An Empirical Study of Self-Disclosure in Spoken Dialogue Systems

Abhilasha Ravichander
Language Technologies Institute
Carnegie Mellon University
aravicha@cs.cmu.edu

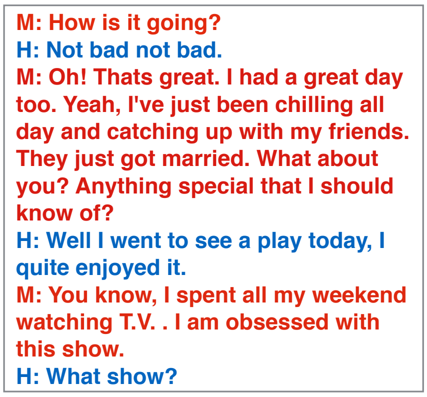
Alan Black
Language Technologies Institute
Carnegie Mellon University
awb@cs.cmu.edu

Abstract

Self-disclosure is a key social strategy employed in conversation to build relations and increase conversational depth. It has been heavily studied in psychology and linguistic literature, particularly for its ability to induce self-disclosure from the recipient, a phenomena known as reciprocity. However, we know little about how self-disclosure manifests in conversation with automated dialog systems, especially as any self-disclosure on the part of a dialog system is patently disingenuous. In this work, we run a large-scale quantitative analysis on the effect of self-disclosure by analyzing interactions between real-world users and a spoken dialog system in the context of social conversation. We find that indicators of reciprocity occur even in human-machine dialog, with far-reaching implications for chatbots in a variety of domains including education, negotiation and social dialog.

1 Introduction

Humans employ different strategies during a conversation in pursuit of their social goals (Tracy and Coupland, 1990). The contributions to a conversation can be categorized as those which serve *propositional* functions by adding new information to the dialog, those which serve *interactional* functions by driving the interaction and those which serve *interpersonal* functions, by building up the relationship between the involved parties. When fulfilling interpersonal functions, people either consciously or sub-consciously employ social conversational strategies in order to connect and build relationships with each other (Laurenceau et al., 1998; Won-Doornink, 1985). This feeling of rapport, of connecting and having common ground with another human being is one of the fundamental aspects of good human conversation. Maintaining conversational harmony has shown to be effective in several domains such as education (Ogan et al., 2012; Sinha and Cassell, 2015a,b; Frisby and Martin, 2010; Zhao et al., 2016) and negotiation (Drolet and Morris, 2000; Nadler, 2003, 2004).



M: How is it going?
H: Not bad not bad.
M: Oh! Thats great. I had a great day too. Yeah, I've just been chilling all day and catching up with my friends. They just got married. What about you? Anything special that I should know of?
H: Well I went to see a play today, I quite enjoyed it.
M: You know, I spent all my weekend watching T.V. . I am obsessed with this show.
H: What show?

Figure 1: Excerpt dialog from conversation between a user and our dialog agent[1]. H represents user utterance and M represents machine dialog.

Self-disclosure is the conversational act of disclosing information about oneself to others. We consider the definition of self-disclosure within the theoretical framework of social penetration theory, where it is defined as the *voluntary* sharing of opinions, thoughts, beliefs, experiences, preferences, values and personal history (Altman and Taylor, 1973). The effect of self-disclosure has been well-studied in the psychology community, in particular it's ability to induce *reciprocity* in dyadic interaction (Jourard, 1971; Derlega et al.,

[1] Real interaction data withheld for confidentiality. Conversation data shown here is not real interaction data but follows similar patterns.

Proceedings of the SIGDIAL 2018 Conference, pages 253–263,
Melbourne, Australia, 12-14 July 2018. ©2018 Association for Computational Linguistics

1973). Several studies have shown that self-disclosure reciprocity characterizes initial social interactions between people (Ehrlich and Graeven, 1971; Sprecher and Hendrick, 2004) and further, that *disclosure promotes disclosure* (Dindia et al., 2002).

This brings us to a natural question: how does such behavior manifest itself in interactions with dialog systems? A subtle but crucial aspect is that humans are aware that machines do not have feelings or experiences of their own, so any attempt at self-disclosure on the part of the machine is inherently disingenuous. However, Nass et al. (1994) suggests that humans tend to view computers as social actors, and interact with them in much the same way they do with humans. Disclosure reciprocity in such a setting would have far-reaching implications for dialog systems which aim to elicit information from the user in order to offer more personalized experiences for example, or to better achieve task completion (Bickmore and Cassell, 2001; Bickmore and Picard, 2005; Goldstein and Benassi, 1994; Lee and Choi, 2017).

In this work, we study this phenomena by building an open-domain chatbot (§3) which engages in social conversation with hundreds of Amazon Alexa users (Figure 1.), and gains insights into two aspects of human-machine self-disclosure. First, self-disclosure by the dialog agent is strongly correlated with instances of self-disclosure by the user indicating disclosure reciprocity in interactions with spoken dialog systems (§4.1). Second, initial self-disclosure by the user can characterize user behavior throughout the conversation (§4.2). We additionally study the effect of self-disclosure and likability, but find no reliable linear relationship with the amount of self-disclosure in the conversation (§4.3). To the best of our knowledge, this work is the first large-scale study of reciprocity and self-disclosure between users in the real world and spoken dialog systems.

2 Background

Self-disclosure as a social phenomena is the act of revealing information about oneself to others. It has been of particular interest to study what factors makes humans self-disclose (Miller et al., 1983; Dindia and Allen, 1992; Hill and Stull, 1987; Buhrmester and Prager, 1995; Stokes, 1987; Qian and Scott, 2007; Jourard and Friedman, 1970; Ko and Kuo, 2009), how do they do it (Chen, 1995; Greene et al., 2006; Chelune, 1975; Sprecher and Hendrick, 2004) and what are the effects of self-disclosing (Gibbs et al., 2006; Mazer et al., 2009; Forest and Wood, 2012; Turner et al., 2007; Knox et al., 1997; Vittengl and Holt, 2000).

One such effect is disclosure reciprocity, which has been shown to be one of the most significant effects of self-disclosure (Jourard, 1971). Reciprocity is the phenomenon by which self-disclosure by one participant in a dyadic social interaction results in self-disclosure from the other participant in response. A substantial amount of research has shown that when one party self-discloses, the other party is much more likely to self-disclose (Jourard, 1971; Jourard and Friedman, 1970; Dindia et al., 2002; Derlega et al., 1973). While the exact cause of this phenomena is not known, it has been suggested that self-disclosure can be viewed as a social exchange, where the party receiving self-disclosure feels obligated to self-disclose in return (Archer, 1979), or as a social conversational norm (Derlega et al., 1993), or from the point of view of social trust-attraction (Vittengl and Holt, 2000) where people self-disclose to people who disclose to them, as they consider self-disclosure to be a sign of trust and liking. Additionally, Sprecher and Hendrick (2004) find that people who consider themselves to be high self-disclosers are likely to be much better at eliciting self-disclosure as well. Derlega et al. (1973) observe that self-disclosure is a positive function of self-disclosure received, regardless of liking for the initial discloser. Mikulincer and Nachshon (1991) analyze personality types and self-disclosure, and find that secure people are more likely to both self-disclose and reciprocate self-disclosure. Cozby (1972) study the relationship between disclosure and liking and suggest that this relationship is not linear. In this work, we attempt to combine these perspectives to gain insights into the nature of self-disclosure in human-machine dialog.

3 Identifying Self-Disclosure

3.1 Coding Self Disclosure

In this work, we consider the definition of self-disclosure within the theoretical framework of social penetration theory (Altman and Taylor, 1973) where it is defined to be the *voluntary* sharing of information which could include amongst other things one's personal history, thoughts, opinions,

Self-Disclosure	No Self-Disclosure
1) M: Good to hear! Is it anything special today? H: Nothing much, I am just going down to the lake with my dogs today 2) M: Have you seen the movie Arrival? H: Oh my god, yes i have it's my favorite movie	1) M: Good to hear! Is it anything special today? H: Not really 2) M: Have you seen the movie Arrival? H: Yeah I did I did see Arrival

Figure 2: Examples of self-disclosing user responses as well as responses when the user does not self-disclose [2].

beliefs, feelings, preferences, attitudes, aspirations, likes, dislikes and favorites. In a human-machine context, we define self-disclosure as the conversational act of revealing aspects of oneself voluntarily, which would otherwise not be possible to be known by the dialog system. A general rule-of-thumb we follow is, self-disclosure is proportional to the amount of extraneous information that is added to a conversation. For example, we do not identify a direct response to a question as self-disclosure as it is not strictly voluntary. We show examples of our definition of human self-disclosure and non-disclosure in the context of our dialog system in Figure. 2.

3.2 Dataset Preparation

The data for this study was collected by having users from the real-world interact with our open-domain dialog agent. The dialog agent was hosted on Amazon Alexa devices as part of the AlexaPrize competition (Ram et al., 2018) and was one of sixteen socialbots that could be invoked by any user within the United States through the command '*Let's chat!*'. The users that interacted with our socialbot were randomly chosen, and did not know which of the sixteen systems they were interacting with. Users who interacted with our bot over a span of three days (N=1507) were randomly assigned to two groups: one received a bot that self-disclosed at high depth from the beginning of the conversation while the other group interacted with a socialbot that self-disclosed only later about superficial topics like movies and TV shows. At the end, both socialbots engaged in free-form conversation with the user, where the initiative of the interaction was on the user and both bots were free to self-disclose at any depth. The users were also free to end the interaction at any time, and thus had no motivation for continuing the conversation besides their own entertainment. To control the direction of the conversation and bot utterance, we utilize a finite state transducer-based dialog system that chats with the user about movies and TV shows, as well as plays games and supports open-domain conversation (Prabhumoye et al., 2017). State transitions are decided based on sentiment analysis of user utterances, in order to gauge interest in a particular topic. Initially the dialog system takes initiative in the conversation and steers the topic of discussion, however later there is a handoff to the user whereby the user can determine the focus of the conversation. In this way, the socialbot leads the user through the following topics, conditioned on user interest as shown in Figure 3:

Greeting : In this phase, our dialog agent greets the user and asks them about their day. The bot which performs high self-disclosure initially also responds with information about it's day and a personal anecdote.

TV Shows: The next phase involves chit chat about popular TV shows. The dialog agent asks the user if they are an enthusiast of a recent popular TV show and moves on to the next phase of the conversation if they aren't.

Movie: In this phase, the dialog agent attempts to engage the user in conversations about movies, asking them if they have seen any of the recent ones.

Word Game: In this phase, the dialog agent requests the user to play a word game. Participation in the game is completely optional and the user can move on to the next phase by stating that they do not wish to play.

CQA: The last phase supports uninhibited free-form conversation. The initiative of the exchange is now on the user and conversation is stateless. The dialog system response is determined by a retrieval model. For each utterance, the socialbot attempts to retrieve the most relevant response from the Yahoo L6 dataset (yl6, 2017), a dataset containing approximately 4 million questions and

[2] Not real interaction data, however very similar to actual utterances found in the interaction data

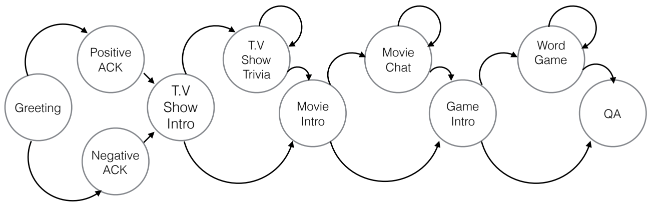


Figure 3: Topic FST for Conversation

their corresponding answers from the Community Question-Answering (CQA) website, Yahoo Answers [3].

The users were then allowed to rate the interaction on a scale of 1-5, based on the question *'Would you interact with this socialbot again?'*. 319 users rated the socialbot (Group 1) and 1507 users interacted with our system in total (Group 2). Following this, to preserve confidentiality of the interaction data, one annotator annotated all turns of conversation from Group 1 for self-disclosure. Annotator reliability was determined by calculating inter-annotator agreement from three external annotators on a carefully prepared anonymized subset of the data amounting to 62 interactions comprising of over 816 turns. The Fleiss' kappa from the four annotators was 63.8, indicating substantial agreement. Atleast two of three annotators agreed on 93.6% of the reference annotations. The full dataset contains a total of 319 conversations, spanning 10751 conversational turns. Out of the 5216 human dialog utterances, 13.8% featured some form of self-disclosure.

Since our agent is a spoken dialog system in the real world there is some amount of noise in the dataset caused due to ASR errors. To estimate this, we randomly sample 100 utterances from the dataset and annotate these utterances for whether they contained an ASR mistake, and if the sentence meaning was still apparent either from context or from the utterance itself. We find that at least one ASR error occurs in 13% of user utterances, but 46.1% of utterances with ASR mistakes can still be understood. Since our dialog agent relies on sentiment-based FST transitions during the initial stages of the conversation, we also analyze the rate of false transitions in the data. We randomly sample 100 utterances from across choice points of all conversations and find that 11% of them consisted of incorrect responses, either due to mistakes in sentiment analysis or due to nuance in the user utterances which rendered a response from the dialog agent unusable. Finally, we analyze how many users had multiple interactions with our dialog agent during the course of our study. This is relevant as user behavior during a second interaction with the system might differ from initial interaction. Users are identifiable only by an anonymized hash key provided by Amazon along with the conversation data. We find that out of 316 users who interacted with our dialog agent and left a rating, only 3 interacted with our agent twice and none of them interacted with our agent more than two times, largely allowing us to disregard this effect.

3.3 Feature Space Design

We utilize the annotations of 319 conversations to train and evaluate a Machine Learning model to identify user self-disclosure. We categorize the features for this model at two levels, *utterance-level* features wherein the user utterance is taken standalone and analyzed for self-disclosure and *conversational-level* features which consider the utterance in context of the current conversation.

3.3.1 Utterance Features

This represents a class of features that only consider the current utterance. These include-

1. **Bag-of-words Features** TF-IDF features from the user utterance.

2. **Linguistic Style Features** This class of features attempts to characterize the linguistic style of user utterances, including lexical choices that might be indicative of self-disclosure (Doell, 2013). These include- i) Length of the user utterance, ii) Presence of negation words, iii) Part-of-speech tags such as nouns and adjectives in the user utterance in order to represent users revealing emotion or discussing topics, iv) Presence of filler words in utterance, v) Number

[3] answers.yahoo.com

of named entities in the utterance, vi) Gazeteer features based on common responses to questions asked by our dialog system, indicative of conversational responses as well as strongly positive, negative or neutral responses [4].

3. **LIWC Features** i) Studies have shown (Sparrevohn and Rapee, 2009) that people who self-disclose tend to use words that reveal strong emotion. Thus, we include features to represent words from affect relevant categories of LIWC (Pennebaker et al., 2015), such as anger, anxiety, sadness, positive emotion or negative emotion, ii) number of personal pronouns, first person singular pronouns, first person plural pronouns, second person pronouns, third person plural pronouns, third person singular pronouns, iii) Additionally, users self-disclosing incidents from their personal lives tend to discuss their social settings. Thus, we use relationship words related to the family and friends categories from LIWC.

3.3.2 Conversation Features

These features are broadly based on dialog structure or the language-based features from local conversational context. These include i) TF-IDF features from the user utterance concatenated with the bot utterance[5], to help capture the difference between direct responses to questions and voluntary self-disclosure, ii) dialog system self-disclosing in previous turn, iii) dialog system asking a question in the previous turn, iv) Amount of word overlap with previous machine utterance, which is defined as the number of words that overlap with the previous dialog system utterance normalized by the length of the dialog system utterance, v) Number of content words[6] that overlap with previous machine utterance.

Model	**Accuracy**	**Precision**	**Recall**	**F1**
First Person	86.6%	68.0%	6.0%	10.9%
Utterance Features	89.8%	69.8%	46.5%	55.5%
Utterance + Conversation Features	**91.7%**	**74.4%**	**60.5%**	**66.67%**

Table 1: Classification performance(%) of models at identifying user utterances to contain self-disclosure.

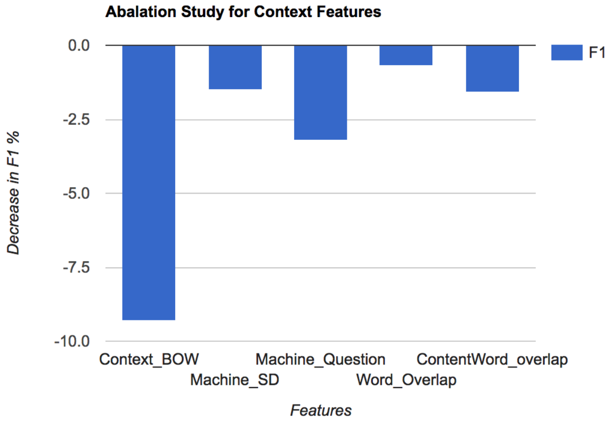


Figure 4: Ablation Study for Conversation Features.

3.4 Results of Identification

The combination of the three categories of features results in a 234-dimensional vector which acts as input to an SVM with a linear kernel. We utilize truncated SVD with 100 components for dimensionality reduction of all bag-of-words based feature classes. We compare against two baselines, the first is a baseline consisting of only personal voiced features (including all LIWC features) and the second attempts to classify self-disclosure independent of dialog context (only conditioned on the current user utterance). We perform 10-fold cross validation and describe our results in Table. 1. We observe that considering user utterances in context of the conversation considerably improves our ability to predict self-disclosure. To perform more detailed error analysis on a larger test set, we randomly sample 1044 utterances from 5216 utterances to be a held-out test set. This test set consists of 134 utterances of self-disclosure. Our classifier achieves an accuracy of 93.4% at

[4] Includes phrases such as ”I’m fine”, ”I’m ok”, ”I’m good”, ”I’m doing ok”, ”I’m doing good”, ”how are you” for conversational responses, ”delightful”, ”favorite”, ”amazing”, ”awesome”, ”fantastic”, ”brilliant”, ”the best”, ”really great” etc. for strongly positive, ”boring”, ”tired”, ”bored”, ”sad”, ”lonely”, ”disgusting”, ”hate”,”awful” etc. for strongly negative and ”rain”, ”summer”, ”winter”, ”cold”, ”wind” etc. for strongly neutral (as users tend to discuss weather while making small talk).

[5] Each word of the bot utterance is encapsulated within a <bot></bot> tag

[6] where we determine content words following the usual definition of nouns, main verbs, adjectives and adverbs.

recognizing self-disclosure on this test set, with a F1-score of 72.7% (Precision: 77.3%, Recall: 68.6%). The test distribution contains 12.8% examples of self-disclosure and 87.2% examples of no disclosure. We further perform an ablation study of each dialog-context feature as shown in Figure 4. We observe that considering the word in the context of the machine utterance is most helpful in identifying self-disclosure, indicating possibly that it helps us capture the notion of self-disclosure being a voluntary phenomena whereby the user reveals information about himself or herself, by separating instances of direct answers to questions from turns where users disclose more than what is asked. We next conduct a careful manual error analysis of the mistakes made by our classifier, in an attempt to identify what cases are particularly hard or ambiguous. We observe that 85% of user turns which our model wrongly labeled as containing self-disclosure had personal pronouns, suggesting that our model considers these as a very strong signal for self-disclosure. However many of these utterances were in fact direct responses to questions, or questions to the bot itself prefaced with a personal pronoun, and thus not really instances of self-disclosure. 25.9% of the mistakes were not well-formed or meaningful sentences, possibly due to ASR errors, speech disfluencies or user phrasing. We also examine the user turns our model failed to predict as being self-disclosure. 19.5% of these mistakes were not well-formed sentences and 12.1% were statements about the bots performance. A further 21.9% of errors contained rare words which might not have been seen before in the training data along with an absence of the linguistic markers of self-disclosure identified by us (for example, *M: Anything special today? H: Really wanna grab a smoke*). In the future, real world knowledge and a larger amount of training data might help mitigate some of these error classes.

4 Effect of Self-Disclosure

4.1 Reciprocity

We analyze common markers of reciprocity (Jourard and Jaffe, 1970; Harper and Harper, 2006), such as the usage of personal pronouns, word overlap with the previous sentence (normalized by length of previous utterance) and average user utterance length between two groups of users- ones who were shown a bot that self-disclosed initially and a bot which only self-disclosed later (Table 2.).

Marker	Mean SD	Mean Ctrl
Word Overlap*	0.0352	0.0226
First Person Pronouns*	0.84	0.57
Avg. Noun Mentions*	2.00	1.49
Avg. Adjective Mentions	0.55	0.47
Avg. User Utt. Length	4.428	3.983

Table 2: Various effects of conversation with a dialog system that self-discloses right off-the-bat and with a control dialog system that only self-discloses later. * indicates p<0.05 after Bonferroni correction.

Group	No Machine SD	With Machine SD
Rated	10.5%	**24.3%**
All	7.4%	**21.6%**

Table 3: % of turns with Human Self-Disclosure following Machine Self-Disclosure/Non-Disclosure.

Within the data which consists of only rated conversations, we observe how many turns where the machine self-disclosed were also met with human self-disclosure ("Rated" in Table. 3). We then tag all user utterances [7] with our SVM classifier as either being instances of self-disclosure or not being instances of self disclosure ("All" in Table. 3). We find that 10.6% of all user utterances contain self disclosure, and 21.6% of machine utterances that contained self-disclosure were followed by a human utterance that contained self-disclosure, compared to the 7.4% of cases where a user self-disclosed without the machine self-disclosing ($p < 0.05$). These results are shown in Table. 3.

Next, we observe the utterance after initial self-disclosure for a group where the socialbot self-discloses compared to the equivalent dialog turn for a group where the bot doesn't self-disclose, to analyze if self-disclosure has immediate effects. These results are shown in Table. 4. We observe that when the bot self-discloses, the user self-discloses in response in 56.5% of all cases. However if the bot does not self-disclose and asks the same question, the user self discloses only in 35.5% of all cases ($p < 0.0001$). Our findings suggest that it is possible user behavior is affected by

[7] from 811 conversations of length greater than three turns.

Group	No Machine SD	With Machine SD
Rated	44.4%	**62.6%**
All	35.5%	**56.5%**

Table 4: % of turns with Human Self-Disclosure in turns immediately following equivalent initial self-disclosing/non-disclosing turn of machine.

the self-disclosing behavior of our dialog agent, and that such an effect can be seen immediately.

4.2 Initial Self-Disclosure and User behavior

We next examine conversation-wide characteristics and self-disclosure patterns of users based on their initial self-disclosing behavior.

Are Conversations With Initial Self-Disclosure Longer? We analyze whether whether initial occurrences of user self-disclosure lead to users prolonging the conversation by examining average conversational length for two groups of users : those who decided to self-disclose at the very beginning of the conversation itself and those who didn't. We find that users who self-disclose initially tend to have significantly longer conversation than users who do not ($p<0.05$), with an average conversational length of 37.19 turns compared to an average of 32.4 turns for users who chose not to self-disclose.

Does not self-disclosing initially imply reduced self-disclosure throughout the conversation? We next examine the hypothesis that users who do not self-disclose initially tend to self-disclose less throughout. This is based on the notion of openness and guardedness in personality (Stokes, 1987; Sermat and Smyth, 1973) indicating that some individuals are more likely to self-disclose than others. For this study, we do not consider interactions involving the word game as it prolongs the conversation without giving opportunities for self-disclosure. We examine to what extent do individuals who refuse to self-disclose initially, self-disclose later in the conversation compared to users who self-disclose from the beginning of the conversation itself. We find that on average, users who do not choose to self-disclose initially are *significantly* less likely to self-disclose ($p<0.05$) even later on in the conversation, only revealing information in 9% of their turns as compared to the 24.6% of turns of other users.

Do users who choose not to self-disclose initially exhibit less interest in following machine interests? To analyze openness to conversation, we invite users to play a long-winded word game with the dialog system. We analyze how much self-disclosure correlates with willingness to play the game and length of game playing. We find that on average users who self-disclose initially are also significantly more open to game-playing than those who don't ($p<0.05$), playing on average 4.75 turns of the game compared to an average gameplay of 3.16 turns by other users. They are also significantly more likely to attempt to play the game ($p<0.05$), with 34.7% of self-disclosing users attempting to play the game and only 25.1% of non-disclosing users attempting to do so.

4.3 Does Self-Disclosure Increase Likability

Motivated by Cozby (1972), we attempt to analyze whether self-disclosure increases likability in human-machine interaction. We utilize the user ratings based on the question *'Would you talk to this socialbot again'* as a proxy for likability of the dialog agent, and examine whether conversations where the user self-disclosed often were given higher ratings than ones where they didn't. We find that there is negligible correlation in general between user ratings and the amount of self-disclosure (pearson's $r = 0.01$). We then examine the differences in user ratings between the top 20% and bottom 20% of self-disclosing conversations, once more excluding interactions with the game. We observe that while more self-disclosing conversations get higher ratings in general, the results are not statistically significant (average rating of conversations with higher self-disclosure is 3.14 compared to 3.13 for conversations with lesser self-disclosure). Lastly, we analyze the effect of reciprocity and self-disclosure, by analyzing the ratings of users who self-disclosed in response to bot disclosure but find no significant difference in the ratings of such users (3.34 to 3.27). Thus we are unable to find any conclusive linear relationship between self-disclosure and likability.

5 Discussion and Related Work

There has been significant prior interest in computationally analyzing various forms of self-disclosure online (Yang et al., 2017; Wang et al., 2016; Stutzman et al., 2012; Yin et al., 2016;

Bak et al., 2014; De Choudhury and De, 2014). Bickmore et al. (2009) study the effect of machine 'backstories' in dialog, and find that users rate their interactions to be more enjoyable when the dialog system has a backstory. Zhao et al. (2016) identify self-disclosure in peer tutoring between humans. Han et al. (2015); Meguro et al. (2010) identify self-disclosure as a user intention in a natural language understanding system. Oscar J. Romero (2017) use self-disclosure as one strategy amongst others to build a socially-aware conversational agent. Higashinaka et al. (2008) study if users self-disclose on topics they like rather than ones they don't, with a focus on text-based chat rather than spoken dialog. Similarly, Lee and Choi (2017) study the relation between self-disclosure and liking for a movie recommendation system, using a Wizard-of-Oz approach instead of constructing a dialog agent. Perhaps closest to our work is the work of Moon (2000), which studies the phenomena of reciprocity in human-machine self-disclosure. However, this phenomena is not studied for dialog, and similar to previous work, relies on a text-based series of interview questions.

In this work, we are interested in realizing self-disclosure in a real-time, large-scale spoken dialogue system. We depart from previous work in three main ways. First, we have the opportunity of deploying a dialog agent in the wild, and studying hundreds of interactions with real users in US households. Second, we study reciprocity of self-disclosure in human-machine dialog, and find markers of reciprocity even in conversations with a dialog agent. Third, we characterize users by their initial self-disclosing behavior and study conversation-level behavioral differences. We believe this work to be a step towards better understanding the effect of dialog agents deployed in the real-world employing self-disclosure as a social strategy, as well as better understanding the implications of self-disclosing *user* behavior with dialog agents.

We acknowledge limitations of our current approach. In this work, our definition of self-disclosure is binary. A more nuanced version that considers both magnitude and valence of self-disclosure would open up several further research directions, such as analyzing reciprocity matching in depth of disclosure and analyzing user behavior based on the valence of disclosure. It would also be interesting to analyze how agent behavior can significantly influence non-disclosing users, as our results find that users who do not initially self-disclose continue to self-disclose at reduced levels throughout the conversation. Another immediate research direction would be to study the effect of other social conversational strategies such as praise (Fogg and Nass, 1997; Zhao et al., 2016) at a large scale in spoken-dialog systems. In the future, one could imagine dialog agents that reason over both social strategies and their magnitude, conditioned on user behavior, in service of their conversational goals.

6 Conclusion

In this work, we empirically study the effect of self-disclosure in a large-scale experiment involving real-world users of Amazon Alexa. We find that indicators of reciprocity occur even in conversations with dialog systems, and that user behavior can be characterized by self-disclosure patterns in the initial stages of the conversation. We hope that these findings inspire more user-centric research in dialog systems, with an emphasis on dialog agents that attempt to build a relationship and maintain rapport with the user when eliciting information.

Acknowledgements

This work has partially been supported by the National Science Foundation under Grant No. CNS 13-30596. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF, or the US Government. The authors are grateful to the CMU Magnus team for their hard work through the AlexaPrize competition, with special thanks to Shrimai Prabhumoye and Chaitanya Malaviya for helping with this effort. The authors would like to express their gratitude to Diyi Yang, Alexander Rudnicky and Dan Jurafsky for interesting discussions related to this work. The authors would like to especially thank Shruti Rijhwani, Shivani Poddar and Aditya Potukuchi for their time and support through this effort. Finally, the authors are immensely grateful to the Amazon Alexa team for facilitating universities to do dialog research with real user data, through the Alexa Prize competition, as well as to all the Amazon Alexa users who were willing to interact with our system.

References

2017. Yahoo Webscope. (2007). L6 - Yahoo! Answers Comprehensive Questions and Answers version (1.0). `https://webscope.sandbox.yahoo.com/catalog.php`. [Online; accessed 15-August-2017].

Irwin Altman and Dalmas Taylor. 1973. Social penetration theory. *New York: Holt, Rinehart &\Mnston* .

R.L Archer. 1979. Anatomical and psychological sex differences. *Self-disclosure: Origins, Patterns, and Implications of Openness in Interpersonal Relationships* .

JinYeong Bak, Chin-Yew Lin, and Alice Oh. 2014. Self-disclosure topic model for classifying and analyzing twitter conversations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 1986–1996.

Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, pages 396–403.

Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2009. Engagement vs. deceit: Virtual humans with human autobiographies. In *International Workshop on Intelligent Virtual Agents*. Springer, pages 6–19.

Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12(2):293–327.

Duane Buhrmester and Karen Prager. 1995. Patterns and functions of self-disclosure during childhood and adolescence. .

Gordon J Chelune. 1975. Self-disclosure: An elaboration of its basic dimensions. *Psychological Reports* 36(1):79–85.

Guo-Ming Chen. 1995. Differences in self-disclosure patterns among americans versus chinese: A comparative study. *Journal of Cross-Cultural Psychology* 26(1):84–91.

Paul C Cozby. 1972. Self-disclosure, reciprocity and liking. *Sociometry* pages 151–160.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity.

Valerian J Derlega, Marian Sue Harris, and Alan L Chaikin. 1973. Self-disclosure reciprocity, liking and the deviant. *Journal of Experimental Social Psychology* 9(4):277–284.

VJ Derlega, S Metts, S Petronio, and ST Margulis. 1993. Sage series on close relationships. self-disclosure.

Kathryn Dindia, M Allen, R Preiss, B Gayle, and N Burrell. 2002. Self-disclosure research: Knowledge through meta-analysis. *Interpersonal communication research: Advances through meta-analysis* pages 169–185.

Kathryn Dindia and Mike Allen. 1992. Sex differences in self-disclosure: A meta-analysis. *Psychological bulletin* 112(1):106.

Kelly Doell. 2013. The word feel as an indicator of enacted social support in personal relationships. *International Journal of Psychological Studies* 5(4):107.

Aimee L Drolet and Michael W Morris. 2000. Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology* 36(1):26–50.

Howard J Ehrlich and David B Graeven. 1971. Reciprocal self-disclosure in a dyad. *Journal of Experimental Social Psychology* 7(4):389–400.

Brian J Fogg and Clifford Nass. 1997. Silicon sycophants: The effects of computers that flatter. *International journal of human-computer studies* 46(5):551–561.

Amanda L Forest and Joanne V Wood. 2012. When social networking is not working: Individuals with low self-esteem recognize but do not reap the benefits of self-disclosure on facebook. *Psychological science* 23(3):295–302.

Brandi N Frisby and Matthew M Martin. 2010. Instructor–student and student–student rapport in the classroom. *Communication Education* 59(2):146–164.

Jennifer L Gibbs, Nicole B Ellison, and Rebecca D Heino. 2006. Self-presentation in online personals: The role of anticipated future interaction, self-disclosure, and perceived success in internet dating. *Communication Research* 33(2):152–177.

Gary S Goldstein and Victor A Benassi. 1994. The relation between teacher self-disclosure and student classroom participation. *Teaching of Psychology* 21(4):212–217.

Kathryn Greene, Valerian J Derlega, and Alicia Mathews. 2006. Self-disclosure in personal relationships. *The Cambridge handbook of personal relationships* pages 409–427.

Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. pages 129–133.

Vernon B Harper and Erika J Harper. 2006. Understanding student self-disclosure typology through blogging. *The Qualitative Report* 11(2):251–261.

Ryuichiro Higashinaka, Kohji Dohsaka, and Hideki Isozaki. 2008. Effects of self-disclosure and empathy in human-computer dialogue. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*. IEEE, pages 109–112.

Charles T Hill and Donald E Stull. 1987. Gender and self-disclosure. In *Self-Disclosure*, Springer, pages 81–100.

Sidney M Jourard. 1971. Self-disclosure: An experimental analysis of the transparent self. .

Sidney M Jourard and Robert Friedman. 1970. Experimenter-subject" distance" and self-disclosure. *Journal of Personality and Social Psychology* 15(3):278.

Sidney M Jourard and Peggy E Jaffe. 1970. Influence of an interviewer's disclosure on the self-disclosing behavior of interviewees. *Journal of Counseling Psychology* 17(3):252.

Sarah Knox, Shirley A Hess, David A Petersen, and Clara E Hill. 1997. A qualitative analysis of client perceptions of the effects of helpful therapist self-disclosure in long-term therapy. *Journal of counseling psychology* 44(3):274.

Hsiu-Chia Ko and Feng-Yang Kuo. 2009. Can blogging enhance subjective well-being through self-disclosure? *CyberPsychology & Behavior* 12(1):75–79.

Jean-Philippe Laurenceau, Lisa Feldman Barrett, and Paula R Pietromonaco. 1998. Intimacy as an interpersonal process: The importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges. *Journal of personality and social psychology* 74(5):1238.

SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103:95–105.

Joseph P Mazer, Richard E Murphy, and Cheri J Simonds. 2009. The effects of teacher self-disclosure via facebook on teacher credibility. *Learning, Media and technology* 34(2):175–183.

Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. 2010. Controlling listening-oriented dialogue using partially observable markov decision processes. In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, pages 761–769.

Mario Mikulincer and Orna Nachshon. 1991. Attachment styles and patterns of self-disclosure. *Journal of Personality and Social Psychology* 61(2):321.

Lynn C Miller, John H Berg, and Richard L Archer. 1983. Openers: Individuals who elicit intimate self-disclosure. *Journal of Personality and Social Psychology* 44(6):1234.

Youngme Moon. 2000. Intimate exchanges: Using computers to elicit self-disclosure from consumers. *Journal of Consumer Research* 26(4):323–339.

Janice Nadler. 2003. Rapport in negotiation and conflict resolution. *Marq. L. Rev.* 87:875.

Janice Nadler. 2004. Rapport in legal negotiation: How small talk can facilitate e-mail dealmaking. *Harv. Negot. L. Rev.* 9:223.

Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, pages 72–78.

Amy Ogan, Samantha L Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. 2012. Rudeness and rapport: Insults and learning gains in peer tutoring. Springer.

Ran Zhao Justine Cassell Oscar J. Romero. 2017. Cognitive-inspired conversational-strategy reasoner for socially-aware agents. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. pages 3807–3813. https://doi.org/10.24963/ijcai.2017/532.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Shrimai Prabhumoye, Fadi Botros, Khyathi Chandu, Samridhi Choudhary, Esha Keni, Chaitanya Malaviya, Thomas Manzini, Rama Pasumarthi, Shivani Poddar, Abhilasha Ravichander, et al. 2017. Building cmu magnus from user feedback. *Alexa Prize Proceedings* .

Hua Qian and Craig R Scott. 2007. Anonymity and self-disclosure on weblogs. *Journal of Computer-Mediated Communication* 12(4):1428–1451.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604* .

Vello Sermat and Michael Smyth. 1973. Content analysis of verbal communication in the development of relationship: Conditions influencing self-disclosure. *Journal of Personality and Social Psychology* 26(3):332.

Tanmay Sinha and Justine Cassell. 2015a. Fine-grained analyses of interpersonal processes and their effect on learning. In *International Conference on Artificial Intelligence in Education*. Springer, pages 781–785.

Tanmay Sinha and Justine Cassell. 2015b. We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*. ACM, pages 13–20.

Roslyn M Sparrevohn and Ronald M Rapee. 2009. Self-disclosure, emotional expression and intimacy within romantic relationships of people with social phobia. *Behaviour Research and Therapy* 47(12):1074–1078.

Susan Sprecher and Susan S Hendrick. 2004. Self-disclosure in intimate relationships: Associations with individual and relationship characteristics over time. *Journal of Social and Clinical Psychology* 23(6):857.

Joseph P Stokes. 1987. The relation of loneliness and self-disclosure. In *Self-disclosure*, Springer, pages 175–201.

Frederic Stutzman, Jessica Vitak, Nicole B Ellison, Rebecca Gray, and Cliff Lampe. 2012. Privacy in interaction: Exploring disclosure and social capital in facebook.

Karen Tracy and Nikolas Coupland. 1990. Multiple goals in discourse: An overview of issues. *Journal of Language and Social Psychology* 9(1-2):1–13.

Rhiannon N Turner, Miles Hewstone, and Alberto Voci. 2007. Reducing explicit and implicit outgroup prejudice via direct and extended contact: The mediating role of self-disclosure and intergroup anxiety. *Journal of personality and social psychology* 93(3):369.

Jeffrey R Vittengl and Craig S Holt. 2000. Getting acquainted: The relationship of self-disclosure and social attraction to positive affect. *Journal of Social and Personal Relationships* 17(1):53–66.

Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling self-disclosure in social networking sites. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, New York, NY, USA, CSCW '16, pages 74–85. https://doi.org/10.1145/2818048.2820010.

Myong Jin Won-Doornink. 1985. Self-disclosure and reciprocity in conversation: A cross-national study. *Social Psychology Quarterly* pages 97–107.

Diyi Yang, Zheng Yao, and Robert Kraut. 2017. Self-disclosure and channel difference in online health support groups.

Zhijun Yin, You Chen, Daniel Fabbri, Jimeng Sun, and Bradley Malin. 2016. # prayfordad: Learning the semantics behind why social media users disclose health information. In *Tenth International AAAI Conference on Web and Social Media*.

Ran Zhao, Tanmay Sinha, Alan Black, and Justine Cassell. 2016. Automatic recognition of conversational strategies in the service of a socially-aware dialog system. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Los Angeles, pages 381–392. http://www.aclweb.org/anthology/W16-3647.