Personalizing Dialogue Agents: I have a dog, do you have pets too?

Saizheng Zhang[†,1]**, Emily Dinan**[‡]**, Jack Urbanek**[‡]**, Arthur Szlam**[‡]**, Douwe Kiela**[‡]**, Jason Weston**[‡]

[†] Montreal Institute for Learning Algorithms, MILA
[‡] Facebook AI Research
`saizheng.zhang@umontreal.ca`, `{edinan,jju,aszlam,dkiela,jase}@fb.com`

Abstract

Chit-chat models are known to have several problems: they lack specificity, do not display a consistent personality and are often not very captivating. In this work we present the task of making chit-chat more engaging by conditioning on profile information. We collect data and train models to (i) condition on their given profile information; and (ii) information about the person they are talking to, resulting in improved dialogues, as measured by next utterance prediction. Since (ii) is initially unknown, our model is trained to engage its partner with personal topics, and we show the resulting dialogue can be used to predict profile information about the interlocutors.

1 Introduction

Despite much recent success in natural language processing and dialogue research, communication between a human and a machine is still in its infancy. It is only recently that neural models have had sufficient capacity and access to sufficiently large datasets that they appear to generate meaningful responses in a chit-chat setting. Still, conversing with such generic chit-chat models for even a short amount of time quickly exposes their weaknesses (Serban et al., 2016; Vinyals and Le, 2015).

Common issues with chit-chat models include: (i) the lack of a consistent personality (Li et al., 2016a) as they are typically trained over many dialogs each with different speakers, (ii) the lack of an explicit long-term memory as they are typically trained to produce an utterance given only the recent dialogue history (Vinyals and Le, 2015);

---

[1]Work done while at Facebook AI Research.

and (iii) a tendency to produce non-specific answers like "I don't know" (Li et al., 2015). Those three problems combine to produce an unsatisfying overall experience for a human to engage with. We believe some of those problems are due to there being no good publicly available dataset for general chit-chat.

Because of the low quality of current conversational models, and because of the difficulty in evaluating these models, chit-chat is often ignored as an end-application. Instead, the research community has focused on task-oriented communication, such as airline or restaurant booking (Bordes and Weston, 2016), or else single-turn information seeking, i.e. question answering (Rajpurkar et al., 2016). Despite the success of the latter, simpler, domain, it is well-known that a large quantity of human dialogue centers on socialization, personal interests and chit-chat (Dunbar et al., 1997). For example, less than 5% of posts on Twitter are questions, whereas around 80% are about personal emotional state, thoughts or activities, authored by so called "Meformers" (Naaman et al., 2010).

In this work we make a step towards more engaging chit-chat dialogue agents by endowing them with a configurable, but persistent persona, encoded by multiple sentences of textual description, termed a profile. This profile can be stored in a memory-augmented neural network and then used to produce more personal, specific, consistent and engaging responses than a persona-free model, thus alleviating some of the common issues in chit-chat models. Using the same mechanism, any existing information about the persona of the dialogue partner can also be used in the same way. Our models are thus trained to both ask and answer questions about personal topics, and the resulting dialogue can be used to build a model of the persona of the speaking partner.

To support the training of such models, we

2204

present the PERSONA-CHAT dataset, a new dialogue dataset consisting of 164,356 utterances between crowdworkers who were randomly paired and each asked to act the part of a given provided persona (randomly assigned, and created by another set of crowdworkers). The paired workers were asked to chat naturally and to get to know each other during the conversation. This produces interesting and engaging conversations that our agents can try to learn to mimic.

Studying the next utterance prediction task during dialogue, we compare a range of models: both generative and ranking models, including Seq2Seq models and Memory Networks (Sukhbaatar et al., 2015) as well as other standard retrieval baselines. We show experimentally that in either the generative or ranking case conditioning the agent with persona information gives improved prediction of the next dialogue utterance. The PERSONA-CHAT dataset is designed to facilitate research into alleviating some of the issues that traditional chit-chat models face, and with the aim of making such models more consistent and engaging, by endowing them with a persona. By comparing against chit-chat models built using the OpenSubtitles and Twitter datasets, human evaluations show that our dataset provides more engaging models, that are simultaneously capable of being fluent and consistent via conditioning on a persistent, recognizable profile.

2 Related Work

Traditional dialogue systems consist of building blocks, such as dialogue state tracking components and response generators, and have typically been applied to tasks with labeled internal dialogue state and precisely defined user intent (i.e., goal-oriented dialogue), see e.g. (Young, 2000). The most successful goal-oriented dialogue systems model conversation as partially observable Markov decision processes (POMDPs) (Young et al., 2013). All those methods typically do not consider the chit-chat setting and are more concerned with achieving functional goals (e.g. booking an airline flight) than displaying a personality. In particular, many of the tasks and datasets available are constrained to narrow domains (Serban et al., 2015).

Non-goal driven dialogue systems go back to Weizenbaum's famous program ELIZA (Weizenbaum, 1966), and hand-coded systems have continued to be used in applications to this day. For example, modern solutions that build an open-ended dialogue system to the Alexa challenge combine hand-coded and machine-learned elements (Serban et al., 2017a). Amongst the simplest of statistical systems that can be used in this domain, that are based on data rather than hand-coding, are information retrieval models (Sordoni et al., 2015), which retrieve and rank responses based on their matching score with the recent dialogue history. We use IR systems as a baseline in this work.

End-to-end neural approaches are a class of models which have seen growing recent interest. A popular class of methods are generative recurrent systems like seq2seq applied to dialogue (Sutskever et al., 2014; Vinyals and Le, 2015; Sordoni et al., 2015; Li et al., 2016b; Serban et al., 2017b). Rooted in language modeling, they are able to produce syntactically coherent novel responses, but their memory-free approach means they lack long-term coherence and a persistent personality, as discussed before. A promising direction, that is still in its infancy, to fix this issue is to use a memory-augmented network instead (Sukhbaatar et al., 2015; Dodge et al., 2015) by providing or learning appropriate memories.

Serban et al. (2015) list available corpora for training dialogue systems. Perhaps the most relevant to learning chit-chat models are ones based on movie scripts such as OpenSubtitles and Cornell Movie-Dialogue Corpus, and dialogue from web platforms such as Reddit and Twitter, all of which have been used for training neural approaches (Vinyals and Le, 2015; Dodge et al., 2015; Li et al., 2016b; Serban et al., 2017b). Naively training on these datasets leads to models with the lack of a consistent personality as they will learn a model averaged over many different speakers. Moreover, the data does little to encourage the model to engage in understanding and maintaining knowledge of the dialogue partner's personality and topic interests.

According to Serban et al. (2015)'s survey, personalization of dialogue systems is "an important task, which so far has not received much attention". In the case of goal-oriented dialogue some work has focused on the agent being aware of the human's profile and adjusting the dialogue accordingly, but without a personality to the agent itself (Lucas et al., 2009; Joshi et al., 2017). For

the chit-chat setting, the most relevant work is (Li et al., 2016a). For each user in the Twitter corpus, personas were captured via distributed embeddings (one per speaker) to encapsulate individual characteristics such as background information and speaking style, and they then showed using those vectors improved the output of their seq2seq model for the same speaker. Their work does not focus on attempting to engage the other speaker by getting to know them, as we do here. For that reason, our focus is on explicit profile information, not hard-to-interpret latent variables.

3   The PERSONA-CHAT Dataset

The aim of this work is to facilitate more engaging and more personal chit-chat dialogue. The PERSONA-CHAT dataset is a crowd-sourced dataset, collected via Amazon Mechanical Turk, where each of the pair of speakers condition their dialogue on a given profile, which is provided.

The data collection consists of three stages:

(i) Personas: we crowdsource a set of 1155 possible personas, each consisting of at least 5 profile sentences, setting aside 100 never seen before personas for validation, and 100 for test.

(ii) Revised personas: to avoid modeling that takes advantage of trivial word overlap, we crowdsource additional rewritten sets of the same 1155 personas, with related sentences that are rephrases, generalizations or specializations, rendering the task much more challenging.

(iii) Persona chat: we pair two Turkers and assign them each a random (original) persona from the pool, and ask them to chat. This resulted in a dataset of 164,356 utterances over 10,981 dialogs, 15,705 utterances (968 dialogs) of which are set aside for validation, and 15,119 utterances (1000 dialogs) for test.

The final dataset and its corresponding data collection source code, as well as models trained on the data, are all available open source in ParlAI[2].

In the following, we describe each data collection stage and the resulting tasks in more detail.

3.1   Personas

We asked the crowdsourced workers to create a character (persona) description using 5 sentences, providing them only a single example:

---
[2]https://github.com/facebookresearch/ParlAI/tree/master/projects/personachat

"I am a vegetarian. I like swimming. My father used to work for Ford. My favorite band is Maroon5. I got a new job last month, which is about advertising design."

Our aim was to create profiles that are natural and descriptive, and contain typical topics of human interest that the speaker can bring up in conversation. Because the personas are not the real profiles of the Turkers, the dataset does not contain personal information (and they are told specifically not to use any). We asked the workers to make each sentence short, with a maximum of 15 words per sentence. This is advantageous both for humans and machines: if they are too long, crowdsourced workers are likely to lose interest, and for machines the task could become more difficult.

Some examples of the personas collected are given in Table 1 (left).

3.2   Revised Personas

A difficulty when constructing dialogue datasets, or text datasets in general, is that in order to encourage research progress, the task must be carefully constructed so that is neither too easy nor too difficult for the current technology (Voorhees et al., 1999). One issue with conditioning on textual personas is that there is a danger that humans will, even if asked not to, unwittingly repeat profile information either verbatim or with significant word overlap. This may make any subsequent machine learning tasks less challenging, and the solutions will not generalize to more difficult tasks. This has been a problem in some recent datasets: for example, the dataset curation technique used for the well-known SQuAD dataset suffers from this word overlap problem to a certain extent (Chen et al., 2017).

To alleviate this problem, we presented the original personas we collected to a new set of crowdworkers and asked them to rewrite the sentences so that a new sentence is about *"a related characteristic that the same person may have"*, hence the revisions could be rephrases, generalizations or specializations. For example *"I like basketball"* can be revised as *"I am a big fan of Michael Jordan"* not because they mean the same thing but because the same persona could contain both.

In the revision task, workers are instructed not to trivially rephrase the sentence by copying the original words. However, during the entry stage if a non-stop word is copied we issue a warning,

Original Persona	Revised Persona
I love the beach.	To me, there is nothing like a day at the seashore.
My dad has a car dealership	My father sales vehicles for a living.
I just got my nails done	I love to pamper myself on a regular basis.
I am on a diet now	I need to lose weight.
Horses are my favorite animal.	I am into equestrian sports.
I play a lot of fantasy videogames.	RPGs are my favorite genre.
I have a computer science degree.	I also went to school to work with technology.
My mother is a medical doctor	The woman who gave birth to me is a physician.
I am very shy.	I am not a social person.
I like to build model spaceships.	I enjoy working with my hands.

Table 1: Example Personas (left) and their revised versions (right) from the PERSONA-CHAT dataset. The revised versions are designed to be characteristics that the same persona might have, which could be rephrases, generalizations or specializations.

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Table 2: Example dialog from the PERSONA-CHAT dataset. Person 1 is given their own persona (top left) at the beginning of the chat, but does not know the persona of Person 2, and vice-versa. They have to get to know each other during the conversation.

and ask them to rephrase, guaranteeing that the instructions are followed. For example, *"My father worked for Ford."* can be revised to *"My dad worked in the car industry"*, but not *"My dad was employed by Ford."* due to word overlap.

Some examples of the revised personas collected are given in Table 1 (right).

3.3 Persona Chat

After collecting personas, we then collected the dialogues themselves, conditioned on the personas. For each dialogue, we paired two random crowdworkers, and gave them the instruction that they will chit-chat with another worker, while playing the part of a given character. We then provide them with a randomly chosen persona from our pool, different to their partners. The instructions are on purpose quite terse and simply ask them to "chat with the other person naturally and try to get to know each other". In an early study we noticed the crowdworkers tending to talk about themselves (their own persona) too much, so we also added the instructions "both ask questions and answer questions of your chat partner" which seemed to help. We also gave a bonus for high quality dialogs. The dialog is turn-based, with a maximum of 15 words per message. We again gave instructions to not trivially copy the character descriptions into the messages, but also wrote explicit code sending them an error if they tried to do so, using simple string matching. We define a minimum dialogue length which is randomly between 6 and 8 turns each for each dialogue. An example dialogue from the dataset is given in Table 2.

3.4 Evaluation

We focus on the standard dialogue task of predicting the next utterance given the dialogue history, but consider this task both with and without the profile information being given to the learning agent. Our goal is to enable interesting directions for future research, where chatbots can for instance have personalities, or imputed personas could be used to make dialogue more engaging to the user.

We consider this in four possible scenarios: conditioning on no persona, your own persona, their persona, or both. These scenarios can be tried using either the original personas, or the revised ones. We then evaluate the task using three metrics: (i) the log likelihood of the correct sequence, measured via perplexity, (ii) F1 score, and (iii) next utterance classification loss, following Lowe et al. (2015). The latter consists of choosing N random distractor responses from other dialogues (in our setting, $N=19$) and the model selecting the best response among them, resulting in a score of one if the model chooses the correct response, and zero otherwise (called hits@1 in the experiments).

4 Models

We consider two classes of model for next utterance prediction: ranking models and generative models. Ranking models produce a next utterance by considering any utterance in the training set as a possible candidate reply. Generative models generate novel sentences by conditioning on the dialogue history (and possibly, the persona), and then generating the response word-by-word. Note one can still evaluate the latter as ranking models by computing the probability of generating a given candidate, and ranking candidates by those scores.

4.1 Baseline ranking models

We first consider two baseline models, an IR baseline (Sordoni et al., 2015) and a supervised embedding model, Starspace (Wu et al., 2017)[3]. While there are many IR variants, we adopt the simplest one: find the most similar message in the (training) dataset and output the response from that exchange. Similarity is measured by the tf-idf weighted cosine similarity between the bags of words. Starspace is a recent model that also performs information retrieval but by learning the

---

[3] github.com/facebookresearch/StarSpace

similarity between the dialog and the next utterance by optimizing the embeddings directly for that task using the margin ranking loss and k-negative sampling. The similarity function $sim(q, c')$ is the cosine similarity of the sum of word embeddings of the query q and candidate c'. Denoting the dictionary of $\mathcal{D}$ word embeddings as W which is a $\mathcal{D} \times d$ matrix, where W_i indexes the ith word (row), yielding its d-dimensional embedding, it embeds the sequences q and c'.

In both methods, IR and StarSpace, to incorporate the profile we simply concatenate it to the query vector bag of words.

4.2 Ranking Profile Memory Network

Both the previous models use the profile information by combining it with the dialogue history, which means those models cannot differentiate between the two when deciding on the next utterance. In this model we instead use a memory network with the dialogue history as input, which then performs attention over the profile to find relevant lines from the profile to combine with the input, and then finally predicts the next utterance. We use the same representation and loss as in the Starspace model, so without the profile, the two models are identical. When the profile is available attention is performed by computing the similarity of the input q with the profile sentences p_i, computing the softmax, and taking the weighted sum:

$$q^+ = q + \sum s_i p_i, \qquad s_i = \text{Softmax}(sim(q, p_i))$$

where $\text{Softmax}(z_i) = e^{z_i} / \sum_j e^{z_j}$. One can then rank the candidates c' using $sim(q^+, c')$. One can also perform multiple "hops" of attention over the profile rather than one, as shown here, although that did not bring significant gains in our parameter sweeps.

4.3 Key-Value Profile Memory Network

The key-value (KV) memory network (Miller et al., 2016) was proposed as an improvement to the memory network by performing attention over keys and outputting the values (instead of the same keys as in the original), which can outperform memory networks dependent on the task and definition of the key-value pairs. Here, we apply this model to dialogue, and consider the keys as dialog histories (from the training set), and the values as the next dialogue utterances, i.e., the replies from the speaking partner. This allows the model

to have a memory of past dialogues that it can directly use to help influence its prediction for the current conversation. The model we choose is identical to the profile memory network just described in the first hop over profiles, while in the second hop, q^+ is used to attend over the keys and output a weighted sum of values as before, producing q^{++}. This is then used to rank the candidates c' using $sim(q^{++}, c')$ as before. As the set of (key-value) pairs is large this would make training very slow. In our experiments we simply trained the profile memory network and used the same weights from that model and applied this architecture at test time instead. Training the model directly would presumably give better results, however this heuristic already proved beneficial compared to the original network.

4.4 Seq2Seq

The input sequence x is encoded by applying $h_t^e = LSTM_{enc}(x_t \mid h_{t-1}^e)$. We use GloVe (Pennington et al., 2014) for our word embeddings. The final hidden state, h_t^e, is fed into the decoder $LSTM_{dec}$ as the initial state h_0^d. For each time step t, the decoder then produces the probability of a word j occurring in that place via the softmax, i.e.,

$$p(y_{t,j} = 1 \mid y_{t-1}, \ldots, y_1) = \frac{\exp(w_j h_t^d)}{\sum_{j'=1}^{K} \exp(w_{j'} h_t^d)}.$$

The model is trained via negative log likelihood. The basic model can be extended to include persona information, in which case we simply prepend it to the input sequence x, i.e., $x = \forall p \in P \mid\mid x$, where $\mid\mid$ denotes concatenation. For the OpenSubtitles and Twitter datasets trained in Section 5.2 we found training a language model (LM), essentially just the decoder part of this model, worked better and we report that instead.

4.5 Generative Profile Memory Network

Finally, we introduce a generative model that encodes each of the profile entries as individual memory representations in a memory network. As before, the dialogue history is encoded via $LSTM_{enc}$, the final state of which is used as the initial hidden state of the decoder. Each entry $p_i = \langle p_{i,1}, \ldots, p_{i,n} \rangle \in P$ is then encoded via $f(p_i) = \sum_j^{|p_i|} \alpha_i p_{i,j}$. That is, we weight words by their inverse term frequency: $\alpha_i = 1/(1 + log(1 + \text{tf}))$ where tf is computed from the GloVe index via

Zipf's law[4]. Let F be the set of encoded memories. The decoder now attends over the encoded profile entries, i.e., we compute the mask a_t, context c_t and next input $\hat{x}_t$ as:

$$a_t = softmax(FW_a h_t^d),$$
$$c_t = a_t^\intercal F; \quad \hat{x}_t = tanh(W_c[c_{t-1}, x_t]).$$

If the model has no profile information, and hence no memory, it becomes equivalent to the Seq2Seq model.

5 Experiments

We first report results using automated evaluation metrics, and subsequently perform an extrinsic evaluation where crowdsourced workers perform a human evaluation of our models.

5.1 Automated metrics

The main results are reported in Table 3. Overall, the results show the following key points:

Persona Conditioning Most models improve significantly when conditioning prediction on their own persona at least for the original (non-revised) versions, which is an easier task than the revised ones which have no word overlap. For example, the Profile Memory generation model has improved perplexity and hits@1 compared to Seq2Seq, and all the ranking algorithms (IR baseline, Starspace and Profile Memory Networks) obtain improved hits@1.

Ranking vs. Generative. Ranking models are far better than generative models at ranking. This is perhaps obvious as that is the metric they are optimizing, but still the performance difference is quite stark. It may be that the word-based probability which generative models use works well, but is not calibrated well enough to give a sentence-based probability which ranking requires. Human evaluation is also used to compare these methods, which we perform in Sec. 5.2.

Ranking Models. For the ranking models, the IR baseline is outperformed by Starspace due to its learnt similarity metric, which in turn is outperformed by Profile Memory networks due to the attention mechanism over the profiles (as all other parts of the models are the same). Finally KV Profile Memory networks outperform Profile Memory Networks in the no persona case due to the ability to consider neighboring dialogue history and next

---

[4]tf $= 1e6 * 1/(idx^{1.07})$

Method	No Persona		Original Persona		Revised Persona	
	ppl	hits@1	ppl	hits@1	ppl	hits@1
Generative Models						
Seq2Seq	38.08	0.092	40.53	0.084	40.65	0.082
Profile Memory	38.08	0.092	34.54	0.125	38.21	0.108
Ranking Models						
IR baseline	-	0.214	-	0.410	-	0.207
Starspace	-	0.318	-	0.491	-	0.322
Profile Memory	-	0.318	-	0.509	-	0.354
KV Profile Memory	-	0.349	-	0.511	-	0.351

Table 3: **Evaluation of dialog utterance prediction with various models** in three settings: without conditioning on a persona, conditioned on the speakers given persona ("Original Persona"), or a revised persona that does not have word overlap.

Method					Persona
Model	Profile	**Fluency**	**Engagingness**	**Consistency**	**Detection**
Human	Self	4.31(1.07)	4.25(1.06)	4.36(0.92)	0.95(0.22)
Generative PersonaChat Models					
Seq2Seq	None	3.17(1.10)	3.18(1.41)	2.98(1.45)	0.51(0.50)
Profile Memory	Self	3.08(1.40)	3.13(1.39)	3.14(1.26)	0.72(0.45)
Ranking PersonaChat Models					
KV Memory	None	3.81(1.14)	3.88(0.98)	3.36(1.37)	0.59(0.49)
KV Profile Memory	Self	3.97(0.94)	3.50(1.17)	3.44(1.30)	0.81(0.39)
Twitter LM	None	3.21(1.54)	1.75(1.04)	1.95(1.22)	0.57(0.50)
OpenSubtitles 2018 LM	None	2.85(1.46)	2.13(1.07)	2.15(1.08)	0.35(0.48)
OpenSubtitles 2009 LM	None	2.25(1.37)	2.12(1.33)	1.96(1.22)	0.38(0.49)
OpenSubtitles 2009 KV Memory	None	2.14(1.20)	2.22(1.22)	2.06(1.29)	0.42(0.49)

Table 4: **Human Evaluation** of various PERSONA-CHAT models, along with a comparison to human performance, and Twitter and OpenSubtitles based models (last 4 rows), standard deviation in parenthesis.

utterance pairs in the training set that are similar to the current dialogue, however when using persona information the performance is similar.

Revised Personas. Revised personas are much harder to use. We do however still see some gain for the Profile Memory networks compared to none (0.354 vs. 0.318 hits@1). We also tried two variants of training: with the original personas in the training set or the revised ones, a comparison of which is shown in Table 6 of the Appendix. *Training* on revised personas helps, both for test examples that are in original form or revised form, likely due to the model be forced to learn more than simple word overlap, forcing the model to generalize more (i.e., learn semantic similarity of differing phrases).

Their Persona. We can also condition a model on the other speaker's persona, or both personas at once, the results of which are in Tables 5 and 6 in the Appendix. Using "Their persona" has less impact on this dataset. We believe this is because most speakers tend to focus on themselves when it comes to their interests. It would be interesting how often this is the case in other datasets. Certainly this is skewed by the particular instructions one could give to the crowdworkers. For example if we gave the instructions "try not to talk about yourself, but about the other's interests' likely these metrics would change.

5.2 Human Evaluation

As automated metrics are notoriously poor for evaluating dialogue (Liu et al., 2016) we also perform human evaluation using crowdsourced workers. The procedure is as follows. We perform almost exactly the same setup as in the dataset collection process itself as in Section 3.3. In that setup, we paired two Turkers and assigned them each a random (original) persona from the collected pool, and asked them to chat. Here, from the Turker's point of view everything looks the same except instead of being paired with a Turker they are paired with one of our models instead (they do not know this). In this setting, for both the Turker and the model, the personas come from the test set pool.

After the dialogue, we then ask the Turker some additional questions in order to evaluate the quality of the model. We ask them to evaluate fluency, engagingness and consistency (scored between 1-5). Finally, we measure the ability to detect the other speaker's profile by displaying two possible profiles, and ask which is more likely to be the profile of the person the Turker just spoke to. More details of these measures are given in the Appendix.

The results are reported in Table 4 for the best performing generative and ranking models, in both the No Persona and Self Persona categories, 100 dialogues each. We also evaluate the scores of human performance by replacing the chatbot with a human (another Turker). This effectively gives us upper bound scores which we can aim for with our models. Finally, and importantly, we compare our models trained on PERSONA-CHAT with chit-chat models trained with the Twitter and OpenSubtitles datasets (2009 and 2018 versions) instead, following Vinyals and Le (2015). Example chats from a few of the models are shown in the Appendix in Tables 7, 8, 9, 10, 11 and 12.

Firstly, we see a difference in fluency, engagingness and consistency between all PERSONA-CHAT models and the models trained on OpenSubtitles and Twitter. PERSONA-CHAT is a resource that is particularly strong at providing training data for the beginning of conversations, when the two speakers do not know each other, focusing on asking and answering questions, in contrast to other resources. We also see suggestions of more subtle differences between the models, although these differences are obscured by the high variance of the human raters' evaluations. For example, in both the generative and ranking model cases, models endowed with a persona can be detected by the human conversation partner, as evidenced by the persona detection accuracies, whilst maintaining fluency and consistency compared to their non-persona driven counterparts.

Finding the balance between fluency, engagement, consistency, and a persistent persona remains a strong challenge for future research.

5.3 Profile Prediction

Two tasks could naturally be considered using PERSONACHAT: (1) next utterance prediction during dialogue, and (2) profile prediction given dialogue history. The main study of this work has been Task 1, where we have shown the use of profile information. Task 2, however, can be used to extract such information. While a full study is beyond the scope of this paper, we conducted some preliminary experiments, the details of which are in Appendix D. They show (i) human speaker's profiles can be predicted from their dialogue with high accuracy (94.3%, similar to human performance in Table 4) or even from the model's dialogue (23% with KV Profile Memory) showing the model is paying attention to the human's interests. Further, the accuracies clearly improve with further dialogue, as shown in Table 14. Combining Task 1 and Task 2 into a full system is an exciting area of future research.

6 Conclusion & Discussion

In this work we have introduced the PERSONA-CHAT dataset, which consists of crowd-sourced dialogues where each participant plays the part of an assigned persona; and each (crowd-sourced) persona has a word-distinct paraphrase. We test various baseline models on this dataset, and show that models that have access to their own personas in addition to the state of the dialogue are scored as more consistent by annotators, although not more engaging. On the other hand, we show that models trained on PERSONA-CHAT (with or without personas) are more engaging than models trained on dialogue from other resources (movies, Twitter).

We believe PERSONA-CHAT will be a useful resource for training components of future dialogue systems. Because we have paired human generated profiles and conversations, the data aids the construction of agents that have consistent per-

sonalities and viewpoints. Furthermore, predicting the profiles from a conversation moves chitchat tasks in the direction of goal-directed dialogue, which has metrics for success. Because we collect paraphrases of the profiles, they cannot be trivially matched; indeed, we believe the original and rephrased profiles are interesting as a semantic similarity dataset in their own right. We hope that the data will aid training agents that can ask questions about users' profiles, remember the answers, and use them naturally in conversation.

References

Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*.

Robin IM Dunbar, Anna Marriott, and Neil DC Duncan. 1997. Human conversational behavior. *Human nature*, 8(3):231–246.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016b. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. *arXiv preprint arXiv:1506.08909*.

JM Lucas, F Fernández, J Salazar, J Ferreiros, and R San Segundo. 2009. Managing speaker identity and user profiles in a spoken dialogue system. *Procesamiento del Lenguaje Natural*, 43:77–84.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. *arXiv preprint arXiv:1606.03126*.

Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017a. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2015. A survey of available corpora for building data-driven dialogue systems. *arXiv preprint arXiv:1512.05742*.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Ellen M Voorhees et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82.

Joseph Weizenbaum. 1966. Elizaa computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2017. Starspace: Embed all the things! *arXiv preprint arXiv:1709.03856*.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Steve J Young. 2000. Probabilistic methods in spoken–dialogue systems. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 358(1769):1389–1402.