A Human-Centered Hierarchical Framework for Dialogue System Construction and Evaluation

Salvatore Giorgi,[1] **Farhan Ahmed,** [2] **Lyle Ungar,** [1] **H. Andrew Schwartz** [2]

[1] University of Pennsylvania
[2] Stony Brook University
sgiorgi@sas.upenn.edu, farhaahmed@cs.stonybrook.edu, ungar@cis.upenn.edu, has@cs.stonybrook.edu

Abstract

We propose a human-centered framework for evaluating dialog systems in which the language of dialog turns are part of a psychologically-grounded hierarchical process: agents' language is generated in the context of stable *traits* such as personality and dynamic *states* that vary from turn to turn. In the context of the Dialogue System Technology Challenge 10 (DSTC10) shared task, we present a framework for evaluating dialog systems through this psychologically-grounded "human" lens, considering both the states expressed in individual turns (and how well they match those of the conversation partner, as humans do), as well as the diversity of traits (e.g., demographics or personality) across agents' entire dialogs. This hierarchy consists of four levels – the dialog system, agents, dialogues, and turns, within which we define five metrics using well validated psychological constructs: emotional entropy (across turns, aggregated to agents), linguistic style and emotion matching (dialog and turn level), as well as agreeableness and empathy (agent level, across dialogues). Evaluating these metrics on five turn-level data sets shows that emotional entropy outperforms baseline systems as well as our other metrics. Within the larger shared task, this same metric comes in first and third place on dimensions of content and grammar, respectively.

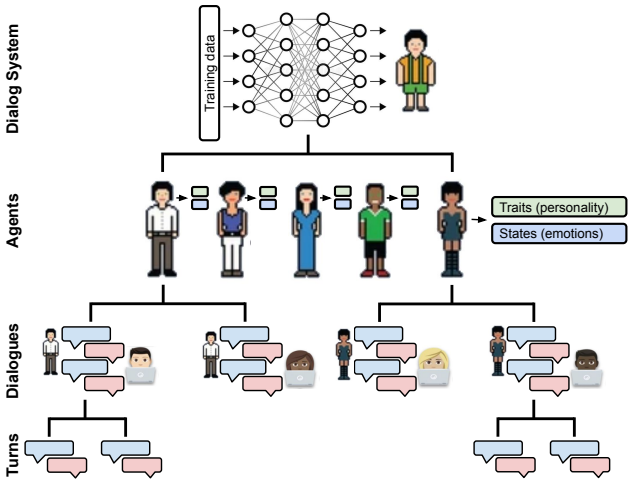

Figure 1: The proposed hierarchical structure of open-domain dialog systems. We note that Turns consist of one utterance from both the agent and entity, but turn level outcomes do not need to be measured with the agent's utterance preceding the entity's (as depicted in the figure).

Introduction

Open-domain dialog systems have been evaluated by both automatic methods and human annotations, both of which have a number of drawbacks. Automatic methods (such as BLEU, METEOR, and ROUGE), which can rely on word overlap, fail to capture the diversity of dialog systems (Liu et al. 2016). On the other hand, human evaluations, which often consider the appropriateness or grammar of a response, are expensive and lack standardization (Sedoc et al. 2019). Compounding these issues is the fact that automatic evaluations often do not correlate with human evaluations (Liu et al. 2016; Deriu et al. 2021). In order to address these issues, the Dialog System Technology Challenge 10 Shared Task aims to develop an automatic evaluation for open-domain dialog systems and asks participants to develop metrics which are both correlated with human judgement and explainable (Chen et al. 2021).

In this work, we propose a framework for evaluating open-domain dialog systems as if they were human, taking queues from Giorgi, Ungar, and Schwartz (2021) which characterized Twitter spambots through a number of human traits. Specifically, we propose a hierarchical framework for dialog system construction and evaluation which consists of four levels, as seen in Figure 1: the dialog system (the high level architecture of the system), agents (specific instances of the system which engage in conversations), dialogues (a back and forth exchange between an agent and another entity), and turns (specific utterances within a dialog). Across these four levels, we propose two general classes of psychologically-grounded measures: state vs. traits metrics (i.e., constructs which remain stable within agents or change dynamically across dialogues) and linguistic matching (i.e., how well do turns and dialogues match the linguistic cues of the other entity in the conversation). For the shared task, we implement five specific instances of these two classes of measures (empathy, agreeableness, linguistic style match-

ing, emotion matching, and emotional entropy) and apply them to both dialog and turn level data sets. Our best system (emotional entropy) ranked 26 out of 37 systems.

Contributions Our contributions include: (1) a hierarchical framework for understanding open-domain dialog systems and (2) a number of psychologically-grounded and human-centered evaluation measures.

Related Work

Our work is aligned with a growing set of methods to embed language processing within human contexts in which they are applied (Volkova, Wilson, and Yarowsky 2013; Hovy 2015; Lynn et al. 2019). In the domain of language generation or dialog agents, a number of works have attempted to *create* agents with human-like traits. This includes embedding agents with empathy (Rashkin et al. 2019; Lin et al. 2020), trust (Novick et al. 2018), and emotion (Zhou and Wang 2018; Huber et al. 2018) as well as general personalizations and personas (Li et al. 2016; Zhang et al. 2018; Mazaré et al. 2018; Roller et al. 2021).

There is also a parallel line of work which aims to *evaluate* dialog agents and conversations as human, with a number of human-like metrics having been proposed. Work by Adiwardana et al. (2020) proposes a metric which jointly measures making sense and being specific, which they note are both basic and important attributes of human conversations. More directly, some have attempted to quantify "humanness" by asking crowd-source annotators: "Which speaker sounds more human?" (Li, Weston, and Roller 2019; Roller et al. 2021). Similarly, Deriu et al. (2020) replace human-bot conversations with bot-bot conversations and ask annotators to label whether or not the bots are human.

Our current proposal builds on the two lines of work outlined above (i.e., creation and evaluation of "human-like" dialog systems) in that we (1) define a hierarchical framework for dialog systems and (2) propose two classes of human-centered measures which can be used to create and evaluate dialog systems, with the goal of working towards a human-like open-domain system (Adiwardana et al. 2020). As such, we believe past work can easily be reframed within our proposed hierarchy. For example, Roller et al. (2021) propose both "engaging talking points" and "consistent persona" as desirable qualities of an open-domain dialog systems. Within our proposed framework "engaging talking points" happen at the turn level while a "consistent persona" can be defined at the agent level, where consistency can be measured across multiple dialogues. Despite the general structure of our proposed framework, we note that a number of dialog systems are task or goal oriented, such as question/answer systems (Chen et al. 2017) or systems designed for highly specific tasks such as trip planning (El Asri et al. 2017) and customer service (Cui et al. 2017). Such systems may be considered outside of the scope of our formulation, in that scheduling a trip is fundamentally different from, for example, a conversational chatbot related to COVID-19 vaccines, which may need additional social and cultural context.

The Dialog System Hierarchy

Figure 1 shows our proposed hierarchical framework for dialog systems which consists of four parts, each of which are defined below: the dialog system, agents, dialogues, and turns. We propose that dialog systems should be created and evaluated with all four levels in mind. This includes such considerations when open-sourcing both the model and data (e.g., is there sufficient information to piece together turns into dialogues and dialogues across agents). We note that the DSTC10 Shared Task asked participants to evaluate dialog system at either the dialog or turn level.

Dialog System This is the overall architecture of the system and the top level of our hierarchy. For example, this could be specified as "a 5-layer LSTM sequence-to-sequence model with attention".

Agent An agent is a specific instance of a dialog system and we note that a single dialog system can produce a number of agents. With this view, a dialog system can be thought of as an agent generator.

Dialog A dialog is a complete back and forth exchange between an agent and another entity, where this second entity could be another dialog agent or a human. A single agent can engage in multiple dialogues.

Turn Finally, a turn is a specific utterance with a dialog. This could include the second entity's preceding or proceeding response.

Human-like Measures

Working under the goal of a human-like open-domain dialog system, we propose two classes of measures: (1) states and traits and (2) linguistic matching. Both classes are rooted in psychological and social sciences and relate to fundamental psychological measurements of humans (i.e., states and traits) and social relationships and interactions (i.e., linguistic matching). These measures can be used to study the dialog system as its own "human" concept and how the dialog system interacts with the world, respectively. In the next section we give examples of how we operationalize these measures and define five metrics which were used for the Dialog System Technology Challenge 10 shared task.

States and Traits The state vs. trait distinction is ubiquitous in psychology, with a long history (Carr and Kingsbury 1938) and, as such, we present standard textbook definitions (Zeigler-Hill and Shackelford 2020): state measures are thoughts, feelings, and behaviors in a specific place and time; while trait measures are those which generalize across situations, remain stable across time, and systematically differ across people. For example, personality is a trait measure while emotions are states. In relation to standard NLP tasks, similar state vs. trait considerations were theorized to relate to sarcasm, stance, and sentiment classification, with stance being a trait-like outcome and sentiment being a state-like outcome (Lynn et al. 2019). It is important to distinguish the measures we use (e.g., personality), grounded against validated psychological instruments, with

proxies for these constructs used in other works (e.g., personas). While proxy measures like "likes" seem to be related to personality (Kosinski, Stillwell, and Graepel 2013), they are not direct assessments of the constructs.

Within our hierarchy, we propose that states and traits be considered as follows: First, at the top level, dialog systems should have the capacity to produce a number of agents with varying traits, while each agent should maintain its given traits across dialogues. That is, one should be able to measure variation in traits across agents from a single dialog system and stability in traits from a single agent across multiple dialogues. On the other hand, states should vary within dialogues and agents should have the capacity to exhibit a range of states.

Linguistic Matching Linguistic matching, while mostly unconscious, has been observed in many settings. It has been shown to predict power differentials (Danescu-Niculescu-Mizil et al. 2012), relationship stability (Ireland et al. 2011), cooperation (Manson et al. 2013), and empathy ratings of therapists (Lord et al. 2015). More generally, the psycholinguistic theory of communication accommodation has studied such unconscious matching tendencies in postures, facial expressions, pitch, pausing, length, and use of function words (Giles, Coupland, and Coupland 1991). To our knowledge, such extensive matching phenonema have yet to be fully studied in open-domain dialog systems, despite being applied in other NLP settings (Danescu-Niculescu-Mizil, Gamon, and Dumais 2011; Danescu-Niculescu-Mizil and Lee 2011). We believe linguistic matching can be a natural extension of the "continually learning" framework proposed by Roller et al. (2020), where agents adapt to new contexts and users. Finally, we note that the sentence embedding similarly metric used as a baseline in the DSTC10 Shared Task can be viewed as a simple matching measure (Zhang et al. 2021).

As measured within our proposed hierarchy, linguistic matching is a property of dialogues and turns. For example, one could measure function word matching in a single turn (i.e., how does well does the agent match the prompt?) or across a dialog (i.e., what is the difference in function word between the agent and entity in a dialog?).

Task Metrics

The DSTC10 Subtask 1 asked participants to submit five metrics which should (1) correlate with human judgement and (2) be explainable. As noted previously, these metrics operationalize the human-like measures and were not specifically design for evaluation diaglog systems, nor are they optimized to correlate with the evaluation metrics in the data sets (e.g., content and grammar). Metric scores were produced at the turn and dialog level (depending on the data set) and then correlated with a number of crowd-sourced human evaluations.[1] We define our five metrics below.

[1]We note that "human evaluations" (i.e., evaluation by a human) as opposed to "automatic evaluations" (i.e., evaluations by a machine with no human judgements) are different than systems being evaluated as "human-like".

Emotional Entropy Using the NRC Hashtag Emotion Lexicon (Mohammad and Kiritchenko 2015) we estimate Plutchik's eight basic emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust (Plutchik 1980). This emotion lexicon, which is a set of weighted words for each emotion category, was automatically derived over tweets with emotion hashtags (e.g., *#anger* and *#joy*). The lexicon is applied to every observation in each data set (i.e., we summed weighted word frequencies which were weighted according to their weight within each emotion category) and then the entropy of the normalized emotion vector is calculated. Emotions (and, thus, emotional entropy) are state measures and can be estimated at multiple levels of the hierarchy: turn, dialog, and agent.

Agreeableness We used a language based personality model to estimate the agreeableness dimension of the Big Five personality traits (Park et al. 2015). This model had an out-of-sample prediction accuracy (Pearson r) of .35 and was built over 1-3grams and 2,000 LDA topics (Latent Dirichlet Allocation; Blei, Ng, and Jordan 2003). Thus, for each turn or dialog, we extracted 1-3grams and loadings for the 2,000 LDA topics and applied the pre-trained regression model, which produced an agreeableness score for each observation. We include agreeableness in our final five metrics since it out performed the other four personality measures (openness to experience, conscientiousness, extraversion, and neuroticism) on the test data. Agreeableness (and personality, in general) is a trait measure that would typically be defined at the agent level or above (e.g., for a given dialog system, does agreeableness vary across agents and is it stable within an agent), though do to lack of agent level data in the task we estimate agreeableness for both the turn and dialog level data sets.

Empathy We build a model to predict empathy, as measured by the Empathic Concern subscale of the Interpersonal Reactivity Index (IRI) (Davis 1983). We use the same data set as Abdul-Mageed et al. (2017) and build a model over 2,805 participants who shared their Facebook status data and answered the IRI questionnaire. Using 10-fold cross validation, we predicted the empathic concern scores from a Ridge penalized linear regression using the same set of 2,000 LDA topics described above. The final model resulted in an out-of-sample Pearson r of 0.26. In order to obtain Empathic Concern estimates for each turn and dialog, we extracted 2,000 LDA topic loadings for each observation and applied the pre-trained regression model. Empathic Concern is a trait level measure. Similar to agreeableness, this would typically be defined at the agent level or above, but for this task we estimate Empathic Concern for turns and dialogues.

Language Style Matching We use the definition provided by Ireland et al. (2011): 1 minus the normalized absolute difference in function word use between the agent and entity. This score was calculated for nine separate function word categories in the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker, Francis, and Booth 2001): personal pronouns, impersonal pronouns, articles, conjunctions, prepositions, auxiliary verbs, high frequency adverbs,

negations, and quantifiers. Turn and dialog level scores were averaged across the nine categories. This is a form of Linguistic Matching which can be measured at the turn, dialog, and agent levels.

Emotion Matching Again, we use the NRC Hashtag Emotion Lexicon (Mohammad and Kiritchenko 2015) and calculate the Spearman rank correlation between the agents emotions and the prompts. Emotion Matching is a form of Linguistic Matching which can be measured at the turn, dialog, and agent levels.

Data and Evaluation

Data The task was evaluated across five separate *turn level* data sets: JSALT, ESL, NCM, DSTC10-Topical, and DSTC10-Persona (Sedoc et al. 2019). Systems are evaluated along the following dimensions, all of which use a 5 point likert scale (Zhao, Lala, and Kawahara 2020):

- **Appropriateness**: "The response is appropriate given the preceding dialogue".
- **Content**: "How much information is provided in the response. / How difficult is it to guess the context of the preceding dialogue from the response."
- **Grammar**: "The quality of the English grammar."
- **Relevance**: "The response content is related to the preceding dialogue."

Evaluation Evaluation is done as follows: First, the Spearman correlation is computed between each dimension (i.e., a human evaluation such as appropriateness) and a given metric within a given data set. Depending on the data set, the Spearman correlation is evaluated across either turns or dialogues. The correlations are then averaged across all dimensions in the data set (e.g., appropriateness, content, grammar, and relevance). Finally, in order to produce a single evaluation accuracy per metric, the correlations are averaged across all data sets. This process is repeated for each metric submitted to the shared task.

Baselines Three baselines metrics are used: AM (Adequacy Metric), FM (Fluency Metric), and Deep AM-FM. AM is a similarity score (i.e., cosine similarity) in a low dimensional embedding space constructed via Latent Semantic Indexing (D'Haro et al. 2019). For the FM metric, the probability for each sentence (i.e., the entity's prompt and the agents response) is calculated via an ngram language model. FM is then calculated as the ratio of the minimum and maximum sentence probabilities. Deep AM-FM is a deep neural network version of the AM-FM framework (Zhang et al. 2021).

Results

Results from the shared task as shown in Table 1. The emotional entropy measure was the best performing system in the Content dimension in the DSTC10-Topical data set and the third best performing system in the Grammar dimension in the DSTC10-Persona data set. In general, entropy measure outperformed all of our other metrics. When compared to the baseline measures, our metrics meet or exceed performance in four out of eleven evaluations, which in addition to the previously mentioned dimenions, include JSALT appropriateness and DSTC10-Persona Content. We note that emotional entropy was our only metric which measured variation in states. This system ranked 26 out of 37 total submissions.

The next two best performing metrics were both related to linguistic matching: style and emotion matching. These ranked 33 and 32, respectively. Our two trait level metrics, agreeableness and empathy, were the lowest performing metrics, ranking 35 and 36, respectively.

Unfortunately, due to the nature of the data, we were unable to evaluate the systems across all levels of our proposed hierarchy. Out of 14 data sets available during the development phase, only 2 were dialog level, with the remaining 12 at the turn level. In the final evaluation phase, all 5 data sets were turn level. Thus, it is not surprising that the two trait level metrics (agreeableness and empathy) did not perform well in the final evaluation. In Table 2 we show the results of our five metrics on the two dialog level data sets available during development: FED-Conversation (Mehri and Eskenazi 2020) and Persona-Chatlog (See et al. 2019). Here we see both of our trait measures outperforming our remaining metrics and performing on the same level as or exceeding the baseline system (empathy and agreeableness, respectively) on the FED-Conversation data. Thus, it is plausible that our metrics depend on the hierarchical level at which they are applied.

Conclusions

In this paper, we proposed a hierarchical framework for evaluating open-domain dialog systems with human-centered measures which consider both trait and state trade-offs (standard measures of human constructs) and linguistic matching (indicators of social relationships and interactions). Five metrics were evaluated, which examined trait level features (agreeableness and empathy), state level variation (emotional entropy), and linguistic matching (style and emotion matching). Due to data limitations, we were unable to fully evaluate the metrics within the proposed hierarchical framework. Despite this, given the turn level focus on the evaluation data, our proposed agent state level variation metric (emotional entropy) outperformed the other four, which is consistent with our hierarchical formulation.

Ethical Considerations There are a number of ethical considerations when constructing and evaluation dialog systems, many of which have been outlined by Roller et al. (2021). These include privacy (since online dialog may contain sensitive information), toxic and offensive content, and, on the part of the researcher, openness to sharing findings. With regard to the current work, imparting system with human qualities such as personality and socio-demographics must be handled with the utmost sensitivity. Biases in training data, misclassifications in downstream tasks, and reliance on outdated social constructs (i.e., binary gender) are just a few examples of how automated systems can fail and further marginalize vulnerable populations (Shah, Schwartz, and Hovy 2020; Xu et al. 2021; Gonen and Goldberg 2019).

	JSALT	ESL	NCM	DSTC10-Topical				DSTC10-Persona				Avg.
	App.	App.	App	App.	Content	Grammar	Relevance	App.	Content	Grammar	Relevance	
AM	.01	.03	.04	.12	.02	.04	.16	.11	.01	.05	.14	.07
FM	.05	.34	.16	.17	.09	.18$^{\diamond}$	.24	.19	.15	.19	.22	.18
Deep AM-FM	.05	.32	.16	.18	.09	.17$^{\diamond\diamond}$	.26	.21	.14	.19	.24	.18
Agreeableness	-.03	.05	.04	.01	-.01	-.01	.00	.01	.00	.01	.01	.01
Style Matching	.05	-.11	-.04	.05	.17	.06	.06	.08	.13	.09	.10	.06
Emotional Entropy	-.02	.07	.09	.02	.25$^{\diamond}$	.10	.02	.14	.28	.21$^{\diamond\diamond\diamond}$	.12	.12
Empathy	-.02	.08	.03	.01	.03	-.01	.00	.00	-.03	-.01	.00	.01
Emotion Matching	.01	-.03	.08	.03	.05	.00	.07	.11	.13	.11	.13	.06

Table 1: Evaluation on the test data. The first three rows are baseline systems. Reported Spearman ρ for each human evaluation metric: Appropriateness (App.), Content, Grammar, and Relevance). $^{\diamond}$, $^{\diamond\diamond}$, and $^{\diamond\diamond\diamond}$ denote first, second, and third place in column-wise scoring results, respectively (with other teams' scores not included in the results).

	FED-Conversation	Persona-Chatlog
Deep AM-FM	.12	.08
Agreeableness	.27	.03
Style Matching	.07	.08
Emotional Entropy	-.07	.01
Empathy	.11	-.01
Emotion Matching	.03	-.01

Table 2: Evaluation on the two dialog level development data sets. Reported Spearman ρ.

On the other hand, the alternative also suffers from similar concerns, namely that dialog systems may exhibit extremely limited variation in such traits. One could imagine a similar situation to the so-called "Wall Street Journal effect" (i.e., part-of-speech taggers are only accurate when applied to language written by white men; Hovy and Søgaard 2015), where dialog system only converse like middle aged white men. Within our proposed framework, dialog systems should produce agents along a spectrum of such trait level constructs.

References

Abdul-Mageed, M.; Buffone, A.; Peng, H.; Giorgi, S.; Eichstaedt, J. C.; and Ungar, L. H. 2017. Recognizing Pathogenic Empathy in Social Media. In *ICWSM*, 448–451.

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Carr, H.; and Kingsbury, F. 1938. The concept of traits. *Psychological Review*, 45(6): 497.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1870–1879.

Chen, Z.; Sedoc, J.; D'Haro, L. F.; Banchs, R.; and Rudnicky, A. 2021. "Automatic Evaluation and Moderation of Open-domain Dialogue Systems.

Cui, L.; Huang, S.; Wei, F.; Tan, C.; Duan, C.; and Zhou, M. 2017. Superagent: A customer service chatbot for e-commerce websites. In *Proceedings of ACL 2017, System Demonstrations*, 97–102.

Danescu-Niculescu-Mizil, C.; Gamon, M.; and Dumais, S. 2011. Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, 745–754.

Danescu-Niculescu-Mizil, C.; and Lee, L. 2011. Chameleons in Imagined Conversations: A New Approach to Understanding Coordination of Linguistic Style in Dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, 76–87. Portland, Oregon, USA: Association for Computational Linguistics.

Danescu-Niculescu-Mizil, C.; Lee, L.; Pang, B.; and Kleinberg, J. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, 699–708.

Davis, M. H. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1): 113.

Deriu, J.; Rodrigo, A.; Otegi, A.; Echegoyen, G.; Rosset, S.; Agirre, E.; and Cieliebak, M. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54(1): 755–810.

Deriu, J.; Tuggener, D.; von Däniken, P.; Campos, J. A.; Rodrigo, A.; Belkacem, T.; Soroa, A.; Agirre, E.; and Cieliebak, M. 2020. Spot The Bot: A Robust and Efficient Framework for the Evaluation of Conversational Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3971–3984. Online: Association for Computational Linguistics.

D'Haro, L. F.; Banchs, R. E.; Hori, C.; and Li, H. 2019. Automatic evaluation of end-to-end dialog systems with adequacy-fluency metrics. *Computer Speech & Language*, 55: 200–215.

El Asri, L.; Schulz, H.; Sarma, S. K.; Zumer, J.; Harris, J.; Fine, E.; Mehrotra, R.; and Suleman, K. 2017. Frames: a

corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 207–219.

Giles, H. E.; Coupland, J. E.; and Coupland, N. E. 1991. Contexts of accommodation: Developments in applied sociolinguistics.

Giorgi, S.; Ungar, L.; and Schwartz, H. A. 2021. Characterizing Social Spambots by their Human Traits. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 5148–5158. Online: Association for Computational Linguistics.

Gonen, H.; and Goldberg, Y. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614.

Hovy, D. 2015. Demographic factors improve classification performance. 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)*, 752–762.

Hovy, D.; and Søgaard, A. 2015. Tagging performance correlates with author age. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: Short papers)*, 483–488.

Huber, B.; McDuff, D.; Brockett, C.; Galley, M.; and Dolan, B. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 277. ACM.

Ireland, M. E.; Slatcher, R. B.; Eastwick, P. W.; Scissors, L. E.; Finkel, E. J.; and Pennebaker, J. W. 2011. Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science*, 22(1): 39–44. PMID: 21149854.

Kosinski, M.; Stillwell, D.; and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences*, 110(15): 5802–5805.

Li, J.; Galley, M.; Brockett, C.; Spithourakis, G. P.; Gao, J.; and Dolan, B. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Li, M.; Weston, J.; and Roller, S. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

Lin, Z.; Xu, P.; Winata, G. I.; Siddique, F. B.; Liu, Z.; Shin, J.; and Fung, P. 2020. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 13622–13623.

Liu, C.-W.; Lowe, R.; Serban, I. V.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132.

Lord, S. P.; Sheng, E.; Imel, Z. E.; Baer, J.; and Atkins, D. C. 2015. More than reflections: Empathy in motivational interviewing includes language style synchrony between therapist and client. *Behavior therapy*, 46(3): 296–303.

Lynn, V.; Giorgi, S.; Balasubramanian, N.; and Schwartz, H. A. 2019. Tweet Classification without the Tweet: An Empirical Examination of User versus Document Attributes. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, 18–28. Minneapolis, Minnesota: Association for Computational Linguistics.

Manson, J. H.; Bryant, G. A.; Gervais, M. M.; and Kline, M. A. 2013. Convergence of speech rate in conversation predicts cooperation. *Evolution and Human Behavior*, 34(6): 419–426.

Mazaré, P.-E.; Humeau, S.; Raison, M.; and Bordes, A. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

Mehri, S.; and Eskenazi, M. 2020. Unsupervised Evaluation of Interactive Dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 225–235.

Mohammad, S. M.; and Kiritchenko, S. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2): 301–326.

Novick, D.; Afravi, M.; Camacho, A.; Hinojos, L. J.; and Rodriguez, A. E. 2018. Inducing rapport-building behaviors in interaction with an embodied conversational agent. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, 345–346. ACM.

Park, G.; Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Kosinski, M.; Stillwell, D. J.; Ungar, L. H.; and Seligman, M. E. 2015. Automatic personality assessment through social media language. *Journal of personality and social psychology*, 108(6): 934.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001): 2001.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, 3–33. Elsevier.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381.

Roller, S.; Boureau, Y.-L.; Weston, J.; Bordes, A.; Dinan, E.; Fan, A.; Gunning, D.; Ju, D.; Li, M.; Poff, S.; et al. 2020. Open-domain conversational agents: Current progress, open problems, and future directions. *arXiv preprint arXiv:2006.12442*.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; et al. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325.

Sedoc, J.; Ippolito, D.; Kirubarajan, A.; Thirani, J.; Ungar, L.; and Callison-Burch, C. 2019. ChatEval: A Tool for Chatbot Evaluation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 60–65. Association for Computational Linguistics.

See, A.; Roller, S.; Kiela, D.; and Weston, J. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1702–1723.

Shah, D. S.; Schwartz, H. A.; and Hovy, D. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5248–5264.

Volkova, S.; Wilson, T.; and Yarowsky, D. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1815–1827.

Xu, A.; Pathak, E.; Wallace, E.; Gururangan, S.; Sap, M.; and Klein, D. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2390–2397. Online: Association for Computational Linguistics.

Zeigler-Hill, V.; and Shackelford, T. 2020. Encyclopedia of personality and individual differences.

Zhang, C.; D'Haro, L. F.; Banchs, R. E.; Friedrichs, T.; and Li, H. 2021. Deep AM-FM: Toolkit for automatic dialogue evaluation. In *Conversational Dialogue Systems for the Next Decade*, 53–69. Springer.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Zhao, T.; Lala, D.; and Kawahara, T. 2020. Designing Precise and Robust Dialogue Response Evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 26–33.

Zhou, X.; and Wang, W. Y. 2018. MojiTalk: Generating Emotional Responses at Scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1128–1137. Melbourne, Australia: Association for Computational Linguistics.