Can You be More Polite and Positive? Infusing Social Language into Task-Oriented Conversational Agents

Yi-Chia Wang[1], Runze Wang[2], Gokhan Tur[1], Hugh Williams[2]
[1]Uber AI, San Francisco, California
[2]Uber Technologies, San Francisco, California
{yichia.wang, runze, gokhan, hugh}@uber.com

Abstract

Goal-oriented conversational agents are becoming ubiquitous in daily life for tasks ranging from personal assistants to customer support systems. For these systems to engage users and achieve their goals in a more natural manner, they need to not just provide informative replies and guide users through the problems but also to socialize with users. To this end, we extend the line of style transfer research on developing generative deep learning models to control for a specific style such as sentiment and personality. This is especially useful and relevant to dialogue generation of conversational agents. In this paper, we first apply statistical modeling techniques to understand human-human conversations. We report that social language used by humans is related to user engagement and task completion. After that, we propose a conversational agent model which is capable of injecting social language into agent responses given user messages as input while still maintaining content. This model is based on a state of the art end-to-end dialogue model using a sequence to sequence deep learning architecture, extended with sentiment and politeness features. We evaluate the model in terms of content preservation and social language level using both human judgment and automatic linguistic measures. The results show that the model can generate social responses that enable agents to address users' issues in a more socially conscious way.

1 Background and Introduction

Conversational agents are becoming part of our lives. These systems generally fall into two categories, *task-oriented assistants* and *chatbots* [31]. Task-oriented assistants are designed to fulfill a specific task by having single-turn or multi-turn conversations with users to retrieve information from them and complete that task (e.g., *Microsoft Cortana*, *Apple Siri*, *Google Assistant*). Chatbots are designed to have chit chat with users, and the goal is usually to mimic human-human conversations and engage users in those conversations for as long as possible (e.g., *ELIZA*[35] and *XiaoIce*).

In order to have human-like and extended conversations, some researchers have studied how to incorporate social language into chatbots to generate proper interpersonal responses and build an emotional connection with users [31]. For example, *XiaoIce* can respond with empathetic language and show caring while chatting with users. However, there are only a handful of studies that focus on incorporating social capabilities into task-oriented assistants [1, 6, 3, 34] even though prior literature has suggested that these factors might play an important role in the process of task-oriented conversations and be associated with user engagement and satisfaction [14, 21, 2, 3]. Thus, we propose this work to answer the following research questions: (1) Can and how do social language used by humans in task-oriented conversations affect user responsiveness and task completion? (2) Can we inject a certain type of social language into the responses of a task-oriented conversational agent?

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

We focus on the customer service domain and the task of the driver partner on-boarding support in a ride-sharing provider since customer service is a typical application area of task-oriented assistants. Moreover, the driver partner on-boarding support is a closed-domain problem and has a well-defined task. We first conduct an empirical study to quantitatively examine the relationship of social language usage to driver partner responsiveness and the completion of their first trip, based on a dataset of driver partner and human agent conversations. After that, we apply the findings to build an end-to-end deep learning model to generate agent responses given driver partner inquiries. Our aim is to train a task-oriented agent that can produce dialogues with the desired level of social language while still maintaining the necessary content to guide driver partners through the on-boarding funnel and lead them to complete their first trip. The main contributions of this work are shown below:

1. Systematically analyzed the relationship between social language and user responsiveness as well as task completion using a real-world conversation dataset.
2. Proposed a deep learning framework for task-oriented dialogue generation which includes a social language understanding component.
3. Evaluated the model for both content preservation and social language generation.

We believe that our findings and approach are also applicable to other deep learning based conversational applications including personal assistants and even chit-chat systems.

2 Related Work

This work lies at the intersection of research in natural language processing (NLP), artificial intelligence (AI), and human-computer interaction (HCI). We start with reviewing the related work of social language in task-oriented assistant, and then summarize the literature of language style transfer, which is an emerging research topic in NLP that motivates our work.

Social Language in Task-Oriented Conversational Agents In 1978, Bloom and Lahey proposed a language development framework which proposes language has three components: *content*, *use*, and *form* [5]. The *use* of language is also called *social language* or *pragmatics*, which is about how to use language in different social settings or contexts. They pointed out that social language is important for interpersonal functions. Humans use various kinds of social language strategies to maintain and develop interpersonal relationships, such as increasing intimacy through self-disclosure [10] and building common ground through small talk [8]. Despite the importance of social language for human-human relationships and interactions, most task-oriented conversational assistants only focus on presenting the right content to users; nevertheless, few of them have tried to look at the use of language and incorporate social language in them. Bickmore and Cassell integrated a theory of social dialogue in a real estate conversational agent (REA) and demonstrated that small talk can help the virtual agent build trust with users [6, 1]. However, REA was not fully automated but controlled by a human wizard who followed scripts during the experiment.

Among all kinds of social language, we have decided to concentrate on *politeness* and *positivity* since we argue that these two types of social language can be important strategies for more natural human-machine conversations. According to the politeness theory [20], politeness is a common social language strategy used for saving "face". It helps to regulate the social distance between two parties and remove face threats (e.g., feeling awkward or embarrassed) from them. Thus, the ability of a conversational agent to respond in a polite manner can protect users from "losing face". We have also included positivity in our study since positivity is contagious [19] and leads to liking [12]. So, we hypothesize that users would prefer a conversational agent using more positive language and thus would be more willing to interact and work with it.

Language Style Transfer Language style transfer is to change the underlying style of a text while still preserving its content. In particular, there is a line of research on developing deep network generative models to control for a specific style, including sentiment [30, 16, 26], personality [25], and politeness [27]. This style transfer idea is especially useful and relevant to dialogue generation of conversational agents. For example, Oraby et al. have combined the idea of style transfer with a task-oriented dialogue model and shown that they can alter the personality of an output utterance by varying the personality parameters in the input vector. However, their model was trained and verified on a small synthetic dataset, so the generalizability and practical use of their model is not clear.

3 Driver Partner On-boarding Process and Data

A ride-sharing company provides rides for real-time requests. Before driver partners can start to take passengers, they need to go through an on-boarding process. Our data originates from an initiative in the ride-sharing provider where new driver partners who just signed up on the platform are paired with a dedicated customer support representative (CSR) to guide them through the on-boarding funnel via Short Message Service (SMS). Typically, the funnel starts when a driver partner consents to a background check and ends when the driver partner completes his/her first trip, which marks the completion of the task.

We collected 4 million on-boarding message pairs that were exchanged between driver partners and the customer support agents in the ride-sharing company. All the messages were de-identified. The dataset was used to 1) analyze the impact of politeness and sentiment in CSR messages on driver partner engagement and task completion and 2) build conversational models to generate agent responses with social language cues.

4 The Relationship between Social Language and User Engagement

The goal of this study is to empirically verify how politeness and positive sentiment in CSR responses correlate with driver partner engagement by building statistical analytics models. We measured politeness and positive sentiment using pre-trained classifiers. We operationalized driver partner engagement in terms of their responsiveness and the completion of their first trip.

4.1 Dependent Variables

Driver partner responsiveness Assuming what CSRs say is of interest to driver partners, then ideally driver partners would be more likely to send a reply message. Based on this assumption, we created a binary variable to measure the **short-term driver partner engagement** as whether driver partners will respond to CSR messages within the next 24 hours.

Completion of first trip In addition to the short-term driver partner engagement metric, we also consider the **long-term impact** of CSR responses on finishing the task, i.e., whether CSRs successfully guide driver partners through the on-boarding funnel, and thus help driver partners complete their first trip. Specifically, this binary measure was set to 1 if a driver partner completed his/her first trip within 7 days after a CSR sent him/her a message, otherwise 0.

4.2 Independent and Control Variables

Given a CSR message, we extracted its politeness and sentiment levels as independent variables. Before that we went through several steps to pre-process and clean CSR messages. Messages were tokenized with the NLTK toolkit [4]. We replaced URLs, email addresses, names, dates, and numbers with tags. After that, we utilized pre-trained text classifiers to extract the features:

Politeness We used a state-of-the-art off-the-shelf politeness classifier to measure the politeness level of a CSR message [13]. The model was trained on a corpus labeled for politeness with domain independent lexical and syntactic features developed from politeness theory. The classifier outputs a politeness score between 0 and 1 and performs almost as well as human raters across domains. Below are a few examples of CSR messages with different levels of politeness scores generated by the classifier:

Example 1 (0.27): Please download the Partner app to confirm your account: <URL>

Example 2 (0.43): Hello <Name>, are you still interested in partnering with us? You're so close to hitting the road and making some money while driving.

Example 3 (0.97): Hello, my name is <Name> your Account Specialist. Good news! It looks like your background check has passed! The final step to earning with us is uploading your registration. Could you please text me a clear photo of your registration so I can upload it to your account?

Table 1: Results of the regression analyses

Dependent Variable Control/Independent Variable	Driver Partner Response Coef.	Driver Partner First Trip Coef.
Signup city	(omitted)	(omitted)
Driver partner age	0.016 ***	0.000
Days since signup	-0.057 ***	0.090 ***
Num of driver partner msg	0.036 ***	0.040 ***
CSR msg length	-0.029 ***	0.002 ***
Politeness	0.029 ***	0.006 ***
Pos sentiment	-0.020 ***	0.001 ***

*:p<0.05, **:p<0.01, ***:p<0.001

Sentiment To measure the sentiment, we used VADER, a rule-based sentiment analyzer [18]. VADER was built with a combination of lexical features and general syntactical and grammatical rules to capture the expression and emphasis of sentiment. The authors compared its performance with eleven benchmarks including both lexicon-based and machine learning approaches (e.g., Linguistic Inquiry and Word Count (LIWC) and Naive Bayes algorithm). They showed that VADER outperforms human judges and is generalizable across contexts. Given a piece of text, VADER produces three dimensional measures to estimate the extent of positive, negative, and neutral sentiment in it. The three sentiment scores represent the proportion of each sentiment in the text and sum up to one as the example shown below. Since positive sentiment score is highly negatively correlated with neutral and negative sentiment, we only included positive score in the models to avoid multicollinearity.

Example 4 (pos=.49, neg=.0, neu=.51): Nice! The 2 links I sent you will be your best friends. Good luck! Let me know how it goes for you.

In addition to some basic demographic information about driver partners such as their age, we also measured the following control variables. By controlling for these variables, we can make claims about the impact of a CSR message rather than about the driver partner himself/herself.

Sign-up city is a dummy variable controlling for the city where a driver partner signed up to become a partner with the ride-sharing company.

Days since signup is the number of days since the driver partner registered to the platform.

Number of previous driver partner messages is a measure of how many messages the driver partner sent to the CSR since he/she signed up. Different driver partners might have different levels of tendency to reply to a CSR message so we used this variable to control for the response variability among driver partners.

CSR message length is the total number of characters in the CSR message.

Except for the binary and dummy variables, all the numerical control and independent variables were standardized and centered, with a mean of zero and standard deviation of one. Additionally, we took the logarithm of the variables *Days since signup* and *Num of driver partner messages* before they were standardized since they had a skewed distribution.

4.3 Analyses and Results

This analysis seeks to statistically test the effect of the level of politeness and positive sentiment in CSR messages on driver partner responsiveness and their completion of the first trip. The unit of analysis was a CSR message, and we built a random-effects linear regression model which grouped CSR messages at the driver partner level (because the same driver partner might receive multiple CSR messages) to deal with non-independence of observations. The results are shown in Table 1. We omitted the *Sign-up city* variables in the table since there are many of them, and we included them in the models mainly for controlling purposes but not their interpretability.

First, considering the control variables, the driver partner responsiveness model shows that older driver partners and those who were more responsive in previous conversations were more likely to reply to the current message; driver partners who had signed up a longer time ago and received longer

CSR messages tended not to respond. On the other hand, all the control variables have a positive effect on the completion of the driver partners' first trip.

Next, we examined the independent variables. CSR messages with a higher level of politeness were more likely to elicit driver partner responses as well as lead driver partners to on-board and complete their first trip. However, although positive sentiment score is positively correlated with driver partner first trips, it negatively predicts driver partner responsiveness, which is counter-intuitive. To further confirm this finding, we replaced the VADER positive sentiment score with LIWC positive emotion measure in the model and still obtained a similar result. One explanation for this surprising finding is that CSRs send a congratulatory message every time when driver partners achieve a milestone. These messages were template-based and crafted with a highly positive tone and thus have a high positive sentiment score, but driver partners usually did not reply to this kind of status update messages.

In general, the significant impact of politeness and positive sentiment on driver partner responsiveness and first trip completion provide evidence that social language in task-oriented conversations can help achieve the task, which motivates us to propose a novel conversational agent framework which can automatically generate agent responses with the desired amount of social language given driver partner messages as input.

5 A Conversational Agent Generates Responses with Social Language

This section presents a conversational agent which can generate responses with more or less social language. The ability to adjust the usage of social language in a conversational agent is essential for the customer support domain, since it not only has positive impact on achieving the task as shown in the last section but also positively correlates with customer satisfaction based on our preliminary analysis.

5.1 Agent Response Generation

In recent years, deep neural networks have become a trend in AI research because of their effectiveness. Among all types of deep learning architectures, a sequence-to-sequence learning approach (*seq2seq*) has been most widely and successfully adopted for natural language processing problems, such as machine translation [e.g. 32], question answering [e.g. 36], text summarization [e.g. 7], and conversational models [e.g. 33, 29, 28]. A typical *seq2seq* model is designed to transform one sequence to another. To do so, it has two sub-modules: an encoder and a decoder. The encoder takes a sequence as input and internalizes it as a vector representation, which is then passed to the decoder to generate a corresponding output sequence. When applied to end-to-end conversational modeling, it generates the next utterance given the previous utterance. In our case, the input sequence is a driver partner message, and the output sequence is a CSR response.

To incorporate social language into a *seq2seq* model, we developed an architecture which is inspired by Huber et al. [17]. They proposed a conversational agent to generate emotionally appropriate responses by extracting features from images attached in conversations. In order to integrate visual information extracted from images into a dialogue model, they modified a *seq2seq* structure that uses visual information together with lexical input for conversational language generation. We modified their architecture by replacing the image understanding layer with a social language layer. The politeness and positivity features are extracted from CSR responses using pre-trained classifiers described earlier. We evaluated this model against the baseline, a typical *seq2seq* model without the social language layer. Figures 1 presents the architectures of the baseline and the proposed model, and we describe the model details below:

Lexical Model Our baseline is a classic *seq2seq* model [32] which transforms a driver partner message to a CSR response. We added an embedding layer for both the encoder and decoder to convert sparse one-hot word representations to dense vector representation [24]. The main advantage of embedding is that it maps words into a latent semantic space so that words with similar meanings and contexts would be closer to each other in that space (e.g., *picture* and *photo*). We built an encoder and decoder recurrent neural network (RNN) with long short-term memory units (LSTM) so that the model can capture word dependencies [15]. The embedding dimension is 300, and the dimensionality of the internal state is set to 512.

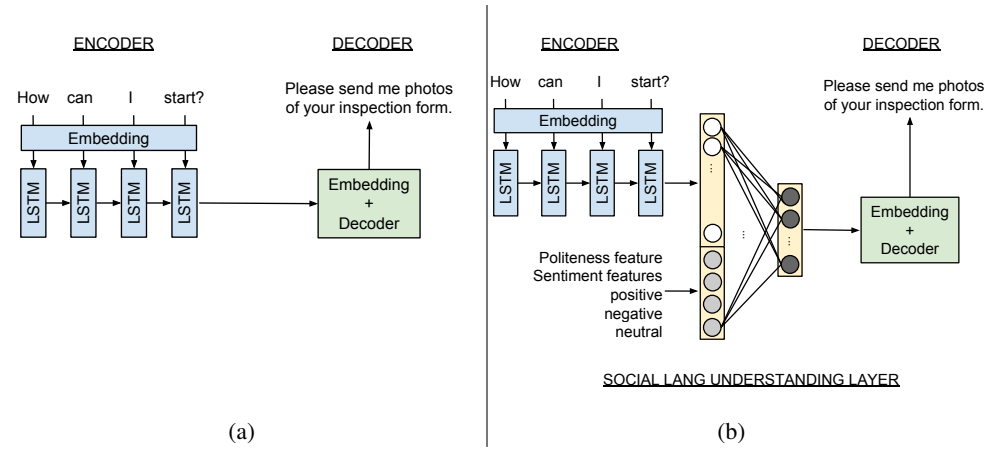

Figure 1: Model architectures of the baseline and the proposed model: (a) The baseline lexical model implements a typical *seq2seq* architecture. It has an embedding layer to convert one-hot word representations to dense representations, and utilizes LSTM cells to capture dependencies among words. (b) The proposed model has a social language understanding layer in-between the encoder and decoder to integrate the politeness and sentiment features with the lexical information.

Lexical and Social Model To take into account of social language in agent responses, we introduced a social language understanding layer in-between the encoder and decoder as shown in Figure 1b. During the model training phase, we applied the pre-trained politeness and sentiment classifiers to extract social language features from CSR responses. We concatenated the social feature vector with the lexical feature vector output by the encoder and passed it to a fully-connected feed-forward neural network. The output values of the fully-connected layer then become the initial state of the decoder. We did not employ attention mechanism or more complex models because our goal is to directly evaluate the impact of the social language layer on the output.

Our dataset consists of 233,571 data points. Each data point is a pair of a driver partner message and a CSR response along with the politeness and sentiment features extracted from the CSR response. Driver-CSR messages were paired together only if the CSR reply was sent within an hour after the driver partner's inquiry. The data was split into the train, validation, and test sets with a ratio of 80%:10%:10% where data points came from the same driver partner would only be assigned in one set. We trained both models on the training set with early stopping based on the validation loss. We used both automatic methods and human judgment to evaluate the agent responses generated by our models. All the evaluation results reported below were based on the hold-out test set.

5.2 Evaluation of Content Preservation

We conducted automatic evaluations to examine qualities of the generated text using the *word2vec* similarity measure [23]. This metric measures the text similarity between the actual CSR responses and the model-generated responses. Our goal is to quantitatively inspect whether and how much the model-generated responses preserve the contents in the ground-truth responses. The idea is that although we introduced a social language layer in the model, we expected that the lexical feature vector output by the encoder should still capture the content, and thus the model should perform at least as well as the baseline model. We used this measure rather than a simple n-gram overlap metric because it can capture a high-level semantic similarity.

***word2vec* Similarity Measure** We computed how similar the model outputs are to the ground truth in terms of their *word2vec* [23] representations. *word2vec* is one of the state-of-the-art word embedding methods, which convert each word to a vector representation in a latent semantic space such that words used in common contexts are positioned close together in the space. Specifically, for each utterance, we mapped its words to their word embedding vectors using a *word2vec* model pre-trained on Google News. Then we averaged the word embedding vectors across the utterance to derive a vector representation for that utterance. We did that for both the model-generated response and its corresponding ground truth, and computed the cosine similarity between their vector representations as their *word2vec* similarity measure.

Table 2: The crowdsourcing task to compare the agent responses generated by the social model with different levels of politeness.

Driver partner message: i need to do the inspection just looking for a place close to my house **Agent response 1**: visit any one of these locations for a free inspection - <url> **Agent response 2**: ok , i 'll send you a link to the nearest free inspection location .
Please answer the following three questions: 1) Does the agent response 1 reply to the driver partner message? (Yes/No) 2) Does the agent response 2 reply to the driver partner message? (Yes/No) 3) Which response is more polite? (1/2/cannot tell)

We computed the *word2vec*-based cosine similarity on a test set of 22,947 pairs of ground truth and model response. We found that adding the social language information significantly improves the similarity score from 0.689 to 0.750, a 9.38% relative increase (pairwise t-test p<.000). This finding suggests that the social model has a better ability to preserve the contents in the ground-truth responses than the baseline *seq2seq* model even though there is a social language understanding layer between the encoder and decoder.

5.3 Social Language Evaluation

After confirming that the social model can maintain content, the next step is to investigate whether we can adjust the level of politeness or sentiment in the model-generated agent responses by changing the value of the politeness or sentiment feature. We utilized crowdsourcing to collect human judgment on the model-generated dialogue responses and also conducted an automatic analysis on the model outputs using the politeness and sentiment classifiers.

Human Judgment In the politeness crowdsourcing task (Table 2), crowdworkers were presented with a driver partner message with two possible agent responses, where the *unboosted* response was generated by the social model with the original level of politeness extracted from the ground truth CSR response; the *boosted* response was also output from the social model but with the politeness feature value increased by one standard deviation. The standard deviation of the politeness feature was calculated based on the test set. To avoid biased judgment, two candidate responses were shown in random order so crowdworkers would not know which one was generated by which model. Crowdworkers were then asked to answer three questions. The first two questions were used to assess whether the two generated responses were replying to the driver partner's message and addressing his/her inquiry. We included these two questions to check whether our model can produce reasonable responses within the context but not something out of nowhere. The third question asked them to compare the two responses and pick the one that is more polite.

We randomly selected 200 data points from the test set and assigned each driver-agent conversation to three different crowdworkers. There was a similar crowdsourcing task for positive sentiment which was evaluated by another three crowdworkers. We measured the inter-rater agreement among the three crowdworkers for each question and both tasks [9, 22, 11]. The result shows that there is a fair to moderate agreement (*Politeness Q1*=.5, *Q2*=.4, *Q3*=.3; *Positivity Q1*=.3, *Q2*=.4, *Q3*=.2).

We evaluated the human judgment by taking a majority vote of the three crowdworkers for each of the three questions (Table 3). The result indicates that there is a significant increase in the politeness level for the responses generated by the boosted model. However, there are two unexpected findings. Firstly, we found no significant difference between the unboosted and boosted responses for the naturalness and the positive sentiment questions. Moreover, the percentages of natural responses for politeness and positivity have a large gap. These two unexpected findings might be due to the small sample, the moderate agreement among crowdworkers, or the confusing definition of the tasks.

Automatic Analyses We further applied the politeness and sentiment classifiers to rate both unboosted and boosted responses automatically for the entire test set. Upon doing so, we observed that the responses generated by the social model with the boosted politeness or positivity input features have significantly higher politeness or positivity scores (Table 4). The results of this evaluation together with the content preservation evaluation confirm that the proposed model can generate social responses that address driver partners' issues.

Table 3: Crowdworker majority vote on the 200 model-generated responses

	Model	% of Natural Responses (Q1, Q2)	% of Social Responses (Q3)
Politeness	Unboosted	34.0%	13.0%
	Boosted	35.0% (2.94%; p=0.916)	44.0% (238%; p<.000)
Positivity	Unboosted	58.0%	35.0%
	Boosted	62.5% (7.76%; p=0.414)	28.0% (-20.0%; p=0.162)

Table 4: Results of average politeness and positive sentiment scores of responses generated by the unboosted and boosted models. These scores were computed on the test set of 22,947 data points.

Model	Avg. Politeness Score	Avg. Positive Score
Unboosted	0.601	0.147
Boosted	0.737 (22.7%; p<.000)	0.259 (76.6%; p<.000)

6 Conclusion

In this paper we investigated whether and how social language is related to user engagement in goal-oriented conversations. We found that CSR messages with a higher level of politeness and positive sentiment were more likely to elicit user responses as well as guide users to accomplish the task. Second, to integrate the findings from the statistical analyses into a dialogue model, we proposed a task-oriented conversational agent framework which can generate agent responses with the desired level of social language by inserting a social language understanding layer into a typical *seq2seq* model. We evaluated the model regarding whether it can preserve the content and boost the level of social language, and found that the model outperforms the baseline on both evaluation approaches.

This work has several implications. First, although we only focus on politeness and positive sentiment, we believe that our proposed modeling framework should be generalizable to incorporate other kinds of social language and task-oriented assistants. The model can also be used to provide driver partners with better experience during their on-boarding process. The customer support services can be improved by utilizing the model to provide suggested replies to CSRs so that they can (1) respond quicker and (2) adhere to the best practices (e.g., using more polite and positive language) while still achieving the goal that the driver partners and the ride-sharing providers share, i.e., getting them on the road.

Future Directions Although we argue that the negative correlation between positivity and driver partner responsiveness is due to the positive tone of milestone notification messages, we need more evidence to support it. One future research direction is to conduct a comprehensive error analysis to understand this finding.

Our current model only takes driver partner messages as input and does not include any context so it sometimes generates unnatural agent responses (Table 3). To avoid out-of-context responses and further constraint the model output, one improvement is to provide more information to the model, such as control for driver partners' document status and incorporate their conversation history. In addition, the model relies on one single input to capture the pattern of a certain social language. Another direction for the model improvement is to consider, for example, to utilize all the raw features which were used to build the politeness classifier as input than just the classifier prediction.

The moderate inter-rater agreement of the crowdsourcing task and the unexpected human evaluation results of positivity suggest that crowdworkers might have some confusion about the task. So, there is a need to investigate the labelling process and refine the task. We would also like to expand our crowdsourcing effort to engage more workers to review and evaluate more model-generated responses.

Although we found that there is a correlation between social language and user engagement, the results are based on the analysis of human-human conversations. We have no idea whether this finding can apply to human-bot interactions, i.e., whether a polite conversational agent would also improve user engagement. Future work should conduct A/B test to examine the effectiveness of a polite and positive conversational agent.

Acknowledgments

We thank Robert E. Kraut, Alexandros Papangelis, and Piero Molino for providing valuable feedback.

References

[1] T. Bickmore and J. Cassell. Relational agents: A model and implementation of building user trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 396–403, New York, NY, USA, 2001. ACM. ISBN 1-58113-327-8. doi: 10.1145/365024.365304. URL http://doi.acm.org/10.1145/365024.365304.

[2] T. W. Bickmore and R. W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Trans. Comput.-Hum. Interact.*, 12(2):293–327, June 2005. ISSN 1073-0516. doi: 10.1145/1067860.1067867. URL http://doi.acm.org/10.1145/1067860.1067867.

[3] T. W. Bickmore, L. M. Pfeifer, and B. W. Jack. Taking the time to care: Empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1265–1274, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1518891. URL http://doi.acm.org/10.1145/1518701.1518891.

[4] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st edition, 2009. ISBN 0596516495, 9780596516499.

[5] L. Bloom and M. Lahey. *Language development and language disorders.* ERIC, 1978.

[6] J. Cassell and T. Bickmore. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13(1):89–132, Feb 2003. ISSN 1573-1391. doi: 10.1023/A:1024026532471. URL https://doi.org/10.1023/A:1024026532471.

[7] S. Chopra, M. Auli, and A. M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, 2016.

[8] H. H. Clark. *Using Language*. Cambridge University Press, 1996.

[9] J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[10] N. L. Collins and L. C. Miller. Self-disclosure and liking: a meta-analytic review. *Psychological bulletin*, 116(3):457, 1994.

[11] A. J. Conger. Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2):322, 1980.

[12] M. Dainton, L. Stafford, and D. J. Canary. Maintenance strategies and physical affection as predictors of love, liking, and satisfaction in marriage. *Communication Reports*, 7(2):88–98, 1994.

[13] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 250–259, 2013. URL http://aclweb.org/anthology/P/P13/P13-1025.pdf.

[14] R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz, and J. Wang. Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1338–1343, Aug 2005. doi: 10.1109/IROS.2005.1545303.

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[16] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, and E. P. Xing. Toward controlled generation of text. *arXiv preprint arXiv:1703.00955*, 2017.

[17] B. Huber, D. McDuff, C. Brockett, M. Galley, and B. Dolan. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 277:1–277:12, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6. doi: 10.1145/3173574.3173851. URL http://doi.acm.org/10.1145/3173574.3173851.

[18] C. J. Hutto and E. Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014.*, 2014. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109.

[19] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, page 201320040, 2014.

[20] P. Levinson, P. Brown, S. C. Levinson, and S. C. Levinson. *Politeness: Some universals in language usage*, volume 4. Cambridge university press, 1987.

[21] Q. V. Liao, M. Davis, W. Geyer, M. Muller, and N. S. Shami. What can you do?: Studying social-agent orientation and agent proactive interactions with an agent for employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, DIS '16, pages 264–275, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4031-1. doi: 10.1145/2901790.2901842. URL http://doi.acm.org/10.1145/2901790.2901842.

[22] R. J. Light. Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological bulletin*, 76(5):365, 1971.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[24] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999959.

[25] S. Oraby, L. Reed, S. Tandon, S. T.S., S. Lukin, and M. Walker. Controlling personality-based stylistic variation with neural natural language generators. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 180–190. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/W18-5019.

[26] S. Prabhumoye, Y. Tsvetkov, R. Salakhutdinov, and A. W. Black. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/P18-1080.

[27] R. Sennrich, B. Haddow, and A. Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 35–40, 2016.

[28] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784, 2016.

[29] L. Shang, Z. Lu, and H. Li. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586. Association for Computational Linguistics, 2015. doi: 10.3115/v1/P15-1152. URL http://www.aclweb.org/anthology/P15-1152.

[30] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems*, pages 6830–6841, 2017.

[31] H.-y. Shum, X.-d. He, and D. Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26, Jan 2018. ISSN 2095-9230. doi: 10.1631/FITEE.1700826. URL https://doi.org/10.1631/FITEE.1700826.

[32] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

[33] O. Vinyals and Q. V. Le. A neural conversational model. *CoRR*, abs/1506.05869, 2015. URL http://arxiv.org/abs/1506.05869.

[34] M. A. Walker, J. E. Cahn, and S. J. Whittaker. Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the first international conference on Autonomous agents*, pages 96–105. ACM, 1997.

[35] J. Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, Jan. 1966. ISSN 0001-0782. doi: 10.1145/365153.365168. URL http://doi.acm.org/10.1145/365153.365168.

[36] J. Yin, X. Jiang, Z. Lu, L. Shang, H. Li, and X. Li. Neural generative question answering. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2972–2978. AAAI Press, 2016. ISBN 978-1-57735-770-4. URL http://dl.acm.org/citation.cfm?id=3060832.3061037.