Effective Social Chatbot Strategies for Increasing User Initiative

Amelia Hardy
Stanford University
amelia@cs.stanford.edu

Ashwin Paranjape
Stanford University
ashwinp@cs.stanford.edu

Christopher D. Manning
Stanford University
manning@cs.stanford.edu

Abstract

Many existing chatbots do not effectively support mixed initiative, forcing their users to either respond passively or lead constantly. We seek to improve this experience by introducing new mechanisms to encourage user initiative in social chatbot conversations. Since user initiative in this setting is distinct from initiative in human-human or task-oriented dialogue, we first propose a new definition that accounts for the unique behaviors users take in this context. Drawing from linguistics, we propose three mechanisms to promote user initiative: back-channeling, personal disclosure, and replacing questions with statements. We show that simple automatic metrics of utterance length, number of noun phrases, and diversity of user responses correlate with human judgement of initiative. Finally, we use these metrics to suggest that these strategies do result in statistically significant increases in user initiative, where frequent, but not excessive, back-channeling is the most effective strategy.

1 Introduction

Chatbots serve in a broad range of roles, from task-based assistants to social companions. However, the experience of dialogue agents differs starkly from human interaction. One key difference is how *initiative*, or conversational control, is shared. In human-human dialogue, participants take turns leading the conversation and introducing new topics (Walker and Whittaker, 1990), but in most human-bot dialogue the bot either leads unilaterally (social chatbots) or responds passively (task-oriented assistants). Socially, it is unpleasant and unengaging to be responsible for all topic selection and equally undesirable to have no agency. However, these one-sided experiences persist as a natural consequence of rule-based bot design, which only works well when user intents are narrowly confined. With advances in neurally generated dialogue, agents can now handle less-restricted user responses, but require the adoption and development of specific mechanisms that encourage the user's initiative. By studying these methods, we seek to create a more human-like and engaging experience.

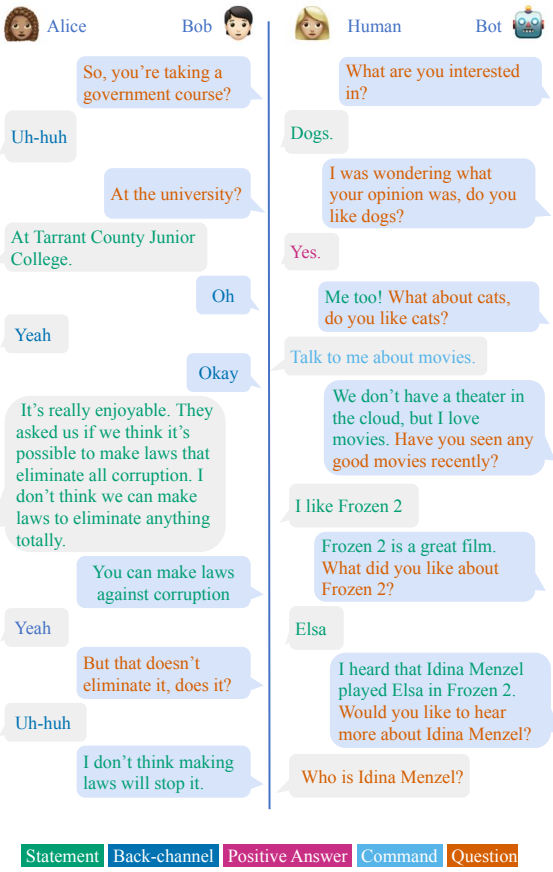

Figure 1: In Human-Human dialogue[1](left), Bob first takes initiative by asking a question, and then uses back-channels to encourage Alice to take initiative, which she does by introducing a new topic: corruption. In a typical current Human-Bot dialogue[2], the bot has initiative and the user responds passively and compliantly, except when interjecting to give a command or ask a question.

[1]From the Switchboard dataset, edited for length and clarity
[2]This dialogue is representative of user conversations with our bot; however, it does not contain any actual user data

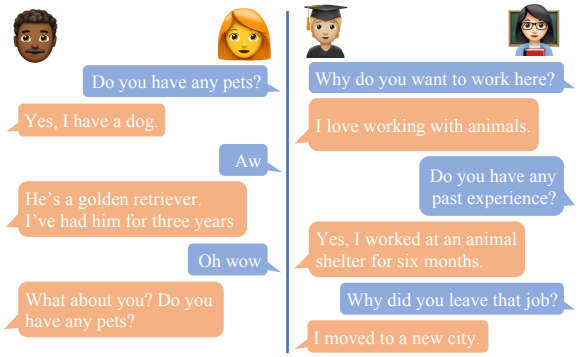

Figure 2: In a cocktail party setting, participants Alice and Bob take turns directing the conversation. First, Alice asks Bob about himself and later, Bob asks Alice about herself. In the job interview scenario, the interviewer sets all topics and the interviewee passively responds, following the interviewer's direction

Existing work on mixed-initiative human-bot dialogue has focused on task-oriented settings, where the space of potential user actions is smaller and success is easier to measure (Horvitz, 1999; Allen et al., 1999; Heeman et al., 2003; Core et al., 2003). Prior work on social dialogue is limited to human-human conversations, which also have different patterns and mechanisms of initiative compared with human-bot social conversations. But neither lines of work effectively transfers to the human-bot social conversations. **Our first contribution is defining granular levels of user initiative in the context of an open-domain social chatbot.**

Current social chatbots designs do not explicitly consider user initiative, neither measuring nor encouraging it. We propose measuring user initiative with automated metrics: utterance length, noun phrases (for meaningful content), and response entropies (for diversity) and validate their correlation with user initiative with a small study (Section 6.2). **Informed by work in sociolinguistics and psychology, our second set of contributions are three strategies for increasing user initiative in open-domain human-bot conversations**.

First, **back-channeling** or giving responses such as "I see" or "Mm-hmm". Discourse research suggests that back-channeling signals the other speaker to continue directing the conversation (Duncan, 1974). Second, **using open-ended statements as prompts**, because repeatedly forcing the user to respond to questions limits their agency. Third, **self-disclosure by the conversational agent**, which has been shown to have a reciprocal effect on users (Lee et al., 2020), since sharing unprompted information indicates higher initiative (Cohen et al., 1999).

We study the effect of these strategies in an Alexa Prize bot, a unique research setting where users engage with the bot socially for the sole purpose of entertainment (Section 4). All three strategies significantly increase user initiative as measured by the automatic metrics. Separately, we annotate a small subset of utterances with the level of initiative taken by the user to validate our metrics (Section 6.2). We find that a simple strategy of back-channeling on one-third of turns encourages many users taking low initiative to start taking high initiative. Replacing questions with statements increases average user utterance length by 23%, in particular, personal statements are very effective in encouraging low initiative taking users to take medium or high initiative. We verify these findings by annotating another set of user utterances, to confirm that the observed increases in automated metrics are truly reflective of increased user initiative (Section 7.5). Our results suggest that incorporating these mechanisms into future chatbot design will facilitate greater user control and more engaging, human-like conversations.

2 Rethinking Initiative

Initiative is a participant's degree of control at a given moment. Consider two dialogue settings with markedly different patterns of control: the cocktail party and the job interview (Figure 2). At a cocktail party participants share the agency to direct the conversation and take initiative in turns, whereas the interviewer takes initiative throughout the interview and retains control of the conversation's direction.

In human-bot social conversation, a user who steers the conversation by suggesting new topics has high initiative, whereas one who follows the bot's lead has low initiative. We examine ideas from prior work on human-human (Section 2.1) and *task-oriented* human-bot (Section 2.2) conversation and build upon them to offer a novel definition (Section 2.3) of initiative in human-bot *social* conversation.

2.1 Human-Human Conversation

Control rules based on dialogue acts have been proposed (Whittaker and Stenton, 1988; Walker and Whittaker, 1990); however they do not account for varying degrees of initiative which are common in social conversations. Addressing this, Cohen et al. (1999) defines initiative on a spectrum. For example, a command ("Let's talk about cats") is stronger than a suggestion ("Maybe we should talk about cats"). **We extend this idea and account for the effect**

of conversational context on the degree of initiative in an utterance. For instance, the answer "I love dogs" displays a lower initiative in response to "What's your favorite animal?" but higher initiative in response to "What would you like to talk about?". In the first case, the other speaker set the overall direction of the conversation to be about animals whereas in the second case it was left open and the topic was chosen from a wider variety of options.

Determining who has initiative also depends on the granularity at which it is being measured. Chu-Carroll and Brown (1998) formalize this notion for *task-oriented* dialogues. One speaker can set the overarching task level initiative (making a reservation) while the other can take utterance level initiative (asking for information, e.g., reservation time). **Such a hierarchy is too restrictive for social dialogue so we consider instead the notion of *local* initiative, which considers how an utterance alters the bot's path.** For example, replying "I like dogs, what about you?" to "What's your favorite animal?" takes more initiative at the utterance level than replying "cats" because the former likely changes the conversation's direction, while the latter stays the course.

2.2 Human-Bot Conversation

Past work on initiative in human-bot conversations has focused on a task-oriented setting (Novick and Sutton, 1997; Horvitz, 1999; Allen et al., 1999; Harms et al., 2019). In this setting, initiative frameworks are based on "collaboration" around a goal, which is accomplished through a series of sub-goals. Although collaborative, social conversation has no clearly-defined objective. The closest analogue is topic, since just as task-oriented conversation breaks down into units of sub-goals, social conversation breaks down into units of topics. **We therefore consider the degree of contribution to topical direction as initiative.**

Defining a dialogue act schema for human-bot social conversations, Yu and Yu (2021) highlight key differences from human-human dialogue acts, most notably the prevalence of user commands as a means of directing conversation. This brings to the fore the asymmetry of the human-bot social setting. Current implementations of social chatbots railroad the user and are less perceptive to implicit cues. This forces the user to use explicit commands to take initiative, which is uncommon in human-human conversations, since humans generally prefer interrogatives over imperatives when making requests (Ervin-Tripp, 1976).

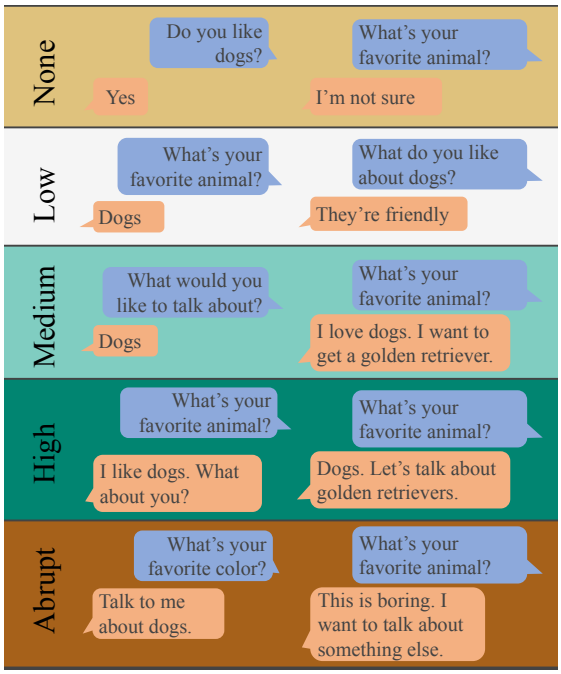

Figure 3: Although the user (orange) and bot (blue) use the same basic dialogue acts in these scenarios, with the bot asking questions and the user replying in statements, their relative levels of initiative differ based on context. The rows of this figure illustrate possible (question, statement) pairs for a given level of initiative.

2.3 Defining Initiative for Social Chatbots

We are now ready to define initiative in the social chatbot domain. Drawing from work on human-human conversations, we define initiative on a spectrum. While dialogue acts are necessary for determining initiative, they are not sufficient. For example, the user (orange) in the scenarios illustrated in Figure 3 always responds with a statement, but has differing levels of initiative. For this reason, we also consider context in our definition.

As these examples show, simple dialogue act-based heuristics fail to capture the more nuanced degrees of initiative. In all of the examples, the bot is asking a question and the user is answering it; however, the user has varying degrees of control. Determining which participant has initiative depends on dialogue act, content, and context.

Definition Based on the extent to which the user is changing the conversation's path, we determine their degree of initiative to be either None, Low, Medium, High, or Abrupt. We say that the user's initiative level is **None** when the user's utterance does not alter the bot's dialogue path. For instance,

this is the case when the user gives a yes/no answer to a yes/no question, since they are choosing between two options pre-defined by the bot. The user also takes no initiative when responding "I'm not sure" to the bot's question, since this answer does not in any way steer the conversation.

The user has more initiative when responding compliantly to the questions shown on the **Low** level. These questions give the user more flexibility than those on the None level, but still limit the response space by confining it to a particular topic. Nonetheless, when answering these questions, the user is able to assert some meaningful direction. On the **Medium** level, the user has greater initiative when answering "What would you like to talk about?" since this question offers even broader control. The user also has greater initiative when answering "What's your favorite animal?" with "I love dogs. I want to get a golden retriever," since they are sharing information outside the expected response and thus contributing to the dialogue's course beyond what they were asked for. The user has **High** initiative both when asking questions and giving commands. These actions directly assert a divergence from the bot's proposed direction. We intentionally distinguish these cases from those on the **Abrupt** level, since in the latter the user is taking initiative in a way that shows discontent and which would be unnatural in human-human conversation.

3 Mechanisms of Initiative

Our goal is to improve the quality and naturalness of social bot conversations by enabling and encouraging the user to take greater initiative. We study three mechanisms for increasing user initiative: statements, back-channels, and personal disclosure.

Statements. In human dialogues, utterance type predicts shifts of control (Whittaker and Stenton, 1988). We focus in particular on the effect of statements. When codifying changes in initiative, Whittaker and Stenton (1988) define four utterance categories: questions, assertions (declarative, factual statements), commands, and prompts (utterances without propositional content, e.g. "uh huh"). Whittaker and Stenton propose control rules based on these categories. Notably, the schemas of both (Walker and Whittaker, 1990) and (Whittaker and Stenton, 1988) do not consider a control shift to take place if the listener is responding compliantly to the speaker's question, since the question is controlling the conversation's direction.

Duncan (1972) associates similar actions with a change in control. He gives six "turn-yielding signals," which are behavioral cues from the speaker to the listener that the listener should start talking. Of these signals, four out of six cannot be replicated on our bot, since they depend on dialogue features that our bot neither gives, nor receives: pitch, intonation, and body language. The remaining two are trailing off sequences, such as "you know" and syntactic completion of a grammatical clause. It follows from the conclusions of (Duncan, 1972) and (Whittaker and Stenton, 1988) that while both statements and questions cue the user to take a turn, statements alone truly provide them with the opportunity for initiative on that turn.

Back-channeling. In addition to statements, we study back-channels as a signal that the user should take initiative. Duncan distinguishes turn-yielding signals from back-channels. Since they do not introduce new content, back-channels do not constitute a turn (Duncan, 1974). Instead, Duncan finds that they are used by the listener to signal that the speaker should continue. Turn-yielding signals, which tell the listener to begin speaking, trigger a change in speaker, while back-channels do not. (Whittaker and Stenton, 1988) also observe that back-channels are used by one participant to give control to the other. However, (Whittaker and Stenton, 1988) frame this slightly differently, with control transferring from the speaker to the listener. Simultaneous back-channelling is a central marker of shifting control in human-human conversations. However, chat bots cannot perfectly replicate this behavior due to technical limitations which allow only one speaker at a time.

Personal Disclosure. The final mechanism we study is the use of personal self-disclosure as a means for increasing user participation. In human-human conversations, self-disclosure not only increases connection, but produces "disclosure-reciprocity effect": when one participant discloses, the other is more likely to disclose as well (Collins and Miller, 1994). This effect has also been measured in human-bot conversations. Chatbot self-disclosure encourages users to share more about themselves than they would otherwise (Lee et al., 2020). Increasing this behavior increases user control, since sharing information without an explicit prompt is a form of initiative (Cohen et al., 1999). Figure 3's Medium level gives an example of how greater user sharing increases initiative.

4 Our Bot

We conducted our experiments using an Alexa Prize competition bot (Khatri et al., 2018). A user saying "let's chat" to an Alexa device is randomly connected to one of the bots participating in the competition. To protect user privacy, teams receive user utterances as text only, so we could not leverage the additional signals, such as intonation, that are present in audio recordings. Explicit evaluation is limited to a single and optional Likert-scale rating at the end of the conversation. Alexa Prize Likert ratings have been shown to be noisy (Khatri et al., 2018); however, the competition rules prevent introducing more fine-grained evaluation questions. Instead, we use other automated metrics, as described in Section 6.

Our bot has a modular design, which allows us to restrict our experiments to the modules that are most compatible. Specifically, these are the modules that are partially or entirely neural, such as our neural chit-chat module, since they are more flexible to changing user behaviors. Amazon user data is confidential, so dialogues shown in this paper are taken from the authors' interactions with the bot. They are representative of typical user conversations, based on an extensive survey of conversation transcripts.

5 Experiment Design and Setup

We conduct four experiments in our bot, studying the effects of combining statements and questions, using personal disclosure, removing questions from responses, and back-channeling.

Comparing Statements and Questions Drawing upon the literature discussed in Section 3, we hypothesize that users will be more likely to take initiative in response to statements rather than questions.

To test whether user initiative is affected by giving a statement, asking a question, or giving a statement and then asking a question, we altered a module of our bot which uses scripted content. We wrote a set of statements and questions that could be combined in coherent pairs (Figure 4). During each conversation, we randomly selected whether users would receive a statement, statement and question, or question alone. To limit variability, we conducted this experiment on a single turn, outside of which we made no other changes.

Using Personal Statements We tested our hypothesis that users would take greater initiative in response to personal statements by randomly selecting the type of statement that users would

Figure 4: Example prompts for comparing **statements vs questions** and example replies. To a question, users generally answer compliantly, in this case by naming foods. To a statement alone, the actions users take in answering are more diverse.

Figure 5: **Statement types** and representative user responses. Users are more likely to reciprocate opinions, reciprocate to or follow up on experiences, and to either agree or disagree with general statements.

receive when given a statement or a combined statement and question. We experimented with three types of statements: personal experience, personal opinion, and general statement, as shown in Figure 5. As with the previous experiment, this was limited to a single turn.

Changing Question Frequency Expanding on our first experiment, we theorize that omitting questions across multiple conversation modules will increase initiative at a conversation-level.

Many modules of our bot rely on appending statements with questions to provide a clear continuation path. To further test the effect of questions in suppressing user initiative, we ran a new experiment across multiple scripted and non-scripted (neural) components of our bot. We removed questions from responses, a fixed percentage (0, 33, 66, or 100) of the time, leaving only the statements. The components of our bot that could not be re-designed to omit questions were not changed.

Introducing Back-channeling In human-human conversation, back-channels are used to signal that

that the listener should either begin or continue speaking (Duncan, 1974), so we hypothesize that back-channeling will increase use initiative.

Back-channeling can break up a long and contentful answer into smaller chunks that are hard for scripted components to analyze. To mitigate this effect, we limited this experiment to our bot's neural chit-chat component, since it has the greatest flexibility and takes many previous turns into account. Within this component, we replaced the generated utterances with back-channels 0, 33, 66, or 100 percent of the time. To avoid a negative and confounding user experience, we did not back-channel in response to utterances less than three words long, or to questions and commands detected by our bot's dialogue act classifier.

Dataset For the **Statement vs. Question** and **Personal Statements** experiments, we collected 8,889 turns of user conversation, which were roughly 40% Question, 40% Statement and Question, and 20% Statement. Responses including a statement were equally divided between the Personal Opinion, Personal Experience, and General Statement categories. We only collect the turn immediately following the bot utterance being studied.

We collected 157,363 turns for the **Frequency of Questions** experiment and 23,783 turns for the **Back-channeling** experiment. Both were equally divided between the 0, 33, 66, and 100 percent categories. We used all turns from a conversation with the Frequency of Questions experiment. Since the Back-channeling experiment only ran in a single module, we only analyzed turns from that module.

6 Evaluation

Although human evaluation can provide high levels of detail and accuracy, it is not scalable. This makes it an impractical method for analyzing large-scale conversational data. We therefore propose and validate a set of automated metrics as a good proxy for our levels of intiative. To evaluate our hypotheses (Section 5), we use several different metrics indicative of user initiative: user utterance length, number of noun phrases in the user utterance, and negative log likelihood of responses. We validated our metrics on a hand-labeled set of user conversations, see Section 6.2.

6.1 Metrics

Utterance Length We used utterance length as a metric, since sharing unprompted information demonstrates higher initiative (Cohen et al., 1999).

Noun Phrases Some long answers may be non-informative, such as "Uhh I'm not really sure about that," thus we also considered the number of distinct noun phrases in user responses, which we detected using spaCy[3].

Negative Log Likelihood If user initiative is truly increasing, then users would have more opportunities to take more conversational directions, so we would expect to see an increase in the diversity of their responses. This increase in diversity can be given by an increase in entropy. To compute entropy, we model the probability of a user response with a language model that had been fine-tuned on a large corpus of user responses. This model gives us the negative log-likelihood (nll) of a user response; we obtain estimated response entropy H_n from nll using a resubstitution estimate:

$$H_n = -\frac{1}{n}\sum_{i=1}^{n} \ln f_n(X_i) \quad (1)$$

where n is the number of responses we sample and f_n is our probability estimate function. If a response is unique and non-generic, then it will be less likely, resulting in a higher nll and higher entropy.

We compute $\ln f_n$ using a GPT2 model (Radford et al., 2019) fine-tuned on user data (see Appendix A.2 for details). For some utterance X_i, $f_n(X_i)$ is the probability our model assigns to that utterance. Since our goal was to test whether users were volunteering more information rather than simply answering a question, we removed turns consisting of the most common non-contentful utterances (see Appendix **??**) before calculating entropy, so that they would not dominate the measurement.

6.2 Validation

To validate that these metrics were correlated with initiative, the authors hand-labeled a set of 245 turns of conversation, where each turn was a pair (bot prompt, user response). We annotated the user's degree of initiative on each turn as either None, Low, Medium, High, or Abrupt, following the instructions in Appendix A.1 and had substantial agreement (Cohen's Kappa 0.71). Figure 6 shows the plots of our metrics' averages for each initiative level. The correlation between the automated metrics and our labeled dataset suggests that they give a reasonable estimate.

[3]https://spacy.io/

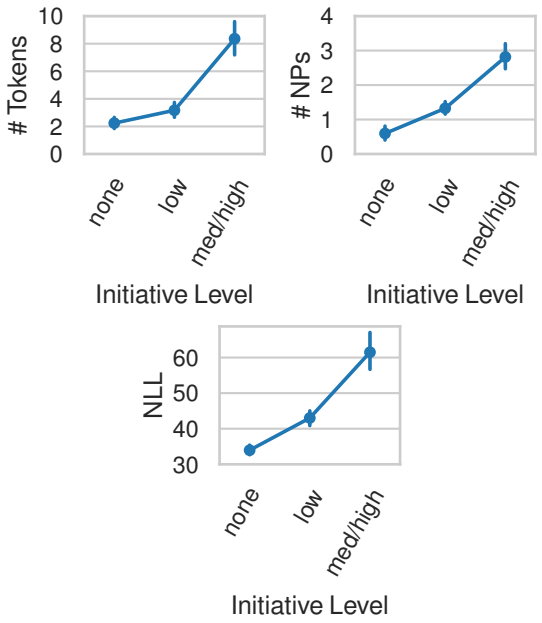

Figure 6: Automated Metrics vs. Hand-Labeled Initiative Levels. Bars show 95% confidence intervals. Due to the small number of High examples in our dataset, we collapsed Medium and High levels in the figure.

Hypo.	#tokens	#NPs	nll
Stmt	$\mathbf{4.36}^{Q,SQ}$	1.39^{SQ}	$\mathbf{21.5}^{Q,SQ}$
Stmt+Ques	$3.74^{S,Q}$	$\mathbf{1.49}^{S,Q}$	19.5^{S}
Ques	$3.55^{S,SQ}$	1.42^{SQ}	19.1^{S}

Table 1: Effect of only statement (S), statement+question (SQ) and only question (Q) on initiative. Superscript indicates significance ($p < 0.05$; paired t-test) w.r.t. other experiment.

7 Analysis and Results

7.1 Statements outperform questions

Table 1 shows the effect of using statements, questions, or combined statements and questions. We found that utterance length was greatest for statements alone and least for questions alone. Using statements increased average nll (entropy), but there was no effect on entropy when comparing questions with and without statements. Number of noun phrases was greatest for the combined statement and question; however that effect is much smaller than the effect on utterance length.

A possible explanation for these results is that the questions in this module were written to elicit entities, so compliant answers would generally be short. When no explicit question is provided, the range of appropriate responses is much larger. We examined a number of conversations where users were given a statement rather than a question, and confirmed that users were disclosing more and not giving longer uninformative answers. Figure 4 shows representative user responses which illustrate this behavior.

Hypo.	#tokens	# NPs	nll.
Per. Exp. (E)	4.25	1.34^{O}	21.1
Per. Opi. (O)	$\mathbf{4.61}^{S}$	$\mathbf{1.52}^{E,S}$	**22.5**
Stmt. (S)	4.15^{O}	1.27^{O}	20.6

Table 2: Effect of personal experience (E), personal opinion (O) and general statement (S) on initiative. Significant ($p < 0.05$; paired t-test) w.r.t. other hypotheses in superscript.

Q Rem.	#tokens	#NPs	#turns	nll
0% (0)	$3.77^{2,3}$	$1.25^{2,3}$	21.3	16.9^{1}
33% (1)	$3.75^{2,3}$	$1.25^{2,3}$	**21.8**	$16.2^{0,2,3}$
66% (2)	$3.91^{0,1,3}$	$1.28^{0,1,3}$	21.0	16.9^{1}
100% (3)	$\mathbf{4.01}^{0,1,2}$	$\mathbf{1.31}^{0,1,2}$	21.1	$\mathbf{17.0}^{1}$

Table 3: Effect of removing an increasing fraction of questions on initiative. Significant ($p < 0.05$; paired t-test) w.r.t. other hypotheses in superscript.

7.2 Personal Statements are reciprocated

We compare the effects of personal opinion, personal experience, and general statements (Table 2). When the statement preceded a question, there was no significant effect based on the type of statement. When a statement was presented alone, user utterances were longer in response to both personal experience and personal opinion-type statements than in response to general statements. Figure 5 gives examples of these types of statements and user responses to them. In general, users reciprocate personal opinions and experiences.

7.3 Fewer questions, greater initiative

We studied the effect of omitting questions across multiple turns (Table 3) and found that utterance length and number of noun phrases increased monotonically as the number of questions decreased. One possible explanation for this result is that our bot's questions are designed to elicit short answers and although users can give longer responses or direct the conversation to a new topic, most do not. As with utterance length and number of noun phrases, negative log-likelihood was greatest when 100% of questions were omitted. Since the question experiments were run across many of the bot's modules, we also measured their effect on number of turns, which was greatest when removing 33% of questions.

Backchan.	#tokens	#NPs	#turns	nll
0% (0)	$4.11^{1,3}$	1.41	24.0	18.9^{1}
33% (1)	$\mathbf{4.39}^{0,2}$	**1.48**	$\mathbf{25.8}^{\mathbf{2,3}}$	$\mathbf{19.8}^{0,2,3}$
66% (2)	4.20^{1}	1.42	23.6^{1}	18.8^{1}
100% (3)	4.30^{0}	1.44	23.5^{1}	19.2^{1}

Table 4: Effect of differing degrees of back-channeling on initiative. Significant ($p < 0.05$; paired t-test) w.r.t. other hypotheses in superscript.

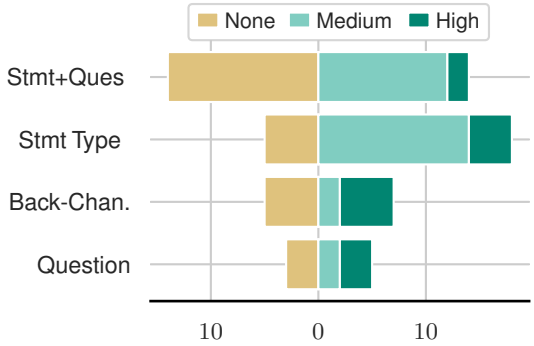

Figure 7: Each bar shows the number of responses which were initially low that converted to low, medium, or high initiative after the intervention. From top to bottom, the number of responses that stayed low initiative after each intervention is: 2, 1, 2, and 0. From a baseline of 0, none is worse than low (toward left), medium and high is better (toward right). See Table 9 for full details.

7.4 Back-channel (but not too much)

Introducing back-channeling had a non-monotonic effect. We found that all of our metrics were greatest when our bot back-channeled 33% of the time. This suggests that there is a point of diminishing returns, after which additional back-channeling leads to decreased engagement. Analyzing user conversations supported this hypothesis. We observed that when the bot always back-channels, some users either back-channel in response (e.g. "oh really?", "yep"), or continue repeating their original utterance.

7.5 To reduce low initiative, be open-ended

While the proposed strategies significantly increased the automated metrics for initiative, what was their effect on levels of initiative as defined in Section 2.3? For each experiment, we identified the most effective strategy as per automated metrics: Statement alone (Table 1), Personal Opinion (Table 2), Question Removal on 100% of turns (Table 3), and Backchanneling on 33% of turns (Table 4). For each of these strategies, we sampled 50 user utterances from turns where it had been used (in the bot's prior utterance) and a corresponding 50 turns where it had not (baseline). Three computer science graduate students without any knowledge of the strategies labelled each turn for the level of user initiative with substantial agreement (Cohen's Kappa of 0.67).

The bot's baseline responses typically asked a question to which the user would generally answer with something short and limited. As expected, when there was no intervention, users tended to take low initiative. All of our interventions replaced questions with different forms of open-ended responses. According to our definition of initiative, "low" initiative can only occur when the user is presented with a relatively small range of options, in the form of a close-ended question. For example, the question "what is your favorite animal?" restricts the range of compliant answers to the space of animals. With our interventions, very few users (at most 2/50, see Figure 7) responded with low initiative, and the rest instead chose between None, Medium, or High initiative. This is expected, since in the absence of questions, users can either direct the conversation themselves by introducing new information (Medium and High levels of initiative), or leave direction up to the bot by giving a non-informative answer such as "I'm not sure" (None level of initiative).

When measuring this effect with our annotations, we found that the bot's personal opinions lead to maximal conversion from low to medium and high initiative (Figure 7). Out of the four strategies, Statement alone performs the worst, but still increases user initiative in half the cases. Interestingly, Backchanneling on 33% of turns and Question Removal on 100% of turns converts a relatively larger fraction of low initiative responses to high initiative. These results indicate that statistically significant improvements in the formal metrics due to the best strategies also translate to a real and qualitative change in user initiative.

8 Discussion

Our goal in experimenting with initiative was to create a more human-like and engaging experience, in which the user had greater agency to direct the conversation. Our results, using both validated automated metrics and manual evaluation (see Figures 4, 5, 7, and 8), show it is possible to encourage the user to share more information by using linguistic cues. These findings suggest that when given the opportunity, many users will choose to take initiative rather than continuing passively.

Alexa Prize Likert ratings are noisy and a poor proxy for overall satisfaction (Khatri et al., 2018).

Since Alexa Prize evaluation is strictly limited to this rating, we were unable to ask more nuanced questions about initiative from the user's point of view and were unable to directly measure improvement in user experience. While we did find a slight reduction in average ratings as we omitted questions (the only experiment affecting large portions of the bot), this result is likely confounded by the particular experience of our bot. As we see in Figure 7, omitting questions leads users to take higher initiative by suggesting topics or asking questions; however, our bot was not initially designed for this behavior and it is likely that it performed worse on these new types of inputs. We studied whether changing one of the bot's utterances affected the subsequent user response; however, we did not study how effectively the bot followed up. In practice, a difficulty with successfully using this strategy remains that it is harder to produce high-quality bot follow-up turns after the user has taken initiative. In general, users appear to share more information in response to our strategies (Section 7.5), which seems likely to reflect a better experience than the brief, passive responses given previously.

Due to user privacy concerns only Alexa Prize team members could label the data in that study. While the relatively small size is indeed a limitation, we believe the qualitative conclusions to be generalizable. More generally, prior work (Reeves and Nass, 1996) suggests that humans expect chat bots to behave like humans. Despite lacking direct empirical evidence for increase in user satisfaction, we believe that more human-like turn taking will likely be satisfying to users.

Another limiting factor to our experiments is that we programmed the bot to back-channel or to omit questions at random. We expect that user preferences for initiative would vary across both individual users and particular topics and that our randomized method was much less natural than one that accounted for context. Both of these factors are likely to have inhibited our effect size. Additionally, as noted in Section 3, we are using a turn-based dialog system and therefore back-channeling cannot be done while the user speaks, but can only be attempted as a turn after they pause. This limits both its usefulness and realisticness as a strategy. Still, the fact that these methods were effective even when timing was chosen at random suggests the strength of their potential for future context-dependent approaches. All of our strategies were tested independently of each other, and we leave it for future work to test their effects in combination.

The question-answer design paradigm is common in open domain chatbot conversations, since it is an easy pattern to engineer. However, it has significant drawbacks. It restricts users' agency, potentially forcing them to discuss topics they aren't interested in. Requiring users to answer questions on every turn can also cause fatigue. In our data, we found that some users would explicitly criticize this behavior, with utterances such as "you ask too many questions." Without mixed-initiative, the bot and user cannot converse as equals. Closing the initiative gap is therefore essential to a truly natural socialbot conversation.

9 Conclusion

We found that it is possible to increase user initiative, as measured by utterance length, number of noun phrases, and response diversity, by giving linguistic cues that the user should steer the conversation. Asking fewer questions produced longer responses with more noun phrases, as did back-channeling 33% of the time. When the bot gave statements, personal ones evoked more engagement than general ones. Natural, human-like dialogue agents must share initiative with the user, and incorporating these strategies is an important step towards that goal.

Acknowledgments

We thank the SIGDIAL reviewers for their helpful feedback. We also thank our colleagues Abigail See, Caleb Chiam, Haojun Li, Mina Lee, Nguyet Minh Phu, and Omar Khattab for their support and guidance.

References

James F. Allen, Curry I. Guinn, and Eric Horvitz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23.

Jennifer Chu-Carroll and Michael K. Brown. 1998. An evidential model for tracking initiative in collaborative dialogue interactions. *User Modeling and User-Adapted Interaction*, 8(3–4):215–254.

Robin Cohen, Coralee Allaby, Christian Cumbaa, Mark Fitzgerald, Kinson Ho, Bowen Hui, Celine Latulipe, Fletcher Lu, Nancy Moussa, David Pooley, Alex Qian, and Saheem Siddiqi. 1999. What is initiative? In S. Haller, S. McRoy, and A. Kobsa, editors, *Computational Models of Mixed-Initiative Interaction*. Kluwer.

Nancy Collins and Lynn Miller. 1994. Self-disclosure and liking: A meta-analytic review. *Psychological bulletin*, 116:457–75.

Mark Core, Johanna Moore, and Claus Zinn. 2003. The role of initiative in tutorial dialogue. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.

Starkey Duncan. 1974. On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3(2):161–180.

Susan Ervin-Tripp. 1976. Is Sybil there? The structure of some American English directives. *Language in Society*, 5(1):25–66.

Jan-Gerrit Harms, Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben. 2019. Approaches for dialog management in conversational agents. *IEEE Internet Computing*, 23(2):13–22.

Peter A. Heeman, Fan Yang, and Susan E. Strayer. 2003. Control in task-oriented dialogues. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 209–212.

Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of CHI*, pages 159–166.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, Ming Cheng, Qinglang Chen, Lauren Stubel, Karthik Gopalakrishnan, Kate Bland, Raefer Gabriel, Arindam Mandal, Dilek Hakkani-Tür, Gene Hwang, Nate Michel, Eric King, and Rohit Prasad. 2018. Advancing the state of the art in open domain dialog systems through the Alexa Prize. *CoRR*, abs/1812.10757.

Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. "I hear you, I feel you": Encouraging deep self-disclosure through a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

David G. Novick and Stephen Sutton. 1997. What is mixed-initiative interaction. In *In Procedings of the AAAI Spring Symposium on Computational Models for Mixed Initiative Interaction*, pages 114–116.

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

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press, USA.

Marilyn Walker and Steve Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*, ACL '90, page 70–78, USA. Association for Computational Linguistics.

Steve Whittaker and Phil Stenton. 1988. Cues and control in expert-client dialogues. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 123–130.

Dian Yu and Zhou Yu. 2021. MIDAS: A dialog act annotation scheme for open domain HumanMachine spoken conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1103–1120, Online. Association for Computational Linguistics.

A Appendix

A.1 Data Labeling

To validate the metrics in Section 6, the authors labeled a set of 245 turns of conversation, where each turn was a pair (bot prompt, user response). The instructions used are shown in Figure 11. For a distribution across labels, see Figure 8. The same instructions were used for the task described in Section 7.5, in which three annotators labeled 400 turns of conversation.

A.2 Model Training Details

To calculate negative log-likelihood and entropy (avg. negative log-likelihood), we used a GPT-2 medium model (Radford et al., 2019), which was pre-trained on the English Webtext dataset and has 345M parameters. We fine-tuned this model on 130,000 examples of dialogue from our bot, where each example contained a single user utterance. This was divided into a training split with 91,000 examples and a validation split with 39,000 examples. During fine-tuning, we used the default hyperparameters and selected the model with the lowest negative log-likelihood loss (3.19) and had been trained for 4 epochs. The model was trained on a Titan RTX using a single GPU and 24 GB of memory. Training took 5 hours and 22 minutes.

Level	# Examples
None	84
Low	77
Medium	50
High	20
Abrupt	14

Figure 8: Label distribution for validation dataset

Experiment	# None	# Low	# Med.	# High	# Abrupt
Stmt + Ques.	22 (+14)	2 (-29)	16 (+12)	5 (+2)	4 (+0)
Stmt Type	15 (+5)	1 (-27)	21 (+14)	6 (+4)	7 (+4)
Back-Chan.	16 (+5)	2 (-15)	21 (+2)	6 (+5)	5 (+3)
Question	18 (+3)	0 (-6)	17 (+2)	5 (+3)	4 (+2)

Figure 9: Each column indicates the number of responses at each level after the intervention. Values in parentheses indicate the difference in number of responses from turns without the intervention.

I don't know, no, yeah, yes, okay, none, uh, cool, what, me too, don't know, not sure, I'm not sure, right, wow

Figure 10: The 15 most common non-informative user responses

Level	Includes	Examples
None	Yes/No responses to binary questions Uninformative answers	**Bot:** have you seen any good movies lately? **User:** Not really. **Bot:** I was wondering, do you like dogs? **User:** Yes. **Bot:** What's your favorite animal? **User:** I don't know.
Low	Responses to closed-ended questions without extra information	**Bot:** What's your favorite color? **User:** Blue. **Bot:** What's your favorite animal? **User:** I like dogs. **Bot:** How was your day? **User:** Pretty good.
Medium	Responses to open-ended questions Responses that share unprompted information	**Bot:** What do you want to talk about? **User:** Dogs. **Bot:** What's your favorite animal? **User:** I love dogs. I want to get a golden retriever. **Bot:** How was your day? **User:** Pretty good. I went for a walk around my neighborhood.
High	Questions Commanding/requesting a topic naturally	**Bot:** What's your favorite color? **User:** Blue. What about you? **Bot:** What's your favorite animal? **User:** I love lions. I want to go to Africa so I can see them. Let's talk about Africa. **Bot:** How was your day? **User:** Pretty good. Tell me about your day.
Abrupt	Commanding/requesting a topic unnaturally Complaints	**Bot:** What's your favorite color? **User:** Let's talk about dogs. **Bot:** What's your favorite animal? **User:** You're boring. **Bot:** How was your day? **User:** I don't want to talk about that.

Figure 11: Instructions used to label validation examples