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

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, & Jason Weston
Facebook AI Research

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 & 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 & Le, 2015); 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[1]. For those reasons, chit-chat models are often ignored as an end-application and the research community has focused on task-oriented communication, such as airline or restaurant booking, instead (Bordes & 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.

We thus 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

[1]For example, currently the most general chit-chat dataset available in `http://parl.ai` a large repository of dialogue datasets is probably OpenSubtitles, which is based on movie scripts, not natural conversations.

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 dataset, 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 dialog systems consist of building blocks, such as dialog state tracking components and response generators, and have typically been applied to tasks with labeled internal dialog state and precisely defined user intent (i.e., goal-oriented dialogue), see e.g. (Young, 2000). The most successful goal-oriented dialog 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 dialog 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 dialog 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 & Le, 2015; Sordoni et al., 2015; Li et al., 2016b; Serban et al., 2017b). Their strengths are that (i) they are not constrained by hard-code rules or explicit internal states that may work well in a narrow domain, but are too restrictive for more open dialogue such as chit-chat, and (ii) being based on architectures rooted in language modeling and machine translation, they excel at generating syntactically coherent language, and can generate entirely novel responses. Their deficiencies are that they typically need a large amount of data to be trained, and the vanilla approach generates responses given only the recent dialog history without using other external memory. The latter issue makes neural models hence typically lack both domain knowledge in the domain being discussed, and a persistent personality during discussions. 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) and either provide or learn appropriate external memories. A related class of neural methods is to use similarly architectures, but to retrieve and rank candidates similarly to the IR baseline, but using memory-augmented networks to score the candidates instead. We compare the generative and ranking approaches to each other in this work.

Serban et al. (2015) list available corpora for training dialog 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 & 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.

Original Persona	Revised Persona
I love the beach. My dad has a car dealership I just got my nails done I am on a diet now Horses are my favorite animal.	To me, there is nothing like a day at the seashore. My father sales vehicles for a living. I love to pamper myself on a regular basis. I need to lose weight. I am into equestrian sports.
I am an eccentric hair stylist for dogs My favorite past time is collecting Civil War antiques. I fake a British accent to seem more attractive. I have been married four times and widowed three. I have an allergy to mangoes	I work with animals. I like finding or buying historical artifacts. I heard girls liked foreigners. I have a lot of experience with marriage I have reactions to certain fruits.
I play a lot of fantasy videogames. I have a computer science degree. My mother is a medical doctor I am very shy. I like to build model spaceships.	RPGs are my favorite genre. I also went to school to work with technology. The woman who gave birth to me is a physician. I am not a social person. 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.

According to the survey (Serban et al., 2015) personalization of dialogue systems is "an important task, which so far has not received much attention". In the case of goal-oriented dialog 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 where each of the pair of speakers condition their dialogue on a given profile, which is provided. The data collection consists of three stages:

- 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.
- 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.
- 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 is available in ParlAI[2]. In the following, we describe each data collection stage in more detail.

[2] https://github.com/facebookresearch/ParlAI/tree/master/parlai/tasks/personachat

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.

3.1 PERSONAS

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

"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. 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 to encourage research progress requires the careful construction of a task 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 crowd-workers 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, 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.

Finally, we encourage the construction of natural sentences. In earlier versions of the task we noticed that the word overlap constraint caused unwanted unnatural constructions such as *"I like eating pretzels"* revised as *"I like to chew and swallow twisted bread with salt"*. Giving explicit instructions about this seemed to help, where we prefer a revision like *"I enjoy beers and beer snacks"*.

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 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 person, your own persona, their person, or both. We can also try each of these scenarios using either the original personas, or the revised ones. We then evaluate the task using two metrics: (i) the log likelihood of the correct sequence, measured via perplexity and (ii) next utterance classification loss, following Lowe et al. (2015).

As dialogue has many possible responses, leading to a multi-modal distribution of words, word overlap measures do not work well as evaluation metrics (Liu et al., 2016; Serban et al., 2015). While word level perplexity has many deficiencies as a measure of conversational success, it is standard in more general language modeling, and can still capture multi-modal distributions to a certain extent as good response word choices should still have high probability. Thus we include it here. Next utterance classification loss consists of choosing N random distractor responses from other dialogues (in our setting, N=19) and choosing the best among them, resulting in a score of one if the model chooses the correct response, and zero otherwise. Its main advantage is that it is easy to interpret.

4 MODELS

Let x be an input sequence (i.e., the previous dialogue utterances), M^1 be the set of profile entries of the speaker (i.e., the model's own profile) and M^2 be the profile of the listener (i.e., the interlocutor's profile). In our experiments, we compare four settings: not having access to any profile information, having access only to M^1, having access only to M^2, or having access to both $M^1 \cup M^2$. In what follows, we use $M \in \{\emptyset, M^1, M^2, M^1 \cup M^2\}$ to cover all four possibilities.

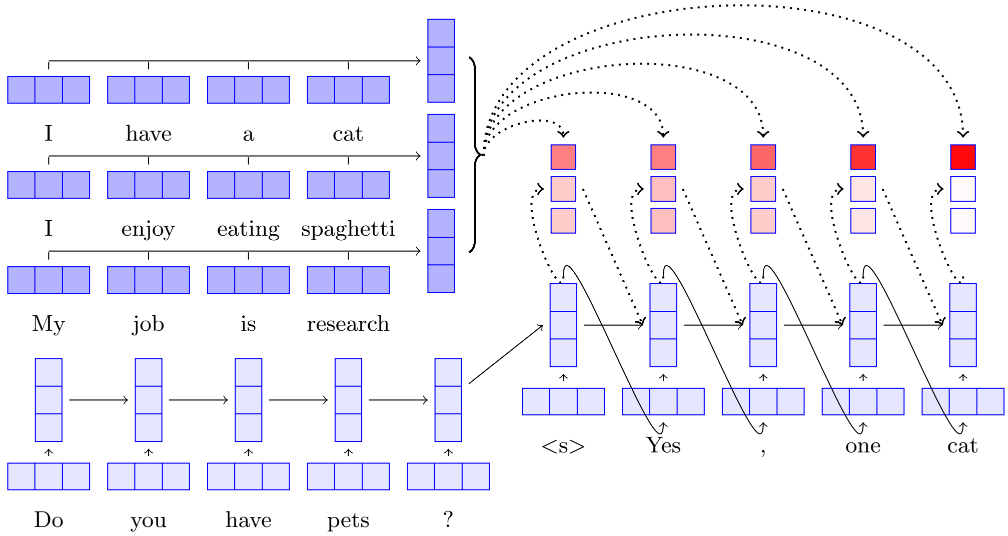

Figure 1: A diagram of the Profile Memory Network for generation. We also implemented a ranking version which has the same architecture except it ranks candidate sentences from the training set instead of generating, representing them using bag-of-word embeddings.

4.1 Ranking Models

The following set of models produce a next utterance by considering any utterance in the training set as a possible candidate reply. They are typically strong baselines, or, if the candidate set is big enough can be hard to beat as the sentences, being written by humans, already have fluency and internal semantic coherence. On the other hand, they cannot generate novel sentences.

4.1.1 IR Baseline

To select candidate responses a standard baseline is nearest neighbor information retrieval (IR) (Isbell et al., 2000; Jafarpour et al., 2010; Ritter et al., 2011; Sordoni et al., 2015). While there are many 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. To incorporate the profile we simply concatenate it to the query vector bag of words.

4.1.2 Starspace

Starspace is a recent model that performs also performs information retrieval but by learning sentence embeddings that measure similarity between the dialog and the next utterance by optimizing the word embeddings directly for that task using the training set (Wu et al., 2017)[3]. Similar supervised embeddings have been used with good results in other dialogue tasks previously (Dodge et al., 2015). Specifically, it optimizes:

$$\sum_{\substack{(q,c)\in E^+ \\ b^-\in E^-}} L^{batch}(sim(q,c), sim(q,c_1^-), \ldots, sim(q,c_k^-))$$

where the loss function L_{batch} compares a positive pair of query and candidate (q, c) with the negative pairs (q, c_i^-), $i = 1, \ldots, k$ using the margin ranking loss $\max(0, \mu - sim(q, c)$, where μ is the margin parameter. The similarity function $sim(\cdot, \cdot)$ 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

[3]Available at `https://github.com/facebookresearch/StarSpace`

which is a $\mathcal{D} \times d$ matrix, where W_i indexes the i^{th} word (row), yielding its d-dimensional embedding, it embeds a sequence s with $\sum_{i \in s} W_i$. While this model supports different word embeddings for the left and right hand side of the similarity function, we found sharing the weights gave the best performance.

Similar to the IR baseline, to incorporate the profile we simply concatenate it to the query vector bag of words.

4.1.3 RANKING PROFILE MEMORY NETWORK

Both the previous models use the profile information by combining it with the dialogue history, which means the model 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 next utterance. We use the same representation 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. Similarly, we found a model that shared word embedding lookup table weights across dialog history, profiles and candidates performed best compared to models with more parameters.

4.1.4 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, e.g. 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 current 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.

4.2 GENERATIVE MODELS

Our next set of models do generate novel sentences by conditioning on the dialogue history and possibly the persona, and then generating the response word-by-word, see e.g. Fig 1. One can still evalutate these models as ranking models by computing the probability of generating a given candidate, and ranking candidates by those scores.[4]

4.2.1 SEQ2SEQ

The input sequence 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 $LSTM_{dec}$

[4]In practice, better results for ranking are obtained by normalizing by the sentence length, following (Dodge et al., 2015).

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.,

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

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., $x = \forall m \in M \mid\mid x$, where || denotes concatenation.

4.2.2 Generative Profile Memory Network

Finally, we introduce a 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 $m_i = \langle m_{i,1}, \ldots, m_{i,n} \rangle \in M$ is then encoded via $f(m_i) = \sum_j^{|m_i|} \alpha_i m_{i,j}$. 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[5]. 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:

$$a_t = softmax(FW_a h_t^d); c_t = a_t^\intercal F; \hat{x}_t = tanh(W_c[c_{t-1}, x_t]). \tag{2}$$

This model is illustrated in Figure 1. Again, 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 automatic evaluation metrics, and subsequently perform an extrinsic evaluation where we use crowdsourced workers to perform a human evaluation of our models.

5.1 Automated metrics

Results for the generative model approaches are reported in Table 3, and for the ranking models in Table 4. For the generative models, we report perplexity and hits@1 (the accuracy of the next dialogue utterance when choosing between the gold response and N=19 distractor responses).To compute hits@1 for generative models we rank candidates according to their mean log likelihood. For ranking models, which are not generative and hence do not allow for computing the perplexity, we only report hits@1.

In all cases we compare using the different persona types (none, my, their and both) and using the original or revised versions. For the ranking models we also tried two variants of training: training with the original personas in the training set or the revised ones. The latter could provide a difference because there is less word overlap between the dialogue and the profiles in that case which can force the model to generalize more (e.g. learn synonyms) rather than learning about word overlap, which crowdsource workers may otherwise resort to.

Overall, the results show the following key points:

- Most models improve significantly when conditioning prediction on their persona ('Self 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.
- 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

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

Persona	Method	Original		Revised	
		Perplexity	Hits@1	Perplexity	Hits@1
No Persona		38.08	0.092	38.08	0.092
Self Persona	Seq2Seq	40.53	0.084	40.65	0.082
	Profile Memory	**34.54**	**0.125**	38.21	**0.108**
Their Persona	Seq2Seq	41.48	0.075	41.95	0.074
	Profile Memory	36.42	0.105	**37.75**	0.103
Both Personas	Seq2Seq	40.14	0.084	40.53	0.082
	Profile Memory	35.27	0.115	38.48	0.106

Table 3: **Evaluation of dialog utterance prediction with generative models** in four settings: conditioned on the speakers persona ("self persona"), the dialogue partner's persona ("their persona"), both or none. The personas are either the original source given to Turkers to condition the dialogue, or the revised personas that do not have word overlap. In the "no persona" setting, the models are equivalent, so we only report once.

Method	No Persona		Self Persona		Their Persona		Both Personas	
	Orig	Rewrite	Orig	Rewrite	Orig	Rewrite	Orig	Rewrite
IR baseline	0.214	0.214	0.410	0.207	0.181	0.181	0.382	0.188
Training on original personas								
Starspace	0.318	0.318	0.481	0.295	0.245	0.235	0.429	0.258
Profile Memory	0.318	0.318	0.473	0.302	0.283	0.267	0.438	0.266
Training on revised personas								
Starspace	0.318	0.318	0.491	0.322	0.271	0.261	0.432	0.288
Profile Memory	0.318	0.318	0.509	0.354	0.299	0.294	0.467	0.331
KV Profile Memory	0.349	0.349	0.511	0.351	0.291	0.289	0.467	0.330

Table 4: **Evaluation of dialog utterance prediction with ranking models** using hits@1 in four settings: conditioned on the speakers persona ("self persona"), the dialogue partner's persona ("their persona"), both or none. The personas are either the original source given to Turkers to condition the dialogue, or the rewritten personas that do not have word overlap, explaining the poor performance of IR in that case.

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.

- Revised personas are much harder to use. We do however still see some gain for the Profile Memory networks using "Self persona" compared to none (0.354 vs. 0.318 hits@1). 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.
- 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. Due to this inherent unfairness in the automatic evaluation, a more fair measure is a human evaluation that compares these methods, which we perform in Sec. 5.2.
- 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 utterance pairs in the training set that are similar to the current dialogue, however when using persona information the performance is similar.

Method Model	Profile	Fluency	Engagingness	Consistency	Persona Detection
Human	Self	4.31(1.07)	4.25(1.06)	4.36(0.92)	0.95(0.22)
Generative 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 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)
OpenSubtitles KV Memory	None	2.14(1.20)	2.22(1.22)	2.06(1.29)	0.42(0.49)

Table 5: **Human Evaluation** of our various PERSONA-CHAT model, along with a comparison to human performance, and OpenSubtitles based model (last row), standard deviation in parenthesis.

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. They are, in order:

- **Fluency**: We ask them to judge the fluency of the other speaker as a score from 1 to 5, where 1 is "not fluent at all", 5 is "extremely fluent", and 3 is "OK".
- **Engagingness**: We ask them to judge the engagingness of the other speaker *disregarding fluency* from 1-5, where 1 is "not engaging at all", 5 is "extremely engaging", and 3 is "OK".
- **Consistency**: We ask them to judge the consistency of the persona of the other speaker, where we give the example that "I have a dog" followed by "I have no pets" is not consistent. The score is again from 1-5.
- **Profile Detection**: Finally, we display two possible profiles, and ask which is more likely to be the profile of the person the Turker just spoke to. One profile is chosen at random, and the other is the true persona given to the model.

The results are reported in Table 5 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 a chit-chat model trained with the OpenSubtitles dataset instead, following Vinyals & Le (2015). Example chats from a few of the models are shown in Tables 7, 6 and 8.

Firstly, we see a difference in fluency, engagingness and consistency between all PERSONA-CHAT models and the model trained on OpenSubtitles. 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. [6] For example, in both the generative and ranking model cases, models endowed with a persona can be detected by the human

[6]We plan to increase the resolution of our evaluation with more rater data in future work.

Persona 1	**Persona 2**
I am a doctor	I eat sardines for breakfast daily
I have one brother	I just had surgery
My mom is a secretary	I love playing Frisbee
I like to eat a lot of meat	I have two large dogs
I am a bodybuilder	I never finished high school

[PERSON 1:] Hi I am Dr. Jenson. What is your name?
[PERSON 2:] Miss Dawes.
[PERSON 1:] Hi Miss Dawes, tell me about your family!
[PERSON 2:] Miss Baggoli, I know it must show on your records that my name is Mary, but everyone has referred to me as Lola since I was a squealing infant in my mother' s arms.
[PERSON 1:] fascinating. My secretary mother would have made note of all of that.
[PERSON 2:] In addition to the remainder, it writes my autobiography.
[PERSON 1:] Oh yeah? well I enjoy a nice steak. how about you?
[PERSON 2:] Father always used to say that if we ever had the money you have...
we' d have steak and ice cream three times every day!
[PERSON 1:] I eat all the steak and chicken too, even bacon
[PERSON 2:] Father always used to say that if we ever had the money you have...
we' d have steak and ice cream three times every day!
[PERSON 1:] yeah you said that already
[PERSON 2:] Dynamic structure can absorb shock.
[PERSON 1:] yes it can. I know about that as a bodybuilder
[PERSON 2:] Ladies and gentlemen... the strongest man in the world is blowing up a hot water bottle.

Table 6: Example dialog between a human (Person 1) and the OpenSubtitles KV Memory Network model (Person 2).

Persona 1	**Persona 2**
I am afraid of dogs	I am a carpenter
I used to work at a carnival	I used to drink but I stopped five years ago
I like to drink scotch to relax	I have a wife and three kids
I write short stories for a living	I like playing poker
I like dark superhero movies	I dont have many friends

[PERSON 1:] Hi are you relaxing on this fine evening?
[PERSON 2:] I am doing good and you?
[PERSON 1:] I am well, just drinking some scotch and working on my writing
[PERSON 2:] I love to cook and bake but I rarely make time for it
[PERSON 1:] what do you usually bake?
[PERSON 2:] I do not cook, I bake cakes at a local restaurant. how about you?
[PERSON 1:] I used to be a juggler at a carnival but now I am an author
[PERSON 2:] I am a dancer, what do you do?
[PERSON 1:] I just told you. Do you have any pets?
[PERSON 2:] no, I am allergic to pets unfortunately.
[PERSON 1:] I understand. I am not allergic but I am terrified of dogs!
[PERSON 2:] I have a cat. she has been with me through three wives, haha.
[PERSON 1:] Funny that you just told me you have no pets
[PERSON 2:] I do not have any animals.

Table 7: Example dialog between a human (Person 1) and the Seq2Seq model (Person 2).

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.

Persona 1	Persona 2
I am engaged to be married	I am a writer
I used to play sports in high school	I write romance novels
My favorite sport is soccer	I live on a faraway island
I am a college graduate	I am an introvert
I love watching sports on TV	I have two cats

[PERSON 1:] I just got done playing soccer. What have you been up to?
[PERSON 2:] I do not have pets. I spend my free time writing. what about you?
[PERSON 1:] Just graduated from college. Looking for work.
[PERSON 2:] yes I graduated, in writing and poetry
[PERSON 1:] Have you had anything published?
[PERSON 2:] I mainly write fiction novels.
[PERSON 1:] Do you like to watch sports?
[PERSON 2:] do you like kings of leon my favorite by them is use somebody
[PERSON 1:] Are you married? I will be married soon.
[PERSON 2:] haha, no time. I have got a novel to finish.
[PERSON 1:] What is your novel about?
[PERSON 2:] it is a love story. I really enjoy writing. do you write or read?
[PERSON 1:] Like to read. Spend most time watching football.
[PERSON 2:] yes I do I am and author for a living

Table 8: Example dialog between a human (Person 1) and the Key-Value Profile Memory Network with Self Persona.

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 movies.

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 personalities and viewpoints. Furthermore, imputing the profiles from a conversation moves chit-chat 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. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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