INSPIRED: Toward Sociable Recommendation Dialog Systems

Shirley Anugrah Hayati[♡] **Dongyeop Kang**[☆] **Qingxiaoyang Zhu**[♡]
Weiyan Shi[♡] **Zhou Yu**[♡]
[♡]Department of Computer Science, University of California, Davis
[☆]University of California, Berkeley
{sahayati, qinzhu, wyshi, joyu}@ucdavis.edu
dongyeopk@berkeley.edu

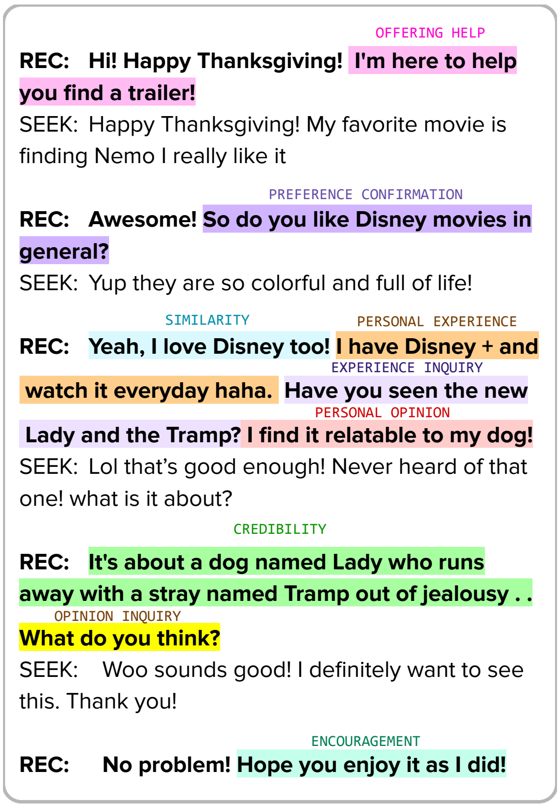
Abstract

In recommendation dialogs, humans commonly disclose their preference and make recommendations in a friendly manner. However, this is a challenge in developing a sociable recommendation dialog system, due to the lack of dialog dataset annotated with such sociable strategies. Therefore, we present INSPIRED, a new dataset of 1,001 human-human dialogs for movie recommendation with measures for successful recommendations. To better understand how humans make recommendations in communication, we design an annotation scheme related to recommendation strategies based on social science theories and annotate these dialogs. Our analysis shows that sociable recommendation strategies, such as sharing personal opinions or communicating with encouragement, more frequently lead to successful recommendations. Based on our dataset, we train end-to-end recommendation dialog systems with and without our strategy labels. In both automatic and human evaluation, our model with strategy incorporation outperforms the baseline model. This work is a first step for building sociable recommendation dialog systems with a basis of social science theories[1].

1 Introduction

Sociable conversational agents build rapport with users, in order to gain trust and favor from them. Social science researchers believe that the rapport influence a more persuasive recommendation to successfully suggest an item that satisfies user needs (Yoo et al., 2012; Gkika and Lekakos; Pecune et al., 2019; Gretzel and Fesenmaier, 2006).

However, existing works on recommendation dialog systems lack a study about communication strategies used by human speakers for making successful and persuasive recommendations. They collect the dataset in scenario-based settings or convert product review datasets into question-answering conversations (Reschke et al., 2013; Yan et al., 2017; Sun and Zhang, 2018; Kang et al., 2019; Li et al., 2018). Common issues with these types of datasets are: (1) homologous utterances, (2) mostly question-answering pairs, and (3) lack of user engagement.



OFFERING HELP
REC: Hi! Happy Thanksgiving! I'm here to help you find a trailer!
SEEK: Happy Thanksgiving! My favorite movie is finding Nemo I really like it

PREFERENCE CONFIRMATION
REC: Awesome! So do you like Disney movies in general?
SEEK: Yup they are so colorful and full of life!

SIMILARITY / PERSONAL EXPERIENCE / EXPERIENCE INQUIRY / PERSONAL OPINION
REC: Yeah, I love Disney too! I have Disney + and watch it everyday haha. Have you seen the new Lady and the Tramp? I find it relatable to my dog!
SEEK: Lol that's good enough! Never heard of that one! what is it about?

CREDIBILITY / OPINION INQUIRY
REC: It's about a dog named Lady who runs away with a stray named Tramp out of jealousy . . What do you think?
SEEK: Woo sounds good! I definitely want to see this. Thank you!

ENCOURAGEMENT
REC: No problem! Hope you enjoy it as I did!

Figure 1: An example snippet of human-human recommendation dialog in INSPIRED. REC refers a person who recommends a movie and SEEK refers a person who looks for a recommendation. Above each recommender's utterance is the recommendation strategy annotated by human workers. Best seen in colors.

In this work, we aim to validate whether sociable recommendation strategies are effective for making a successful recommendation in a dialog. To do so,

[1] Dataset and code are available at https://github.com/sweetpeach/Inspired

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 8142–8152,
November 16–20, 2020. ©2020 Association for Computational Linguistics

Dataset	INSPIRED	CONVREC (Sun and Zhang, 2018)	GORECDIAL (Kang et al., 2019)	REDIAL (Li et al., 2018)
Naturalness	✓	✗	✗	✓
Sociable Strategies	✓	✗	✗	✗
Movie Information	✓	✗	✓	✗
Conversation Types	Mixed	QA	Mixed	Mixed
#Dialogs	1,001	385	9,125	10,006
#Utterances	35,811	-	160,904	182,150

Table 1: Comparison of related recommendation dialog datasets. "QA" refers to question-answer pairs. "Mixed" indicates that the conversations contain both statements and question-answer pairs. CONVREC collected 385 human-curated dialogs, but only released 875,721 simulated dialogs.

we propose INSPIRED, a recommendation dialog dataset of two-paired crowd-workers in a natural setting, with additional annotations for sociable recommendation strategies. The dataset consists of 1,001 dialogs, and each utterance is manually annotated with the sociable strategies based on social science theory. To encourage more natural dialog flow, we do not set any restrictions on the number of movies or the type of movies to recommend. Figure 1 shows an example of annotated dialog. More examples are in Table 11 and 12 in the Appendix.

Our analyses show that sociable recommendation strategies are correlated with successful recommendation in dialogs. These insights motivate us to build a more sociable recommendation dialog system to achieve better persuasion outcomes.

For extrinsic evaluation, we build two end-to-end dialog systems trained on the INSPIRED dataset: one is encoded with recommendation strategies and the other is not. We find that the model encoded with our strategy annotations performs better in both automatic and human evaluation.

We believe that enriching the intersection between social science and computational linguistics in INSPIRED opens plenty of rooms for future studies on sociable recommendation dialog.

2 Related Work

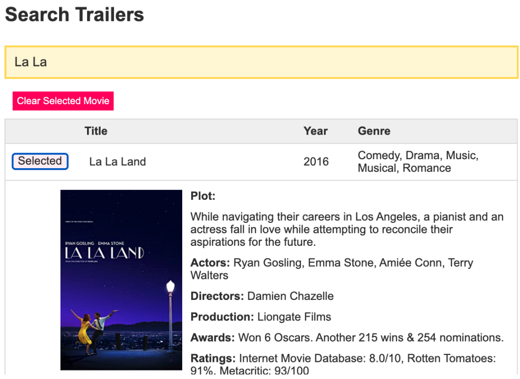
Social science theories on recommendation. Psychological researchers believe that interactions with recommendation systems should not only be seen from a technical perspective but should also be examined from a social and emotional perspective (Zanker et al., 2006). Yoo et al. (2012) propose that credibility, likeability, friendliness, humor, and other language styles are significant factors for persuasive recommendations. Pecune et al. (2019) has studied modeling social explanation for movie recommendation, such as personal opinion and personal experience. Häubl and Murray (2003) find that more information on recommendation may help consumers make better purchase decisions, but leave them overwhelmed with the abundant information. Inspired by these theories, we borrow such principles in the design of our sociable recommendation strategies.

Conversational recommendation systems. While studies on conversational recommendation systems have been done, none of them focus on the sociable recommendation strategies for persuasive outcome. This is is due to the lack of existing datasets for studying effective strategies in recommendation dialog. Table 1 compares different factors across the recommendation dialog datasets including INSPIRED.

Prior works on recommendation dialogs collect data based on template-based question-answering pairs from user reviews (Thompson et al., 2004; Reschke et al., 2013; Sun and Zhang, 2018; Zhang et al., 2018b). These datasets contain structured utterances where the recommender continuously asks for the seeker's product preference.

Kang et al. (2019) collected goal-driven recommendation dialogs (GORECDIAL) in a gamified setting where both speakers are given a small set of movies with descriptions to find the best recommendation. This role-play game setting may not effectively reflect the real-world situation since the seeker pretends that they like the given movies.

The most similar work to ours is Li et al. (2018)'s REDIAL dataset which consists of chit-chats for movie recommendation. However, the recommendations are conditioned on the movies mentioned in the dialog, and not directly on the language usage. Also, they tend to mention only movie names rather than an in-depth discussion on the movie



	Title	Year	Genre
Selected	La La Land	2016	Comedy, Drama, Music, Musical, Romance

Plot:

While navigating their careers in Los Angeles, a pianist and an actress fall in love while attempting to reconcile their aspirations for the future.

Actors: Ryan Gosling, Emma Stone, Amiée Conn, Terry Walters

Directors: Damien Chazelle

Production: Liongate Films

Awards: Won 6 Oscars. Another 215 wins & 254 nominations.

Ratings: Internet Movie Database: 8.0/10, Rotten Tomatoes: 91%. Metacritic: 93/100

Figure 2: Movie search interface for recommenders.

preference.

Our work is also closely related to Radlinski et al. (2019) on movie preference elicitation and Fabian Galetzka1 (2020) on movie discussion in the dialog setting. Preference elicitation is an important step for the human recommender to comprehend seeker's taste before recommendation, but these datasets are not recommendation conversations.

Meanwhile, dialogs in INSPIRED have both stages: preference elicitation and recommendation. INSPIRED also captures sociable recommendation strategies in conversations and measures recommendation with ratings.

Sociability in dialog systems. In human-human conversations, people engage in a talk that does not only contain task-oriented topics (Bickmore and Cassell, 2005). Thus, sociability has raised more attention in dialog systems as they become more sociable, engaging, and user-adaptive (Zhang et al., 2018a; Shi and Yu, 2018; Göker and Thompson, 2000).

Zhang et al. (2018a) proposed a chit-chat dataset and presented the task of more personalized dialogs system conditioned on user profile information. Sociability leads to a more persuasive conversation (Yoo et al., 2012), so social skills are essential for dialog systems to make successful recommendations.

Communication strategies on specific tasks, such as donation and product price negotiation, have been found useful for task completion (Wang et al., 2019; Zhou et al., 2019). In this work, we connect different sociable strategies with recommendation in dialog settings and show that sociable strategies have a positive impact on recommendation success.

3 Recommendation Dialog Collection

3.1 Movie Database Creation

To ensure that the recommended movie has trailers and metadata information, we curate a database with all movie trailers from Movieclips Trailer[2] released between 2008 and 2020, and movies from MovieLens dataset (Harper and Konstan, 2015). In total, we have 17,869 movies with trailers and metadata information. We design a simple movie search interface (Figure 2) to assist recommenders in searching for a movie.

3.2 Recommendation Task

We recruit crowd-workers from Amazon Mechanical Turk. In each conversation, two workers are randomly paired and assigned different roles: one as a recommender and another as a seeker. Our collection set-up is more realistic compared to prior works as (1) recommenders have no limitations of the number of movies to recommend, (2) seekers accept or reject a movie following their true preference, and (3) we record if seekers actually watch the video trailer or not.

Recommender. Recommenders' task is to recommend a movie successfully to the seeker. Before chatting, we show them tips for sociable recommendation strategies with example utterances. Then they chat with the seekers in two phases: user information gathering and movie recommendation. In the user information gathering phase, recommenders are asked to understand the seekers' movie tastes. In the recommendation phase, the recommenders can still request seekers' preference while browsing movies to recommend. We encourage the recommenders to continue the conversation until seekers accept a movie.

Seeker. Seekers are asked to talk about movie recommendations without any strategy support. After they complete the conversation, seekers can opt to accept or reject the provided movie recommendations. If the seekers accept the recommendation, they can watch the entire recommended movie trailer or part of it, or simply skip it after the conversation. We record how long seekers watched the recommended movie trailer and ask them to rate the trailer on 5-Likert scale in the post-task survey.

[2] youtube.com/user/movieclipsTRAILERS

Dataset Statistics	
# Dialogs	1,001
# Utterances	35,811
Average turns per dialog	10.73
Average tokens per utterance	7.93
# Unique tokens	18,316
Recommender's Statistics	
# Utterances	18,339
Average tokens per turn	14.64
# Unique tokens	13,753
Seeker's Statistics	
# Utterances	17,472
Average tokens per turn	12.12
# Unique tokens	10,097

Table 2: INSPIRED's statistics. # denotes the number.

Cases	**#Dialogs**	
Accept (Rating 4-5)	532	(53.1%)
Accept (Rating 3 or lower)	45	(4.5%)
Accept (Other Reasons)	289	(28.9%)
Accept Uninterested	123	(12.3%)
Reject	12	(1.2%)

Table 3: Statistics of dialogs when the seekers accept or reject the final recommended movie. "Accept (Rating 4-5)" means that the seekers accept the recommendation and give rating 4 or 5, and the same is for "Accept (Rating 3 or lower)". "Accept (Other Reasons)" suggests that the seeker gives other reasons for not finishing the video. "Accept Uninterested" indicates that the seekers accept the recommendation, do not finish watching the video, and explains in the post-task survey that they are not interested in the recommended video.

3.3 Dialog Data Collection Details

We use ParlAI platform (Miller et al., 2017) and hire 1,594 US crowd-workers from Amazon Mechanical Turk with a minimum of 90% task acceptance rate. The dialog collection process lasted from November 2019 to March 2020.

Workers first fill out questionnaires related to their personality traits and values before their conversations. The questionnaire consists of three personality trait models: the Big Five personality traits (15 questions) (Goldberg, 1993), the Schwartz Portrait Value (10 questions) (Schwartz, 2003), and the Decision Making Style (2 questions) (Hamilton et al., 2016)[3]. Then, recommenders start the conversation and both workers should chat for a minimum of 10 turns or until a recommendation is made. After the conversation ends, both workers will answer a post-task survey of demographic questions such as age, and gender. Seekers are asked to rate the trailer with a high score (4 or 5 stars) on a 5-Likert scale and provide the reason of why they reject or do not finish watching the video. Both workers receive a bonus up to $2 if they complete the entire process in addition to the base pay of $0.5.

Table 2 presents statistics of the collected dataset[4]. Even though our dataset has relatively small number of samples compared to REDIAL or GORECDIAL, it has human annotations on each sociable strategy. Moreover, our dataset can be used in combination with other datasets in a semi-supervised setting, as shown in our implementation of recommendation dialog systems in §6.

The statistics of accept and reject cases are shown in Table 3. We have higher number of successful cases (79.7%) compared to failure cases. This shows that people tend to accept recommendations, and it is not surprising since watching a video trailer is an entertaining, low-risk activity. For training the dialog model, we use every dialog from all cases so that the dialog system will be able to respond to diverse responses.

4 Recommendation Strategy Annotation

4.1 Strategy Definition

After conversations are collected, two experts, trained with linguistics background, develop an annotation scheme using content analysis method (Krippendorff, 2004) and from past study on human behavior in making recommendations. Similar approaches have been done in prior studies on work for persuasion task (Wang et al., 2019) or negotiation task (Zhou et al., 2019). We divide the recommendation strategies into two categories: sociable strategies and preference elicitation strategies. Sociable strategies are also derived from our literature study on the social science theories.

Sociable strategies contain eight strategies related to the recommendation task. These strategies relate to the recommenders trying to build rapport with the seekers.

- **Personal opinion** refers to a condition when recommenders express their subjective opinion

[3] We also release this personality information in our dataset for future work

[4] Dialog collection interfaces are in appendix H in Appendix

Category	Example
Personal Opinion	“I really like Disney’s more recent princesses”
Personal Experience	“I have Disney+ and watched it everyday!”
Similarity	“Oh, I love Disney as well.”
Encouragement	“You should definitely watch it!”
Offering Help	“I’m here to help you find a trailer!”
Preference Confirmation	“So do you like Disney movies in general?”
Credibility	“It’s about a dog named Lady who runs away with a stray named Tramp”
Self-Modeling	“We are planning to go see Maleficent, we heard it was a very good movie.”
Experience Inquiry	“Have you seen the new Lady and the Tramp?”
Opinion Inquiry	“What do you like about the Avengers: End-game?”
Recommendation	“You should check out Shazam!”

Table 4: Example utterances for each strategy.

about a movie, including its plot, actors, or other movie attributes.

- **Personal experience** refers to the use of sharing personal experience related to a movie. For example, recommenders may say that they watch the movie several times to convince the seekers that the movie is good. Both personal opinion and personal experience are part of self-disclosure that leads to establishing rapport with the seekers (Altman, 1973).

- **Similarity** refers to a condition when the recommenders are empathizing and being like-minded toward seekers about their movie preference to produce similarity among them. Similarity is believed to influence the seekers’ liking for the source that leads to trust the recommenders’ judgment more (O’Keefe, 2004), following Lazarsfeld and Merton (1964)’s homophily theory that states humans like other people who are similar to them.

- **Encouragement** is the use of praise of the seekers’ movie taste and encouragement to watch a recommended movie to build rapport and promote the recommended movie.

- **Offering help** is a strategy when the recommenders disclose explicit intention to help the seeker or being transparent. It is a part of “transparency” strategy from Gretzel and Fesenmaier (2006).

- **Preference confirmation** is a strategy when the recommenders ask or rephrase the seeker’s preference. This strategy is also a part of “transparency” strategy which states that the recommenders disclose their thinking process of understanding the seekers’ preference.

- **Self-modeling** is a strategy when the recommender becomes a role model to do something first so that the Seeker would follow (Dowrick, 1999).

- **Credibility** happens when the recommender shows expertise and trustworthiness in providing information to persuade the seeker (Fogg, 2002; O’Keefe, 2004; Rhoads and Cialdini, 2002). In our study, a recommender is doing credibility appeal when they provide factual information about movie attributes, such as the plot, actors, or awards that the movie has.

Preference elicitation inquiries include the following inquiries that are asked by the recommenders to know the seekers’ movie tastes.

- **Experience inquiry** asks for seeker’s experience on movie watching, such as whether a seeker has watched a certain movie or not.

- **Opinion inquiry** asks for seeker’s opinion on movie-related attributes. Example answers for this inquiry is the seeker’s explanation on what they like about the plot or if they admire the actors’ acting skill.

Other kinds of utterances, such as greetings or thanks, fall into non-strategy category. We also label sentences which are recommendation. Recommendation is defined as when the recommender

Category	#Utterances	
Sociable Strategies		
Credibility	2,687	(13.7%)
Personal Opinion	2,599	(13.9%)
Encouragement	1,975	(10.6%)
Similarity	957	(5.1%)
Offering Help	953	(5.1%)
Preference Confirmation	950	(5.1%)
Personal Experience	564	(3%)
Self-Modeling	449	(2.4%)
Preference Elicitation Inquiries		
Experience Inquiry	1,505	(8.1%)
Opinion Inquiry	2,120	(11.3%)
Non-strategy		
No Strategy	2,566	(13.7%)
Acknowledgment	1,354	(7.2%)
Recommendation	2177	(6.1%)

Table 5: Statistics of the number of utterances annotated with strategies in INSPIRED.

suggests a new movie title for the first time for the seeker. 30% of the recommendation sentences are "experience inquiries", 27% are "encouragement", and 14% are "personal opinion". Example annotated utterances are displayed in Table 4. Meanwhile, Table 5 shows the number of annotated utterances in INSPIRED.

4.2 Annotation Quality

To ensure annotation quality, we separate our annotation study in two steps. First, we hire two experts with linguistics training to perform annotation, in order to test the validity of the scheme. The two experts annotated 30 randomly selected conversations and reached a Kappa agreement of 0.77, suggesting that our scheme is possible to replicate.

Our dataset contains more than 18k utterances, so it's too costly to hire experts to annotate all of them. In the second step, We hire US-based crowd-workers (95% task acceptance) from Amazon Mechanical Turk for the annotation tasks. In each task, a worker was given a tutorial of the annotation and then they were given 10 dialogs to annotate. One of the dialogs was labeled by experts to calibrate the quality of the worker's annotation, called as evaluation dialog. Five workers work on the same task. We filter out workers whose score is below the threshold 0.60 on the evaluation dialog. To set this threshold in a reasonable value, we conducted the following study. This time we ran onetask in which all the dialogs are already labeled with the experts including the evaluation dialogs. We found that if the workers' score on the evaluation dialog is above 0.60, their agreement score with the expert's annotation on the rest of the dialogs in this task is 0.77.

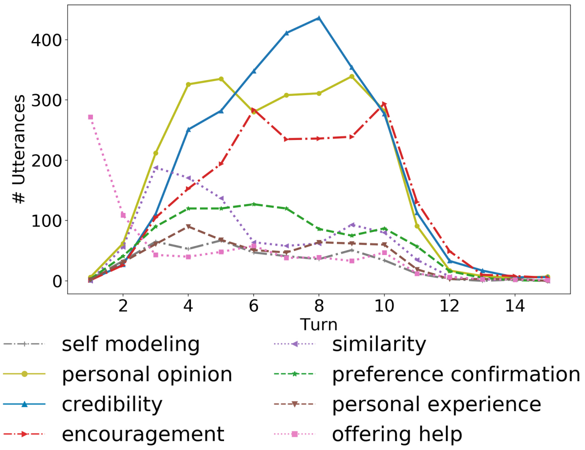


Figure 3: Distribution of sociable strategies over the dialog turns. Best viewed in color.

These selected high quality crowd-workers annotate the rest of the dialogs. We still have five workers annotate the same dialog. If more than one worker disagrees on a utterance's annotation, the experts are then involved to annotate them as quality control. The inter-annotator majority agreement among all workers is 0.78 over all dialogs. The annotation scheme for the crowd-workers are provided in Figure 12 in the Appendix.

5 Strategy Analyses

5.1 Distribution of Strategies over Dialog

As shown in Figure 3, we observe that different sociable strategies are unequally distributed across conversation turns. Most notably, "offering help" and "similarity" often happen at the beginning, indicating that recommenders strategically attempt to build rapport with seekers at the early stages. Then, "credibility" and "personal opinion" frequently appear in the conversations, as recommenders seek to persuade. Moreover, "encouragement" mostly appears in the middle and at the end of conversations.

5.2 What Strategies Contribute to Successful Recommendations?

We study the association of sociable strategies and successful recommendations. A recommendation is considered successful if seekers finish watching a substantial portion of the recommended movie trailer and rate the trailer with a high score (4 or 5 stars). We set a threshold that seekers need to watch at least more than 50% of the video duration since some videos have advertisements at the end, etc. On the other hand, a recommendation is considered unsuccessful if the seekers reject the recommendation (“Reject”) or skip watching the trailer (“Accept Uninterested”). Thus, for our analysis, we use 532 successful dialogs and 135 unsuccessful dialogs for our analysis on association of strategies in successful recommendations.

To analyze the effect of our sociable recommendation strategies on success of recommendation, we run a logistic regression model to predict the success of recommendation (1 = successful, 0 = unsuccessful). We use frequency of the strategy in a dialog as the feature value.

Table 6 shows the coefficients of each strategy with respect to the recommendation. We observe that “personal opinion”, “similarity”, “encouragement”, and “credibility” strategies have a significant positive effect on successful recommendations. This confirms with the previous studies that more sociable recommenders are more likely to be successful in the recommendation.

“Similarity” strategy has the highest coefficient value which suggests that if the recommender is conforming to the seeker’s preference, the seeker is more likely to favor the recommendation. This also supports the theory in O’Keefe (2004) that likeability helps in recommendation. We also observe that all the preference elicitation inquiries are not significantly contributing to the successful recommendation. From this result, we are not saying that recommenders need not to query seekers’ preferences since it is crucial to understand their tastes. However, a more sociable approach is necessary for a more successful recommendation.

Category	**Coefficient**
Sociable Strategies	
Personal Opinion	0.12*
Personal Experience	0.05
Similarity	**0.23***
Encouragement	0.20**
Offering Help	0.03
Preference Confirmation	0.05
Self-Modeling	0.02
Credibility	0.09*
Preference Elicitation	
Experience Inquiry	−0.01
Opinion Inquiry	0.06

Table 6: Associations between different strategies and successful recommendation. $*p < 0.05$, $**p < 0.01$

5.3 Are Sociable Strategies Still Significant with the Presence of Movie Attributes?

In a recommendation task, a natural question to ask is how big a role the recommended product plays in the acceptance of recommendation. If the quality of the product matters more than how you recommend, it makes more sense to improve the products rather than the recommendation skills. Therefore, we also analyze if adding movie attributes, such as the genre, recent release date, and the number of likes of the movie trailer have an impact on successful recommendation along with the eight sociable strategies and two preference elicitation inquiries.

For the popularity, we categorize the top 10% movies in terms of the number of likes to be popular and the rest to be non-popular in our database. A movie is said to be recent if it is released in 2019 or 2020. For the genre, we select the top five most popular genres in the movie database. When we check with the recommended movies in INSPIRED, 96% of recommended movies are covered by the top five genres.

Results of the analysis between the strategies and movie attributes are shown in Table 8 in the Appendix. Sociable strategies remain significantly correlated with successful recommendations. Recommenders who perform “similarity” strategy, express “personal opinion”, and show “encouragement” are more likely to successfully recommend a movie ($p < 0.05$). Surprisingly, none of the movie attributes has significant effect on successful recommendations. A possible reason is that the seekers’ movie tastes are so diverse that movie attributes such as genre do not have a significant impact on the recommendation success.

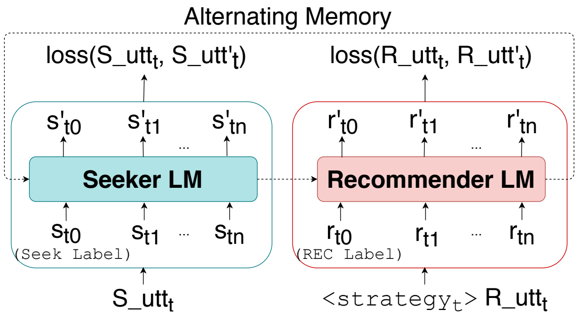

Alternating Memory
loss(S_utt$_t$, S_utt'$_t$)
loss(R_utt$_t$, R_utt'$_t$)
s'$_{t0}$ s'$_{t1}$ ... s'$_{tn}$
Seeker LM
s$_{t0}$ s$_{t1}$ ... s$_{tn}$
(Seek Label)
S_utt$_t$
r'$_{t0}$ r'$_{t1}$ ... r'$_{tn}$
Recommender LM
r$_{t0}$ r$_{t1}$ ... r$_{tn}$
(REC Label)
<strategy$_t$> R_utt$_t$

Figure 4: The Seeker's language model (Seeker LM) and the Recommender's language model (Recommender LM) are separate memory. The Seeker LM input at turn t is the seeker's utterance S_utt_t consisting of a sequence of tokens $s_{t0}, s_{t1}, ...s_{tn}$. The Recommender LM input at turn t is the recommender's utterance R_utt_t consisting of a sequence of tokens $r_{t0}, r_{t1}, ..., r_{tn}$. The <strategy$_t$> prepended as a special token. For the baseline, the recommender's input does not contain the strategies.

6 Recommendation Dialog Systems

To evaluate how the strategies in INSPIRED are useful in creating a more engaging and persuasive recommendation dialog, we develop a generative dialog model as our baseline to compare against our strategy-incorporated dialog system. We split the dialogs into 801/100/100 for train/validation/test split. We use external recommendation system from TMDB[5] with heuristics to select the movies. More details for heuristics and training set-up are in the Appendix.

6.1 Baseline Model

The baseline dialog model uses two separate Transformer-based pretrained language models (Vaswani et al., 2017; Radford et al., 2019; Wu et al., 2019) to learn the recommender's and seeker's language models separately in alternating order. Both language models are trained to maximize the likelihood of generating ground truth utterance on the alternating memory as shown in Figure 4. The model is pretrained on non-task related corpus, WebText, and task-related corpus: recommendation dataset from REDIAL (Li et al., 2018) and movie preference elicitation dataset (Radlinski et al., 2019). Then, we fine-tune the model with INSPIRED.

We replace movie attributes such as titles, actors, and genres with indexed placeholders. It is because in a single conversation, multiple attributes may be mentioned several times. The replacement with placeholders improves factual correctness as we replace them back with the original movie attributes later. At the end of the sentence, we append the attribute information as below:

Original: "If you like La La Land, you should also see Amazing Spiderman with Emma Stone"

With placeholder: "If you like [MOVIE_TITLE_0], you should also see [MOVIE_TITLE_1] with [MOVIE_P_ACTOR_0]; movies: La La Land (2016), The Amazing Spider-Man (2012); people: Emma Stone"

6.2 Strategy-incorporated Model

We prepend the strategy as a special token to the input utterance so that the model does not only generate sentences but also strategies. Similar method was used to control text generation style (Rashkin et al., 2019) as a simple and effective way to incorporate the strategies. The input to the encoder is as follows:

Prepend: "encouragement If you like [MOVIE_TITLE_0], you should also see [MOVIE_TITLE_1] with [MOVIE_P_ACTOR_0]; title: La La Land (2016), The Amazing Spider-Man (2012); people: Emma Stone"

The model first generates five candidate sentences. Then, it randomly selects a generated candidate that either contains "encouragement" strategy or has the greatest sentence length. In our experiment, we have tried various combinations of the top three strategies (e.g., "encouragement" only, "encouragement" and "similarity"), and it turns out that "encouragement" only model gave the best result. Moreover, the sentence length selection is based on our intuition when chatting with the system. This aligns from our findings, "encouragement" is the second most frequently used strategy when humans make recommendations (§4.1), and "recommendation" is associated positively with successful recommendation (Table 8)[6].

To decide if a sentence is a recommendation or not, we train a BERT-based recommendation classifier that receives an input of recommender's current utterance and seeker's utterances from previous turn with 95.4% accuracy and 91.2 % F1-score. While the index in the placeholder may become a

[5] https://www.themoviedb.org/

[6] We tried multi-task learning for utterance generation and strategy prediction, and the automatic metric result is comparable with the prepend method.

Model	PPL↓	BLEU-4↑
Baseline	9.28	5.11
Strategy	8.93	6.63

Table 7: Results for automatic metrics.

proxy to decide whether the system needs to recommend a movie or not, it is not strictly supervised. Thus, if a generated sentence is labeled as "recommendation", we enforce our dialog system to recommend a new movie.

6.3 Results

We compare the baseline dialog model without strategy supervision against our dialog model with strategy supervision. We use both automatic metrics and human evaluation.

For automatic metrics, we compute perplexity and BLEU scores (Papineni et al., 2002), suggesting that prepending strategies improves the model performance as shown in Table 7. For human evaluation, twenty-eight participants chat with both models for 2-3 times for a more reliable judgment. We randomize which model they will chat first, in order to avoid exposure bias. After chatting, they are asked to decide which model is better in these five aspects: fluency, consistency, naturalness, persuasiveness, and engagingness. If they are unable to distinguish the dialog systems, they are allowed to choose "can't tell" option.

Results in Figure 5 suggest that human users prefer the model with strategy over the baseline in all aspects[7]. It is interesting to see that although the strategy model is preferred on all metrics, people find the two model differs the most in engagingness, followed by naturalness. This supports our hypothesis that human users will find the conversations more engaging and more natural with sociable strategies incorporated in recommendation dialog systems.

7 Conclusion and Future Work

In this work, we have introduced INSPIRED, a new recommendation dialog dataset collected in natural setting and annotated with sociable recommendation strategies. We analyze the connection between different strategies and the recommendation results. Our findings show that sociable strategies do have a positive impact on the acceptance of recommendation and dialog quality. This work opens up several directions for future studies in building sociable and personalized recommendation dialog systems as follows:

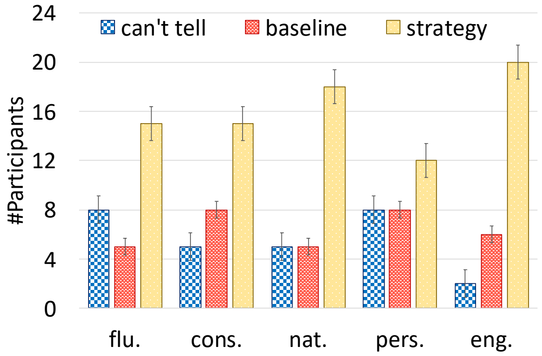


Figure 5: Human evaluation result. "Flu." stands for fluency, "cons.": consistency, "nat.": naturalness, "pers.": persuasiveness, and "eng.": engagingness.

First, we will explore more ways of utilizing the strategies, including dynamic strategy selection after decoding. Then, we plan to investigate the strategy patterns for people with different personalities and movie preferences to make dialog system more personalized. Finally, another interesting exploration is to extend the model with a jointly trainable movie recommendation and movie information modules.

Acknowledgments

We would like to thank members of the NLP lab at UC Davis for discussion and participation in the pilot study. We are also grateful for human evaluation participants and Mechanical Turk workers on contributions of building this dataset.

References

Irwin Altman. 1973. Reciprocity of interpersonal exchange. *Journal for the Theory of Social Behaviour*, 3(2):249–261.

Timothy Bickmore and Justine Cassell. 2005. Social dialongue with embodied conversational agents. In *Advances in natural multimodal dialogue systems*, pages 23–54. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Peter W Dowrick. 1999. A review of self modeling and

[7] We also run additional user study with five-scale ratings on these five aspects with results in Table 10 in the Appendix

related interventions. *Applied and preventive psychology*, 8(1):23–39.

David Schlangen Fabian Galetzka1, Chukwuemeka U. Eneh. 2020. A corpus of controlled opinionated and knowledgeable movie discussions for training neural conversation models. In *Language Resources and Evaluation Conference (LREC)*, volume 12.

B. J. Fogg. 2002. Persuasive technology: Using computers to change what we think and do. *Ubiquity*, 2002(December).

Sofia Gkika and George Lekakos. The persuasive role of explanations in recommender systems.

Mehmet H Göker and Cynthia A Thompson. 2000. Personalized conversational case-based recommendation. In *European Workshop on Advances in Case-Based Reasoning*, pages 99–111. Springer.

Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist*, 48(1):26.

Ulrike Gretzel and Daniel R. Fesenmaier. 2006. Persuasion in recommender systems. *International Journal of Electronic Commerce*, 11(2):81–100.

Katherine Hamilton, Shin-I Shih, and Susan Mohammed. 2016. The development and validation of the rational and intuitive decision styles scale. *Journal of personality assessment*, 98(5):523–535.

F. Maxwell Harper and Joseph A. Konstan. 2015. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19.

Gerald Häubl and Kyle B Murray. 2003. Preference construction and persistence in digital marketplaces: The role of electronic recommendation agents. *Journal of Consumer Psychology*, 13(1-2):75–91.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. 2019. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. 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 1951–1961, Hong Kong, China. Association for Computational Linguistics.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433.

Paul F. Lazarsfeld and Robert King Merton. 1964. Friendship as social process: a substantive and methodological analysis. pages 18–66, New York. Van Nostrand.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems*, pages 9725–9735.

Kaihui Liang, Austin Chau, Yu Li, Xueyuan Lu, Dian Yu, Mingyang Zhou Zhou, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen, and Zhou Yu. 2020. Gunrock 2.0: A user adaptive social conversational system. In *3rd Proceedings of Alexa Prize (Alexa Prize 2020)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Daniel James O'Keefe. 2004. Trends and prospects in persuasion theory and research. In *Readings in persuasion, social influence, and compliance gaining*, pages 31–43. Pearson/Allyn and Bacon.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Florian Pecune, Shruti Murali, Vivian Tsai, Yoichi Matsuyama, and Justine Cassell. 2019. A model of social explanations for a conversational movie recommendation system. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, HAI '19, pages 135–143, New York, NY, USA. ACM.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding movie preferences. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*.

Kevin Reschke, Adam Vogel, and Dan Jurafsky. 2013. Generating recommendation dialogs by extracting information from user reviews. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 499–504, Sofia, Bulgaria. Association for Computational Linguistics.

Kelton V. Rhoads and Robert B. Cialdini. 2002. The business of influence. In *Persuasion handbook: Developments in theory and practice*, pages 513–542, London, United Kingdom. Sage.

Shalom H Schwartz. 2003. A proposal for measuring value orientations across nations. *Questionnaire Package of the European Social Survey*, 259(290):261.

Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1509–1519, Melbourne, Australia. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 235–244. ACM.

Cynthia A Thompson, Mehmet H Goker, and Pat Langley. 2004. A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research*, 21:393–428.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. 2019. Alternating recurrent dialog model with large-scale pre-trained language models. *arXiv preprint arXiv:1910.03756*.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Kyung-Hyan Yoo, Ulrike Gretzel, and Markus Zanker. 2012. *Persuasive Recommender Systems: Conceptual Background and Implications*, 1st edition. Springer Publishing Company, Incorporated.

Markus Zanker, Marcel Bricman, Sergiu Gordea, Dietmar Jannach, and Markus Jessenitschnig. 2006. Persuasive online-selling in quality and taste domains. In *International Conference on Electronic Commerce and Web Technologies*, pages 51–60. Springer.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W Bruce Croft. 2018b. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 177–186. ACM.

Yiheng Zhou, He He, Alan W Black, and Yulia Tsvetkov. 2019. A dynamic strategy coach for effective negotiation. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 367–378, Stockholm, Sweden. Association for Computational Linguistics.