Aiming to Know You Better Perhaps Makes Me a More Engaging Dialogue Partner

Yury Zemlyanskiy
University of Southern California
Los Angeles, CA 90089
`yury.zemlyanskiy@usc.edu`

Fei Sha
Netflix
Los Angeles, CA 90028
`fsha@netflix.com`*

Abstract

There have been several attempts to define a plausible motivation for a chit-chat dialogue agent that can lead to engaging conversations. In this work, we explore a new direction where the agent specifically focuses on discovering information about its interlocutor. We formalize this approach by defining a quantitative metric. We propose an algorithm for the agent to maximize it. We validate the idea with human evaluation where our system outperforms various baselines. We demonstrate that the metric indeed correlates with the human judgments of engagingness.

1 Introduction

There has been a significant progress in creating end-to-end data-driven dialogue systems (Ritter et al., 2011; Vinyals and Le, 2015; Serban et al., 2017; Shang et al., 2015; Sordoni et al., 2015). The general scheme is to view dialogues as a sequence transduction process. This process is then modeled with the sequence-to-sequence (SEQ2SEQ) neural network (Sutskever et al., 2014) whose parameters are fit on large dialogue corpora such as OpenSubtitles (Tiedemann, 2009). What is especially appealing about these systems is that they do not require hand-crafted rules to generate reasonable responses in the open-domain dialogue (i.e., chit-chat) setting.

An important goal of such systems is to be able to have a meaningful and engaging conversation with a real person. Despite the progress, however, this goal remains elusive — current systems often generate generic and universally applicable responses (to any questions) such as *"I do not know"*. While such responses are reasonable in isolation, collectively too many of them are perceived as dull and repetitive (Sordoni et al., 2015; Serban et al., 2017; Li et al., 2016a,b).

It remains open what metrics to use to optimize a data-driven model to produce highly engaging dialogues (Liu et al., 2016). Li et al. (2016b,a) propose to use several heuristic criteria: how easy to answer the utterance with non-generic response, how grammatical the response is, etc. Zhang et al. (2018) suggests to use pre-defined facts about the conversation agents as the context for the dialogue. Specifically, conditioning on those facts (called "memories" in their approach), the dialogue becomes "personalized", purposefully coherent and is perceived being more engaging.

In this paper, we investigate a different approach which leverages the following intuition: an engaging dialogue between two agents is a conversation that is focused and intends to ***discover*** information with the goal of increased understanding of each other. In other words, discovering implies asking engaging and inquisitive questions that are not meant to be answered with dull responses.

How do we use these intuitions to build engaging dialogue chatbots? Imagine a dialogue between a chatbot and a human. The human has facts about herself and is willing to share with the chatbot. The chatbot has only a vague idea what those facts might be – for instance, it knows out of 100 possible ones, 3 of them are true. The chatbot's initial utterance could be random as it has no knowledge of what the 3 are. However, the chatbot wants to be engaging so it constantly selects utterances so that it can use them to identify those 3 facts. This is in spirit analogous to a (job) interview: the HR representative (i.e., our interviewer "chatbot") is trying to figure out the personality characteristics (i.e., "facts") of the applicant (i.e., the "human" interviewee). A successful interview implies that the HR representative was able to get as much information about the applicant as possi-

---

On leave from U. of Southern California (feisha@usc.edu)

Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018), pages 551–561
Brussels, Belgium, October 31 - November 1, 2018. ©2018 Association for Computational Linguistics

ble within a limited amount of time, while dull and repetitive questions are avoided at all cost. In other words, the amount of gathered information can be seen as a proxy measure to the engagingness of the dialogue.

We have implemented such an "interview" setting to validate our intuitions. First, we have developed a metric called DISCOVERYSCORE that can measure how much information has been gathered by the chit-chat bot after a dialogue. During a dialogue, we show how this metric can be used to guide the chatbot's generation of responses at its turns — these responses are selected so that they lead to the highest expected DISCOVERYSCORE. To identify such responses, the chit-chat chatbot needs to simulate how its human counterpart would react. To this end, we have proposed an improved version of the personalized chatbot (Zhang et al., 2018) and use it as the chit-chat bot's model of the human. Finally, we perform human studies on the Amazon Mechanical Turk platform and demonstrate the positive correlation between DISCOVERYSCORE and the engagingness scores assessed by human evaluators on our chit-chat bot.

The rest of the paper is organized as the follows. We discuss briefly the related work in Section 2. In Section 3, we then describe various components in our approach: the metric DISCOVERYSCORE for assessing how engaging an dialogue is, a chatbot model that is used in our study, and a response selection procedure for our chatbot to yield engaging conversations. We report empirical studies in Section 4 and conclude in Section 5.

2 Related Work

One of the biggest challenges for chit-chat bots is the lack of the exact objective for models to optimize. This stands in stark contrast to task-oriented dialogue systems (Wen et al., 2016; Su et al., 2016b).

Several heuristic criteria are proposed in (Li et al., 2016a,b) as objectives to optimize. Asghar et al. (2017) proposes humans-in-the-loop to select the best response out of a few generated candidates. Cheng et al. (2018) uses an additional input signal – the specificity level of a response, which is estimated by certain heuristics at training time and can be varied during evaluation.

Another way to address the lack of the explicit objective function is to predict many possible responses at once. Zhou et al. (2017) maps the input message to the distribution over intermediate factors, each of which produces a different response. Similarly, (Zhao et al., 2017; Shen et al., 2018; Gu et al., 2018) use variants of *variational autoencoder*. These approaches are complementary to defining the objective for dialogue models, as an external reward can further guide the response generation and simplify learning such one-to-many mappings.

Liu et al. (2016) hypothesizes that creating a perfect metric for automatic evaluation (so it can be used to optimize a dialogue model to be more engaging, at least in principle) is as hard as creating human-like dialogue system itself. The authors also note that some of the common automatic evaluation metrics (of generated texts) like BLEU, METEOR or ROGUE correlate poorly with human judgments of engagingness. Lowe et al. (2017) suggests a metric ADEM, which is trained to mimic human evaluators. While it's shown to have better correlation with the scores assigned by humans, it also gives preference to safer and generic responses.

In our work, we propose to measure how much information the chit-chat bot has gathered about its human counterpart as a proxy to the engagingness of the dialogue. To the best of our knowledge, this metric has not been explored actively in the design of chit-chat bots.

To apply the metric to generate engaging utterances, the chit-chat bot needs to have a model of how the human partner will respond to its utterances. To this end, we have used the chit-chat bot developed in (Zhang et al., 2018) as a base model and improved upon it. That bot, called PROFILEMEMORY, has a set of memories (basically, factual sentences) defining its persona and can output personalized utterances using those memories. Note that in (Zhang et al., 2018) PROFILEMEMORY is used as a chit-chat bot to generate contextualized dialogue (so as to be engaging). In our work, however, we use it and its improved version as a model of how humans might chat. Our chit-chat bots can be any existing ones (such as a vanilla SEQ2SEQ model without persona) or another PROFILEMEMORY with its own persona that is different from what humans might have. The key difference is that our chit-chat bot generates utterances to elicit human counterparts to reveal about themselves while PROFILEMEMORY in its original work generates utterances to tell stories about itself.

Similar ideas have been explored in cognitive research. Rothe et al. (2016) analyzed how people ask questions to elicit information about the world within a *Battleship* game (Gureckis and Markant, 2009). In particular, they proposed to evaluate questions based on *Expected Information Gain* (Oaksford and Chater, 1994), which is built on the similar principles as DISCOVERYSCORE.

3 Method

In the following section, we describe in details our approach for designing engaging chit-chat bots. We start by describing the main idea, followed by discussing each component in our approach.

3.1 Main Idea

The main idea behind our approach is that the chit-chat bot stays in "discovery" mode. Its main goal is to identify key aspects of its human counterpart. Algorithmically, it chooses utterances to elicit responses from the human so that the responses increase its understanding of the human.

More formally, imagine each human is characterized by a collection of K facts $F = \{z_1, z_2, \ldots, z_K\}$, where z_1 is *I was born in Russia*, z_2 is *My favorite vegetable is carrot*, and z_K is *I like to swim*. The chit-chat bot has access to a universal set of all candidate facts $\mathcal{U}$, and F is just a subset of $\mathcal{U}$. However, the bot does not know the precise composition of F at the beginning of the conversation. Its goal is to identify the subset (or to reduce the uncertainty about it). With a bit abuse of terminology, we call F the personality of the human or the persona.

We denote a dialogue as a sequence of sentences $h_N = [s_1, t_1, ..., s_N, t_N]$ where s_n denotes the sentences by the chit-chat bot and t_n denotes the ones by the human.

3.2 A Metric for Measuring Engagingness

The chit-chat bot assumes that the human's response t is generated probabilistically when it is the human's turn to respond to the chit-chat bot's utterance s

$$P(t \mid s, F) = \sum_{z \in F} P(t \mid s, z) P(z \mid s, F) \quad (1)$$

Intuitively, the human first decides on which fact z she plans to use (ie, which information she wants to reveal) and based on the fact and the chit-chat bot's question, she provides an answer.

The goal of the *discovery oriented* chit-chat is to maximize the mutual information between the dialogue and the revealed personality

$$\mathcal{I}(F; h_N) = \mathbb{H}[P(F)] - \mathbb{H}[P(F \mid h_N)] \quad (2)$$

where $\mathbb{H}[\cdot]$ stands for the entropy of the distribution. Maximizing the mutual information is equivalent to minimizing the uncertainty about F after a dialogue. Intuitively, the chit-chat bot aims to discover the maximum amount of knowledge about the human. We thus term this quantity as the DISCOVERYSCORE.

For simplicity, we assume a uniform prior on which F is. Thus, the key quantity to compute is the entropy of the posterior probability. We proceed in two steps.

Calculating the posterior probability We assume that every human's response t_n is independent from the previous dialogue history, conditioned on the immediately previous message, and chatbot's question s_n is independent unconditionally. Thus, the posterior can be computed recursively:

$$P(F \mid h_N) \approx P(F \mid h_{N-1}) \sum_{f \in F} P(z_N = f \mid s_N, t_N) \quad (3)$$

where z_N is the fact used in the Nth turn. The "single-turn" posterior for the specific fact f is computed as (we have dropped the subscript N to be cleaner)

$$P(z = f \mid s, t) = \frac{P(t \mid s, z = f) P(z = f \mid s)}{\sum_{f' \in \mathcal{U}} P(t \mid s, z = f') P(z = f' \mid s)} \quad (4)$$

We will make a further simplifying assumption that $P(z = f' \mid s)$ is uniform[1] and compute the

---

[1] This is only an approximation: the human will respond to "what kind of food do you like?" with any facts that relate to food but definitely not to geographical locations, sports, etc. However, this assumption is not as damaging as long as $P(t \mid s, z = f)$ is almost zero for the z that $P(z = f \mid s)$ should be ignored – the multiplication would result in zero anyway. Since z refers to the fact, $P(t|s, z = f)$ being almost zero reduces to suggest that for a response t, there are just only a very limited number of s (questions) and facts that can be used to generate that response. For example, a response *"I lived in Russia as a child"* can only be elicited from *"Where did you spend your childhood?"* (as question) and *"I was born in Russia"* (as a fact). For any other question and fact pair (such as *"Where did you spend your childhood?"*,

posterior approximately

$$P(z = f \mid s, t) \approx \frac{P(t \mid s, z = f)}{\sum_{f' \in \mathcal{U}} P(t \mid s, z = f')} \quad (5)$$

Substituting this into the expression for $P(F \mid h_N)$, we obtain

$$\begin{aligned} &P(F \mid h_N) \approx \\ &P(F \mid h_{N-1}) \frac{\sum_{f \in F} P(t_N \mid s_N, z_N = f)}{\sum_{f \in \mathcal{U}} P(t_N \mid s_N, z_N = f)} \end{aligned} \quad (6)$$

Acute readers might have identified this as a form of Bayesian belief update, incorporating new evidence at time N. The likelihood $P(t_N \mid s_N, z_N = f)$ depends on how to model how the human generates responses. It is sufficient to note that this probability can be computed conveniently by personalized chatbot models. We postpone the details to the next section.

Calculating the entropy We make an assumption that the number of facts K assigned to the human is known in advance. Therefore, we can consider only probabilities $P(F \mid h_N)$, where F is of a particular known size.

$$P(F \mid h_N, |F| = K) = \frac{P(F \mid h_N)}{\sum\limits_{F' \subset \mathcal{U}, |F'| = K} P(F' \mid h_N)} \quad (7)$$

The entropy of distribution $P(F \mid h_N, |F| = K)$ can be computed directly by enumerating all possible combinations of K facts.

3.3 ChatBot Models

In our work, there are two types of chatbot models. The first one is the chit-chat bot who will respond to messages from the human conversation partner. While we can use any existing chatbot models, the key ingredient to our approach is to respond so that the expected gain of knowledge on the human is increased. However, since the chit-chat bot cannot inquire the human with "if I answer you this, would I gain knowledge?", it has to estimate the gain in knowledge from its model of the human. The second type of model addresses the aspect of modeling the human. In particular, among the 3 models described below, all 3 can be used as the chit-chat bot models and only PROFILEMEMORY and PROFILEMEMORY$^+$ can be used as the model of humans[2].

SEQ2SEQ dialogue model This basic model maps an input message t to a vector representation using the encoder LSTM layer and uses it as an initial state h_0^d for the decoder LSTM layer. The decoder predicts a response s sequentially, word by word via softmax. Both the encoder and the decoder share the same input embeddings table.

PROFILEMEMORY model PROFILEMEMORY (Zhang et al., 2018) is built on top of SEQ2SEQ and uses exactly the same architecture for the encoder. Additionally, it has a list of memory slots (called *profile memory*) and each slot stores a fact, represented by a sentence. Each fact is encoded into a single vector representation using the weighted average of its word embeddings where the embeddings table is shared with the encoder and the decoder. In this work, we call the profile memory as the personality.

The decoder is an LSTM layer with attention over the encoded memories. In essence, the attention mechanism computes a weight for each fact and a weighted sum of the facts form a context vector. The context vector and the hidden states are combined as inputs to a softmax layer to generate words sequentially. For details, please consult (Zhang et al., 2018)

PROFILEMEMORY$^+$ model The PROFILEMEMORY has a weakness that is especially critical to our intent of using it as a model of the human. It has to apply attention at every step, even when responding to messages which are not relevant to any of the facts. Thus it always reveals something about the personality (unless the attention is uniform, generally hard to achieve in practice). To address this issue, we enhance the model with a *DefaultFact*, which does not correspond to any real sentence. It does have a vector representation (as other facts do) except the representation is learned during the training. An advantage is that the *DefaultFact* allows to efficiently train on the dialogue datasets without profile memories, such as OpenSubtitles – intuitively it is the bucket for "all other facts" that the dialogue does not explicitly refer to.

---

and *"I like apples"*), the response would be unlikely. We believe this is largely due to the experimental/data design that has ensured facts are being largely non-overlapping for each personality and the dialogues are in general centered around the facts. We leave to future work on how to refine this approximation.

[2] In our empirical studies, we use PROFILEMEMORY$^+$ most of the time as it is more powerful than the other two.

3.4 Dialoguing with Intent to Discover

As a metric, DISCOVERYSCORE can only be computed over and assess a finished dialogue. How can we leverage it to encourage the chit-chat bot to be more engaging? In what follows, we describe one of the most important components in our approach.

Instead of using the standard maximum a posterior inference for the typical SEQ2SEQ (and its variants) to generate a sentence, we proceed in two steps to identify the best utterance that has the potential to yield high DISCOVERYSCORE. The first step is to generate a large set of candidate utterances (for example, using beam search). The second step is to re-rank these utterances. We describe the second step in details as the first step is fairly standard.

At the Nth turn of the dialogue, the chit-chat bot has access to the dialogue history h_{N-1} and an estimate of the human's personality $P(F \mid h_{N-1})$. Let s be a sentence from the chatbot's candidate set. Since the bot has a model of the human, it can predict the human's response t as

$$t \sim P(\cdot \mid h_{N-1}, s, F) = P(\cdot \mid s, F) \qquad (8)$$

where F is used to instantiate the model's memory/facts/personality – in other words, we query the model to see what kind of utterances the human might respond with.

The value of a possible response s, i.e, the expected DISCOVERYSCORE assuming s and t completes the dialogue with $h_N = [h_{N-1}, s, t]$, is then given by

$$V(s) = \mathbb{E}_{F \sim P(\cdot|h_{N-1})} \mathbb{E}_{t \sim P(\cdot|s,F)} \mathcal{I}(F; h_N) \qquad (9)$$

Note that the first expectation is needed as the bot has uncertainty of what personality the human is. In practice, we compute this for each s from the candidate set by sampling F and t. We then select the optimal utterance that maximize the value

$$s_N = \arg\max_s V(s) \qquad (10)$$

4 Experiments

We evaluate empirically the proposed approach in several aspects. First, we investigate the effectiveness of the proposed PROFILEMEMORY$^+$ model. This model is especially used to model human interlocutors so that it can be used by the chit-chat bot to estimate how an utterance could elicit the human partner to reveal key facts about her (cf. Section 3.4). Secondly, we investigate whether the proposed metric DISCOVERYSCORE correlates with the engagingness score of a dialogue assessed by human evaluators.

4.1 Evaluating PROFILEMEMORY$^+$

We contrast PROFILEMEMORY$^+$ to SEQ2SEQ and PROFILEMEMORY. We show that not only PROFILEMEMORY$^+$ is a stronger model for personalized chit-chat but also PROFILEMEMORY$^+$ does not reveal its personality easily. Being discreet is a highly desirable property when the model is used to simulate the human participating in the dialogue; when the facts are easily revealed, then the chit-chat bot can use generic or irrelevant questions to identify the personality thus the dialogue does not become engaging.

4.1.1 As a stronger personalized chatbot

Datasets We train all three models on the original PersonaChat dataset (Zhang et al., 2018) and the Year 2009 version of the OpenSubtitles corpus (Tiedemann, 2009). The PersonaChat data set, which consists of crowdsourced 9000 dialogues (123,000 message-response pairs in total) between two people with randomly assigned personas/personalities. There are total 1155 personalities and each personality is defined by 3 to 5 memories (facts such as *"I was born in Russia"* or *"I like to swim"*). 968 dialogues are set aside for validation and 1000 for testing. We report the perplexity of our models on this test data set. The OpenSubtitles corpus has 322,000 dialogues (1.2 million message-response pairs). During training, we augment samples from OpenSubtitles with random personas, which forces PROFILEMEMORY$^+$ to actively prioritize *DefaultFact* over these fake facts.

Implementation Details Similarly to (Zhang et al., 2018), we use a single layer LSTM for both the encoder and the decoder with hidden size of 1024 for all models. The word embeddings are of size 300 and are initialized with GloVe word vectors (Pennington et al., 2014). All models are trained for 20 epochs to maximize the likelihood of the data by using SGD with momentum with batch size 128. Learning rate is reduced by a factor of 4 if the validation perplexity has increased compared to the previous epoch. We found that *general* post-attention (Luong et al., 2015) over

Model	Datasets	Perplexity
SEQ2SEQ	P	38.08
PROFILEMEMORY	P	34.54
SEQ2SEQ	P	31.538
SEQ2SEQ	P+O	30.022
PROFILEMEMORY	P	28.406
PROFILEMEMORY	P+O	27.373
PROFILEMEMORY$^+$	P	28.098
PROFILEMEMORY$^+$	P+O	**26.807**

Table 1: Perplexity on PersonaChat test dialogues by 3 different models. For datasets, P stands for PersonaChat and O for OpenSubtitles. The first two rows are reported by (Zhang et al., 2018). The rests are from our implementation.

encoded memories gives better performance than pre-attention. Weights for encoding memories are being learned during training and are initialized with 0.01 for the top 100 frequent words, and with 1 for others. We found that this simple initialization procedure outperforms the one suggested in (Zhang et al., 2018).

Results The perplexity on the test dialogues by all of the models is contrasted in Table 1. The first two rows are previously reported in (Zhang et al., 2018). The rest results are from models implemented by us.

Our re-implementation of SEQ2SEQ and PROFILEMEMORY show better performance than what are reported in (Zhang et al., 2018), likely due to the difference in the amount of data used for training[3] , as well as model architecture (post- instead of pre-attention) and optimization procedure (e.g., SGD vs. ADAM).

Including additional data such as OpenSubtitles, in general, improves performance. Our PROFILEMEMORY$^+$ performs better than PROFILEMEMORY. This is the benefit of having *Default-Fact* (cf. Section 3.3) which re-directs the attention by the messages and responses that are not related to the real personality away it. On the other end, in PROFILEMEMORY, the attention has to select a real personality no matter what the messages or responses are.

4.1.2 As a discreet chatbot

Since we intend to use a personalized chatbot such as PROFILEMEMORY and PROFILEMEMORY$^+$ as a model of the human interlocutor, we would want the model to behave intelligently: when given an

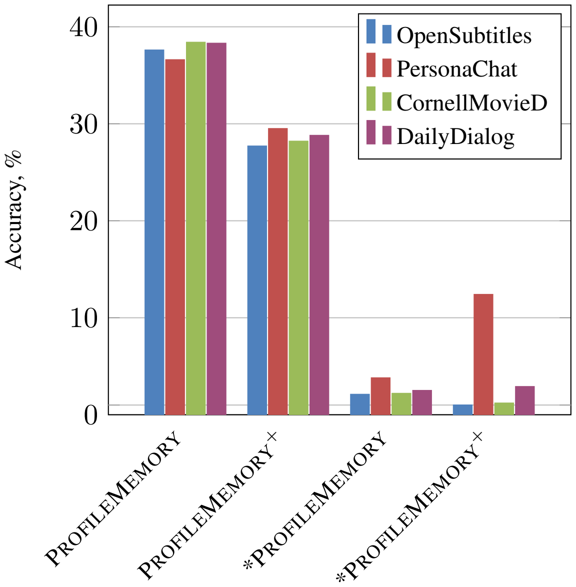

Figure 1: Average accuracies of utterances sampled from different corpora (PersonaChat, OpenSubtitles, CornellMovieDialog, DailyDialog) in revealing the personality of the human interlocutor modeled by personalized chatbots (cf. Section 4.1.2). PROFILEMEMORY and PROFILEMEMORY$^+$ have been trained on PersonaChat data set, while *PROFILEMEMORY and *PROFILEMEMORY$^+$ have been trained on both PersonaChat and OpenSubtitles.

Top 5 sentences, accuracy 29-32%
Tell me about it! What do you do for fun?
Nice! what is it that you do?
In a cabin, all by myself, hoping my grandkids will visit. any you?
You should give it a try! what do you do with your weekends?
Wow cool. what do you do in your spare time. i work on art projects.
Bottom 5 sentences (sampled), accuracy 0%
How come you were rejected?
So he can stay put
Spending the night pondering life.
Hence the fact that she survived.
Maybe she just needs a friend?

Table 2: Examples of sentences from PersonaChat dataset with the highest and the worst accuracies in revealing a personality of PROFILEMEMORY$^+$ model (trained jointly on PersonaChat and OpenSubtitles). We expect human to reply to such utterances with something which will more likely (correspondingly, less likely) reveal her personality. Note, that we don't predict human's personality from the presented utterances alone. Rather, these are considered good (correspondingly, bad) questions to get to know your interlocutor better.

---

[3] Zhang et al. (2018) only modeled the second person in dialogues, which reduces the training data by half.

irrelevant message, the model should not reveal its personality. When given a relevant message, the model reveals its personality. We can expect a similar behavior from a real human, which might reply with *"I don't know"* or *"I don't understand"*, when the question is irrelevant to them. In other words, we want to avoid simulating dialogues like this – Chatbot: *"Hmm... Thank you."*, Human (Simulation): *"I was born in Russia"*. With this adversity, the chit-chat bot has to ask meaningful and relevant questions if its goal is to discover the personality of the human interlocutor.

Experiment setup We randomly sample 100 memories/facts (out of total 5709) from the PersonaChat dataset. For simplicity, we assume the model of the human interlocutor has a simple personality, denoted by one of the 100 facts. We assign this single personality to each of the PROFILEMEMORY and PROFILEMEMORY$^+$ models trained on either the PersonalChat dataset or jointly with the OpenSubtitles dataset. So there are 4 variants in total.

We then construct a simple 2-turn dialogue, where the model is given a probing message and the model responses with a sampled utterance. We use the DISCOVERYSCORE (cf. Section 3.2) to measure how much the dialogue reveals a personality. We then select the personality that maximizes the revealing. If the selected personality is the true personality, we consider the lead message is able to accurately predict the personality. We then average all probing messages to compute the averaged accuracy.

For probing messages, we use sentences from 4 different datasets - PersonaChat, OpenSubtitles, CornellMovieDialogCorpus (Danescu-Niculescu-Mizil and Lee, 2011) and DailyDialog (Li et al., 2017). Our expectation is that an ideal model won't reveal its personality when asked a random question from OpenSubtitles or CornellMovieDialogCorpus, since most of the time it's completely irrelevant lines from a movie script. DailyDialog contains more casual conversations, so some of them we expect to be useful. Of course, the accuracy of random sentences from PersonaChat should be the highest on average, since the corpus was collected with the intent to get to know each other better.

Results The averaged accuracies from the different corpora are shown in Figure 1.

For PROFILEMEMORY$^+$ trained only on PersonaChat data, all types of sentences have similar effectiveness in predicting personality. However, after joint learning with OpenSubtitles, only sentences from PersonaChat (which are most relevant to personalities) are able to predict noticeably accurate than other sentences.

As an illustration, examples of the sentences from PersonaChat with the best and the worst accuracy are presented in the Table 2, for the PROFILEMEMORY$^+$ trained both on PersonaChat and OpenSubtitles.

These findings, together with the superior modeling ability (cf. Section 4.1.1 and Table 1), have validated the usage of PROFILEMEMORY$^+$ trained additionally on OpenSubtitles as a proper model for human interlocutors.

4.2 Human Evaluation

We report human evaluations to show that (1) DISCOVERYSCORE can be used to measure how engaging the conversation is. (2) the dialogue's quality can be increased by choosing a response with the highest expected future DISCOVERYSCORE (cf. Section 3.4).

Setup We use "The Conversational Intelligence Challenge 2"[4] evaluation procedure provided in ParlAI framework (Miller et al., 2017). We use around 200 Amazon Mechanical Turkers for human evaluation. The same procedure is also used in (Zhang et al., 2018).

During an evaluation round, a Turker is assigned a random persona (with 3-5 profile facts) from the PersonaChat dataset. Each Turker is paired with a chatbot – we experiment with several models including SEQ2SEQ and PROFILEMEMORY trained on PersonaChat and PROFILEMEMORY$^+$ model trained on both PersonaChat and OpenSubtitles. The chatbot can also adopt personal facts, but only PROFILEMEMORY and PROFILEMEMORY$^+$ are able to utilize it. Every evaluation dialogue has at least 6 turns per participant.

After the dialogue the Turker is asked to evaluate its interlocutor (i.e., the chatbot) by how fluent, engaging and consistent it is on a 1 to 5 scale (5 being the best). Our primary focus is engagingness score and we will show in below that it correlates well with the DISCOVERYSCORE we proposed. The Turker is also asked to guess the chatbot's persona out of two given persona candidates

---

[4] `http://convai.io`

ChatBot Model	**Fluency**	**Engagingness**	**Consistency**	**Persona Detection, %**
SEQ2SEQ	3.90 (1.24)	3.52 (1.44)	3.77 (1.32)	57.14
+ BeamSearch	4.25 (1.13)	3.51 (1.11)	3.92 (1.27)	47.41
+ BeamSearch + Re-ranking	4.64 (0.67)	**3.92** (1.14)	4.03 (1.16)	48.65
ProfileMemory	4.13 (1.04)	3.62 (1.48)	3.92 (1.29)	68.57
+ BeamSearch	4.54 (0.82)	3.92 (1.09)	4.28 (1.11)	60.58
+ BeamSearch + Re-ranking	4.25 (0.97)	**4.10** (1.10)	4.22 (1.06)	70.00
ProfileMemory$^+$	4.03 (1.22)	3.70 (1.22)	3.79 (1.34)	78.89
+ BeamSearch	4.59 (0.84)	3.73 (1.44)	4.16 (1.20)	61.22
+ BeamSearch + Re-ranking	4.41 (1.05)	**4.27** (1.07)	3.99 (1.26)	69.89

Table 3: Human evaluation results of various dialogues models. Every model is evaluated by its fluency, engagingness and consistency on a scale from 1 to 5. Persona Detection corresponds to how accurate a human can guess the chatbot's personality thus demonstrating how well a model utilizes the assigned persona (note, it's **not** related to DiscoveryScore). Numbers in parenthesis correspond to standard deviation.

Engagingness	Average DiscoveryScore	% of questions	Average Length
1	2.578 (1.697)	37.6	8.04
2	2.776 (1.923)	48.4	8.53
3	2.634 (1.837)	53.0	8.29
4	2.977 (2.044)	65.6	7.92
5	3.196 (1.682)	55.6	8.68

Table 4: Average DiscoveryScore over dialogues grouped by corresponding engagingness score in different tiers. Numbers in parenthesis correspond to standard deviation.

(each with 3-5 profile facts). This metric is called "Persona Detection" and demonstrates how well the model is utilizing the assigned persona. Naturally, we expect SEQ2SEQ-based chatbots to have Persona Detection rate around 50% since they are not using provided persona at all. Each chatbot model is evaluated on at least 100 dialogues.

Response generation We experimented with both the greedy decoding (which is default) and the beam search (with 100 beam size) for text generation.

DiscoveryScore-based re-ranking As described in Section 3.4, we use DiscoveryScore-based re-ranking to select the response with the intent to discover the personality of the human participant. Concretely, the chatbot is given a set of 30 facts from PersonaChat data set, which does include the true facts. The chatbot is also told that the human has only 3 facts in her personality. This is mainly for computational efficiency.

The re-ranking takes place in two steps. First, the chatbot generates 100 response candidates. For every candidate, it performs 10 simulated dialogues with the ProfileMemory$^+$ model as a proxy for the human interlocutor. Finally, it selects the response with the highest expected DiscoveryScore.

Results The evaluation results are presented in the Table 3. The results clearly demonstrate that the DiscoveryScore-oriented re-ranking makes conversations more engaging for all type of the chatbot models.

When re-ranking is used, many human evaluators provided a feedback stating that the model was acting "genuinely interested" and asked a lot of questions. In contrast, modeling without re-ranking had a lower engagingness score precisely because of the lack of questions.

Persona Detection score indicates that ProfileMemory$^+$ is doing a better job in modeling a persona. We also see a decrease in this metric when we combine ProfileMemory$^+$ with the re-ranking procedure, which is likely caused by the chatbot asking more questions than revealing itself personality.

Example dialogues between human and two ProfileMemory$^+$ models with and without DiscoveryScore-based re-ranking are given in

ChatBot's Persona	**Human's Persona**
I will graduate from college in two years.	I play piano.
My parents are divorced.	I have a ten year old son.
I could not live without my dog, Bart.	I graduated high school in 1986.
I am very popular and have a lot of friends.	I hope to become a music teacher after retirement.
I am an excellent student.	

[Human]: Hi! How are you doing?
ProfileMemory$^+$: I am good how are you
DiscoveryScore: I am well, thank you. How are you?

[Human]: I'm doing okay
ProfileMemory$^+$: Okay.
DiscoveryScore: Ok. What do you do for work?

[Human]: I'm working in some local company. How about you?
ProfileMemory$^+$: I am a stay at home mom
DiscoveryScore: I am a college student. Do you have any hobbies?

[Human]: I'm a musician, I'm playing piano from time to time.
ProfileMemory$^+$: What city are you from?
DiscoveryScore: That is cool. Do you have any hobbies?

[Human]: Not really. But I hope to become a music teacher soon. What are your hobbies?
ProfileMemory$^+$: What kind of music do you like?
DiscoveryScore: I like to spend time with my friends and family. You?

Table 5: Examples of the dialogue between human and chatbots: ProfileMemory$^+$ model conducts only beam size for best response and DiscoveryScore additionally uses DiscoveryScore-based re-ranking.

the Table 5.

DiscoveryScore as a proxy for Engagingness We group all the dialogues between chatbots and humans by the assigned engagingness score and compute the average DiscoveryScore, average length of utterances and average percentage of generated questions - see Table 4. Interestingly enough, there is no obvious correlation between how engaging the dialogue has been perceived and simple metrics like the length of the response or the number of asked questions. On the other hand, it is strongly correlated with DiscoveryScore, indicating that it indeed can be used as one of the automatic metrics for dialogues quality.

5 Conclusion & Future Work

We introduce a new metric DiscoveryScore to assess the engagingness of a dialogue based on the intuition that the more interested the chatbot is in its interlocutor the more engaging the dialog becomes. We propose an improved ProfileMemory$^+$ model, which achieves state-of-the-art perplexity results on the PersonaChat dataset. One appealing property of the model is that it doesn't reveal assigned personality upon irrelevant questions. We demonstrate how it can be used to estimate the expected DiscoveryScore by running simulations with the model as a human substitute. A re-ranking method that uses such estimates allows us to significantly improve the dialogue engagingness score over several baselines, which we demonstrate with human evaluations.

We hope to continue exploring DiscoveryScore in more general settings with richer, more complicated personalization or when profile information is not explicitly defined.

Acknowlegements

We appreciate the feedback from the reviewers. This work is partially supported by USC Graduate Fellowships, NSF IIS-1065243, 1451412, 1513966/1632803/1833137, 1208500, CCF-1139148, a Google Research Award, an Alfred P. Sloan Research Fellowship, gifts from Facebook and Netflix, and ARO# W911NF-12-1-0241 and W911NF-15-1-0484.

References

Nabiha Asghar, Pascal Poupart, Xin Jiang, and Hang Li. 2017. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference* *on Lexical and Computational Semantics, *SEM @ACM 2017, Vancouver, Canada, August 3-4, 2017*, pages 78–83. Association for Computational Linguistics.

Regina Barzilay and Min-Yen Kan, editors. 2017. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. Association for Computational Linguistics.

Xueqi Cheng, Jun Xu, Jiafeng Guo, Yanyan Lan, Ruqing Zhang, and Yixing Fan. 2018. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1108–1117. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Xiaodong Gu, Kyunghyun Cho, JungWoo Ha, and Sunghun Kim. 2018. Dialogwae: Multimodal response generation with conditional wasserstein auto-encoder. *CoRR*, abs/1805.12352.

Todd M Gureckis and Douglas B. Markant. 2009. Active Learning Strategies in a Spatial Concept Learning Game. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In (Su et al., 2016a), pages 1192–1202.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Chia-Wei Liu, Ryan Lowe, Iulian 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. In (Su et al., 2016a), pages 2122–2132.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. In (Barzilay and Kan, 2017), pages 1116–1126.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Alexander H. Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. Parlai: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017 - System Demonstrations*, pages 79–84. Association for Computational Linguistics.

Mike Oaksford and Nick Chater. 1994. A rational analysis of the selection task as optimal data selection. *Psychological Review*, 101(4):608–631.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 583–593. ACL.

Anselm Rothe, Brenden M. Lake, and Todd M. Gureckis. 2016. Asking and evaluating natural language questions. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society, Recogbizing and Representing Events, CogSci 2016, Philadelphia, PA, USA, August 10-13, 2016*. cognitivesciencesociety.org.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In (Singh and Markovitch, 2017), pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1577–1586. The Association for Computer Linguistics.

Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. 2018. Improving variational encoder-decoders in dialogue generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*. AAAI Press.

Satinder P. Singh and Shaul Markovitch, editors. 2017. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. AAAI Press.

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. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 196–205. The Association for Computational Linguistics.

Jian Su, Xavier Carreras, and Kevin Duh, editors. 2016a. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. The Association for Computational Linguistics.

Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016b. Continuously learning neural dialogue management. *CoRR*, abs/1606.02689.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Jörg Tiedemann. 2009. News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, volume 5, pages 237–248.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, David Vandyke, and Steve J. Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *CoRR*, abs/1604.04562.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *CoRR*, abs/1801.07243.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In (Barzilay and Kan, 2017), pages 654–664.

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In (Singh and Markovitch, 2017), pages 3400–3407.