialog. The objective is usually to characterize interaction behavior (Johansson, 2004; Trippas et al., 2017) and to understand users' attitude and expectations towards an automatic agent (Vtyurina et al., 2017).

Task-based user studies commonly create collections using WoZ methodology (Li et al., 2018). A participant engages in conversation for some task (e.g. schedule a bus ride). A wizard acts as intermediary to an existing non-conversational system. This frees dialog state tracking and conversation understanding from current practical limitations. Yet the conversations intend to solve tasks that discourage natural information flow (Serban et al., 2018). Moreover, the Wizard interacts with an existing system, often strongly basing them by the *existing interface* and its terminology.

3 Coached Wizard-of-Oz User Studies

As we have seen, most dialogues backed by real systems are biased by that existing system. These systems, in turn, are often biased by the *metadata available* rather than *natural* user preferences. For instance, if a Wizard is presented with an existing categorization of possible answers, it is normal for them to ask the user to select among these.

Meanwhile, we aim to understand desirable qualities of *future* conversational search and recommendation systems and desire to understand natural user preferences. We ask which properties users express preferences about, and also in what way. Our methodology is thus closer to coaching the user, through questions that avoid suggesting particular terminology or answers. Rather, open-ended questions are used to obtain preferences, requesting *examples*, and questioning *what aspect* of the expressed preferences or examples the system should pay attention to. By using a WoZ approach, with human operators simulating the system (who we refer to as *Wizards*), we similarly allow for human-level natural language understanding. This renders linguistically rich utterances. We also design for "users" (who we refer to as *Requesters*) to have an experience as consistent as possible to interacting with a fully automated digital assistant.[2]

To make this concrete, we introduce our validation setting: Movie preference elicitation. In each conversation, the Wizard was instructed to elicit the Requester's preferences following a general script, while keeping the exchanges as natural as possible. While the full instructions are presented in the Appendix, at a high level these are to:

1. Ask *what sort of movies* the Requester likes.
2. Ask for an *example* of a liked movie.
3. Ask *what in particular* was appealing.
4. Ask for an example of a disliked movie.
5. Ask what in particular was not appealing.
6. Select example movies, and for each:
   (a) Ask if the user has heard of / seen it.
   (b) If so, ask for similar preferences.

Importantly, the flow is permitted to evolve naturally and may be adapted to the Requester.

Compared to existing corpora, the dialogues collected are not slot-filling, nor do they resemble "20 questions" with repetitive yes/no questions. They also differ from past unstructured dialogues, having clear preference structure. This makes our CCPE method unique in providing rich yet tractable conversational exchanges.

4 Methodology

The Wizard was provided the written dialog flow template, and given occasional feedback on their conversations. Unique to our setup among WoZ systems, the Wizard typed their input, which was played to the Requester using text-to-speech consistent with that used by a commercial digital assistant. Thus, from the perspective of the Requester, the system resembled today's speech-based digital assistants as closely as possible, aiming to preserve the distinctive nature of spoken dialogue (Chafe and Tannen, 1987).

The Requesters were paid crowd workers on a crowdsourcing platform, invited to talk about their movie preferences. There we informed that an assistant would guide them with questions. They spoke using a microphone, with the audio played directly to the Wizard.

To collect the corpus, each Wizard had a succession of conversations, matched to a sequence of Requesters. After each conversation, the Requester's audio was transcribed by a separate crowd worker, then combined with the known typed text of the Wizard. An example partial dialog is provided in the Appendix.

Elements that are not relevant to preference understanding were removed from the transcribed

[2] While Requesters were not told that they are conversing with a Wizard-of-Oz system, it is possible they suspected it.

conversations. These include pleasantries, confirmation of the Requester's task, resolution of technical issues or task interruptions. On the other hand, the transcribed speech was kept as uttered, including filler words, disfluencies and discourse markers. Conversations that ended prematurely were kept (where of non-trivial length). While relatively rare, conversations where the Requester only gave single-word answers were removed as they only provided minimal insight into natural recommendation dialog. Finally, all utterances were annotated, as described below.

4.1 Methodological Notes

We briefly discuss three common challenges seen. (1) Audio failures occurred at times, where one of the Wizard and Requester could not correctly hear the other. Other times, there was also out-of-context background communication. (2) Some Requesters had poor engagement, with very short answers. While the Wizard attempted to elicit richer answers, this did not always succeed. We hypothesize that some crowd workers acted lazily, although perhaps some also did not have particular preferences to express. (3) Undesirable prompting by the Wizard saw some Requesters prompted for specific properties. Other times, the Wizard interjected their own preferences. While this biased the Requester, it is also natural and sometimes led to richer exchanges. We therefore allowed it, but attempted to filter it in our analysis by associating each named item or attribute with the first speaker who mentioned it. We are thus able to differentiate prompted and unprompted terminology.

4.2 Semantic Annotation

Our key contribution is a methodology for preference elicitation. To better allow characterizing how users naturally express preferences in the example movie setting, we also annotated the dialogues by identifying preference statements.

As developing robust annotation guidelines that yield consistent labels is known to be complex, annotation was performed by the authors of this paper.[3] In particular, we sub-sampled 510 of the dialogues collected to annotate. These have a median of 22 turns and median duration of 3 minutes and 36 seconds. During annotation, 8 conversations were identified as of too poor quality, yielding a final set of 502 conversations. The conversations consist of 11,972 utterances and were annotated with 15,646 annotations.

[3]Most conversations were transcribed by a single author, with an equal number completed by each author. A fraction were annotated by two different authors to measure inter-judge agreement, reported below.

4.3 Annotation Ontology

In the corpus, we first annotate **Anchor items**: names of movies or series, genres or categories, people, and other entities. These provide the anchor points for preferences, i.e., what is being talked about.

Preferences by a Requester or Wizard were also annotated. These were partitioned by what the preference was about (matching the anchor items), and the information conveyed in three categories:
Preference statements about an anchor item indicate that the person does or does not like the relevant item, or some aspect of it. It most closely matches the popular meaning of a *preference*.
Descriptions of an anchor item consist of neutral information about an anchor item. Bringing attention to specific parts of a movie (for instance), they tell us what this person finds as key characteristics.
Other statements about an anchor item convey relevant information but do not provide an explicit sentiment, such as "I haven't seen that." While not telling us if the user likes or dislikes the movie, these convey relevant information for a recommendation system.

In summary, the annotations identify statements that a conversational recommender should be able to interpret. See Appendix for an example.

5 Annotation Analysis

At least one movie was named in 99.6% of conversations, and at least one movie genre or category was named in 95%. A person was named in just 33% of conversations. Other statements, usually about whether the Requester had seen a movie, were present in 66% of conversations. We identified on average 12.5 preferences about specific movies, and 5.5 genre preferences in each, as well as 0.3 preferences about a person. Neutral descriptions of movies were found in 40% of conversations. In total, 6,297 movie preferences were found, along with 2,775 genre preferences, 2,545 movie names and 1,714 genre or category names.

5.1 Inter-Judge Agreement

A random subset of 80 conversations (15%) were independently annotated by two annotators. As

our ontology is on two dimensions, and spans between labels can overlap, Krippendorff's α_U does not apply (Artstein and Poesio, 2008). Due to space constraints, we report agreement uncorrected for chance agreement. In the 4,094 annotations, 58% matched exactly and 17% had one annotator select a substring of that selected by the other, with the same type. We thus find 75% inter-judge agreement. A further 6% of annotations consisted of the same text being annotated with different labels, most often due to disagreement between neutral description and preference labels.

5.2 User-Generated Anchor Items

In one step, the Requesters were asked to name specific likes and dislikes. They did not find it difficult: Only 4% of did not provide any movies, while 70% named at least two. Analyzing the movies named, we find a heavy tailed distribution: 643 distinct movies were named (1.3 distinct movies each). No movie was mentioned by more than 18 distinct Requesters, and all but 18 movies or series were mentioned 6 or fewer times. That is, Requesters often gave examples of less well-known movies, characterizing their uniqueness.

We find a similar heavy-tailed pattern among mentions of other named entities, such as people (actors, directors) and genres. However, people (actors or directors) are only mentioned in 33% of conversations. On the other hand, users often refer to fine-grained movie sub-genres.

5.3 Conversational Preference Relationships

The dialog collection also illustrates how preferences build upon each other. E.g., consider:

ASST	Have you seen the movie *Arrival*?
USER	Yes.
ASST	Did you like that movie?
USER	Yes, I did.
ASST	What did you enjoy about it?
USER	I liked the narrative, I liked that it didn't pull punches and didn't have unnecessary action scenes. I thought [...]

To interpret each utterance, the full context needs to be taken in account. This also provides an opportunity to use the CCPE-M dataset to study contextual natural language understanding.

5.4 Non-rating preferences

In the above, we also see the user provide information that is not a rating of a movie. Rather, we first learn that the user has *seen* a given movie. In other conversations, we observe that a user has *not heard* of some classic movie, or has seen *all* the movies in some series. Such statements, known to be informative (Steck, 2010; Marlin et al., 2007), were seen in 66% of conversations.

5.5 Details Present in Preferences

We saw that when Requesters were asked an open-ended question about the type of movie that they like or dislike, they most often first characterized themselves by movie genre. These genres were sometimes expanded with details such as example movies, yet it is interesting to note that people were much more rarely mentioned here.

5.6 Disfluences

We note that many spoken preferences are naturally disfluent. This requires flexible approaches to semantic interpretation. For example *I really like the action and all that like the like I really like like the action in that movie was pretty great.*

5.7 Final Observations

We find that in the movie domain, when users express preferences naturally, these are very rich. The items suggested *by users* follow a heavy-tailed distribution. The natural language observed is often both complex and disfluent, and requires the full conversational context to interpret. Preferences refer to rich properties, with emphasis on the story, plot, characters and acting.

6 Conclusion

This paper presented a new methodology for obtaining natural conversational preferences. By asking questions in a "coaching" format, where the assistant avoids prompting the user with specific terminology, the collected data allows a quantitative analysis of the structure of preferences. This analysis can then inform the design of conversational recommendation systems, providing a basis for realistic natural language understanding and natural language generation challenges.

This work opens a number of avenues. It identifies challenges in natural language understanding of realistic preference statements, and provides a datasets for addressing them. Assuming that the output of a system should reflect users' language, the methodology and data also provide guidance for development of future conversational systems. Finally, our method could be used to obtain similar datasets in other domains.

References

Satoshi Akasaki and Nobuhiro Kaji. 2017. Chat detection in an intelligent assistant: Combining task-oriented and non-task-oriented spoken dialogue systems. In *Proc. of ACL'17*, pages 1308–1319.

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. *SIGIR Forum*, 46(1):2–32.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Christina L. Bennett and Alexander I. Rudnicky. 2002. The Carnegie Mellon Communicator corpus. In *Proc. of INTERSPEECH'02*.

Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR*, abs/1605.07683.

Wallace Chafe and Deborah Tannen. 1987. The relation between written and spoken language. *Annual Review of Anthropology*, 16:383–407.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 19(2):25–35.

Yun-Nung Chen, Asli Çelikyilmaz, and Dilek Z. Hakkani-Tür. 2018. Deep learning for dialogue systems. In *COLING (Tutorials)*, pages 25–31.

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). *SIGIR Forum*, 52(1):34–90.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H. Miller, Arthur Szlam, and Jason Weston. 2016. Evaluating prerequisite qualities for learning end-to-end dialog systems. In *Proc. of ICLR'16*.

Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Proc. of HLT'90 workshop*, pages 96–101.

Pontus Johansson. 2004. *Design and Development of Recommender Dialogue Systems*. Ph.D. thesis, Linköping University.

Hideo Joho, Lawrence Cavedon, Jaime Arguello, Milad Shokouhi, and Filip Radlinski. 2018. Cair'17: First international workshop on conversational approaches to information retrieval at sigir 2017. *SIGIR Forum*, 51(3):114–121.

Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Understanding user satisfaction with intelligent assistants. In *Proc. of CHIIR'16*, pages 121–130.

Qing Li, Jia Wang, Yuanzhu Peter Chen, and Zhangxi Lin. 2010. User comments for news recommendation in forum-based social media. *Inf. Sci.*, 180(24):4929–4939.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Proc. of NeurIPS'18*, pages 9748–9758.

Ewa Luger and Abigail Sellen. 2016. "like having a really bad pa": The gulf between user expectation and experience of conversational agents. In *Proc. of CHI'16*, pages 5286–5297.

Benjamin Marlin, Richard Zemel, Sam Roweis, and Malcolm Slaney. 2007. Collaborative filtering and the missing at random assumption. In *Proc. UAI'07*.

Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. *Recommender Systems Handbook*, 2nd edition.

Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proc. of AAAI'16*, pages 3776–3783.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems: The journal version. *Dialogue Discourse*, 9(1):1–49.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proc. of NAACL-HLT'15*, pages 196–205.

Harald Steck. 2010. Training and testing of recommender systems on data missing not at random. In *Proc. KDD'10*, pages 713–722.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *Proc. of SIGIR'18*, pages 235–244.

Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How do people interact in conversational speech-only search tasks: A preliminary analysis. In *Proc. of CHIIR'17*, pages 325–328.

Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proc. of CHI EA'17*, pages 2187–2193.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proc. of AAAI'17*, pages 4618–4626.

Steve J. Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. of the IEEE*, 101(5):1160–1179.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards conversational search and recommendation: System ask, user respond. In *Proc. of CIKM'18*, pages 177–186.

Appendix

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.
ASST	OK. Now, tell me about a movie you didnt like.
USER	I guess Spaceballs, because it was kind of just annoying in some parts. I mean, it was all right, but most of the time, you know, it just It needed to have a little bit different kind of comedy in it. I didn't like the some of the characters in it, so
ASST	Generally speaking, what did you not like about that particular type of movie?
USER	The storytelling wasn't really on pace.
USER	So the plot could have been better.

Figure 1: Partial example dialog between a Wizard and an elicitation participant (Requester).

General Instructions
The goal of this type of dialog 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 2: Written instructions provided to Wizards, also provided feedback on practice conversations.

ASST	What kind of movies do you like, and why do you like this type of movie?
USER	I like science fiction movies. I like science fiction movies because they always have nteresting stories, and they deal with crazy new technologies or futuristic technologies.
	`Name_of.Genre_or_Category`: science fiction
	`Preference_about.Genre_or_Category` : I like science fiction movies
	`Preference_about.Genre_or_Category` : interesting stories
	`Preference_about.Genre_or_Category` : deal with crazy new technologies or futuristic technologies
ASST	Is there a specific movie that you heavily dislike?
USER	The Exorcist
	`Name_of.Movie_or_Series` : The Exorcist
	`Preference_about.Movie_or_Series` : The Exorcist
ASST	What do you dislike about this movie?
USER	I don't like how anxious it makes me.
	`Preference_about.Movie_or_Series` : I don't like how anxious it makes me

Figure 3: Example semantic annotations on two segments of conversations