Improving Factual Consistency Between a Response and Persona Facts

Mohsen Mesgar **Edwin Simpson*** **Iryna Gurevych**
Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science, Technical University of Darmstadt
https://www.ukp.tu-darmstadt.de

Abstract

Neural models for response generation produce responses that are semantically plausible but not necessarily factually consistent with facts describing the speaker's persona. These models are trained with fully supervised learning where the objective function barely captures factual consistency. We propose to fine-tune these models by reinforcement learning and an efficient reward function that explicitly captures the consistency between a response and persona facts as well as semantic plausibility[1]. Our automatic and human evaluations on the PersonaChat corpus confirm that our approach increases the rate of responses that are factually consistent with persona facts over its supervised counterpart while retaining the language quality of responses.

1 Introduction

Response generation models should ideally generate an appropriate *response* to a given context consisting of utterances previously exchanged between dialogue partners and facts describing the speakers' persona. These models have applications in developing dialogue systems as user interfaces for digital assistants (Bobrow et al., 1977) and also in asynchronous interactions in social media in which speakers define themselves by their profiles.

In this work, we focus on the aspects of persona that can be captured by a set of factual statements, a.k.a., profiles. Table 1 illustrates the persona of the speaker who should respond to the given message. The first response is *topically coherent with the message* and also *linguistically fluent* (or in general, *semantically plausible*) but *factually inconsistent*, unlike the second response, with the second fact in the speaker's persona. We aim to improve the response quality in terms of its factual consistency with facts about the given speaker's persona while retaining its semantic plausibility.

Persona
fact 1: i hate my job
fact 2: i ' m 40 years old
fact 3: i work as a car salesman
fact 4: my wife spends all my money
fact 5: i am planning on getting a divorce
Dialogue History
message: hi , want to be friends?
Generated Responses:
inconsistent: i ' d love to be friends . i ' m 50 years old
consistent: sure , i am 40 , i can tell you about myself

Table 1: A speaker's persona, dialogue history, one inconsistent, and one consistent possible response.

Recent approaches to this problem (Zhang et al., 2018; Dinan et al., 2019; Wolf et al., 2019) generate a response conditioned on persona facts and dialogue history and then use human-generated responses as demonstrations to train their models by fully supervised learning (SL). While this strategy has led to markedly improved performance, there is still a misalignment between this training objective – maximizing the likelihood of human-written responses – and what users care about – generating semantically plausible and factually consistent outputs as determined by humans. This misalignment has several reasons: the maximum likelihood objective considers no distinction between primary errors (e.g. inconsistent responses) and unimportant errors (e.g. selecting the precise word from a set of synonyms); models are incentivized to place probability mass on all human-generated responses, including those that are low-quality; and distributional shift during sampling can degrade performance. Optimizing for targeted quality factors is a principled approach to overcome these problems

* Now at the Intelligent Systems Lab, Dept. of Computer Science, University of Bristol.

[1] https://github.com/UKPLab/EACL21-personalized-conversational-system

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 549–562
April 19 - 23, 2021.

(e.g., Gao et al. (2019) optimize text summarization systems for quality factors relevant to that task).

Our goal is to advance methods for training response generation models on objectives that closely capture the behavior users care about. We first define a reward function to explicitly assesses the quality of a generated response according to *factual consistency with persona facts*, *topical coherence with dialogue history*, and *language fluency*. We then train a policy via reinforcement learning (RL) to maximize the score given by our reward function; the policy generates a token of response at each "time step", and is updated using the Actor-Critic learning approach (Mnih et al., 2016) based on the "reward" our reward function gives to the entire generated response.

We evaluate our approach on PersonaChat (Zhang et al., 2018), a benchmark corpus of English dialogues designed to evaluate the factual consistency between a response and persona facts. We assess the language quality and the factual consistency of responses our RL-based model generates using automatic metrics and human evaluations . Our core contributions are twofold:

- We propose to fine-tune a transformer-based response generation model by an RL method including an efficient reward function that ensures factual consistency with persona facts as well as semantic plausibility of a response.
- We use automatic and human evaluations to show that our RL-based method generates a response that is factually consistent with persona facts more frequently than its SL-based counterpart (Wolf et al., 2019).

The method we present in this paper is motivated in part by long-term concerns about the misalignment of NLP systems with what humans want them to do. When misaligned response generation models generate facts inconsistent with background knowledge like persona facts, their mistakes are relatively low-risk and easy to catch. However, as these systems become more popular to solve essential tasks, their mistakes will likely become more subtle, making this an important area for further research.

2 Method

Let $d = (u_1, ..., u_{T-1})$ be the exchanged utterances between dialogue partners until turn $T-1$, and $p = \{f_1, ..., f_{|p|}\}$ be a persona expressed by a set of facts (i.e. short sentences) about the speaker who should generate a response. Our goal is to generate a response $r = (t_1, ..., t_M)$ consisting of M tokens so that r is consistent with the facts in persona p, topically coherent with u_{T-1}, and linguistically fluent.

2.1 TransferTransfo-SL

We use the *TransferTransfo* (Wolf et al., 2019) dialogue model which is pre-trained and then fine-tuned with fully supervised learning (SL). TransferTransfo is a multi-layer transformer (Vaswani et al., 2017) based on the Generative Pre-trained Transformer (GPT) (Radford et al., 2018). Each transformer layer uses constrained self-attention where every token can only attend to its left context. Generation was performed using beam search with sampling, and an n-gram filtering is used to ensure the model does not directly copy from the persona facts nor former utterances. This model significantly improves over the traditional seq-to-seq, memory-based, and information-retrieval baselines in terms of (1) topical coherence of the response, (2) consistency with a predefined persona, and (3) grammaticality and fluency as evaluated by the automatic metrics in the ConvAI2 competition (Dinan et al., 2019). Since this agent uses transformers, it copes with different lengths of dialogue history.

The transformer layers' parameters in this model are transferred from the pre-trained GPT and then are fine-tuned in a supervised scenario to optimize the losses for the response classification and response generation tasks. The former loss measures if the model distinguishes a correct response appended to the input sequence from a set of randomly sampled distractors, which are randomly selected. The latter one is the language modeling loss that measures how well the model can generate a response similar to the human-generated response. The generative loss is estimated as follows: the self-attention model's final hidden state is fed into an output softmax over the vocabulary to obtain the next response token probabilities. These probabilities are then scored using a negative log-likelihood loss, where the gold next tokens are taken as labels.

2.2 TransferTransfo-RL

Besides the remarkable improvement achieved by TransferTransfo-SL, its generated responses are not necessarily factually consistent with persona

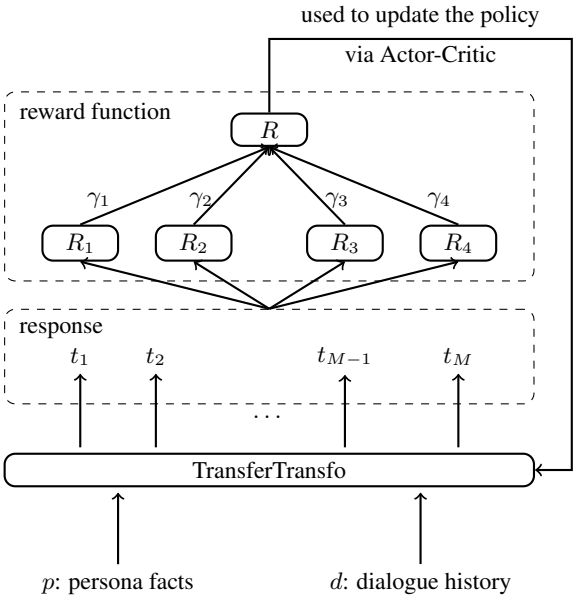

Figure 1: An abstract view of our RL approach.

facts. For example, the inconsistent response in Table 1 is generated by this system. We propose to fine-tune the parameters of this model using reinforcement learning (RL). The TransferTransfo model generates a response token-by-token for a given persona and dialogue history. After generating the last token, i.e. '<EOS>', or reaching the maximum length allowed for a response, a reward model assesses the quality of the response (Figure 1). The reward value is used to fine-tune the parameters of TransferTranfo towards the policy that generates a response that is factually consistent with persona facts and also semantically plausible.

Action We consider generating each token of a response as an action performed by the TransferTransfo model:

$$\mathcal{P}_\theta\left(r|s\right) = \mathcal{P}_\theta(t_1|s)\prod_{k=2}^{M}\mathcal{P}_\theta(t_k|t_{1..k-1},s)\ , \quad (1)$$

where t_k is the kth token in response r and $t_{1..k-1}$ indicates the sequence of tokens generated prior to token k. For the sake of brevity, we use the notation s to refer to (p,d). The function $\mathcal{P}_\theta(r|s)$ is the policy with the parameters θ of TransferTransfo.

Reward function A response generation system should ideally generate a response that is factually consistent with the persona facts, topically coherent with the former interactions, and linguistically fluent. Thus, we propose a compound reward consisting of four sub-rewards: R_1 ensures factual consistency with the persona facts. R_2 accounts for topical coherence with the former utterance. R_3 and R_4 reinforce fluency. We use a weighted sum of these sub-rewards as the training signal:

$$R = \gamma_1 R_1 + \gamma_2 R_2 + \gamma_3 R_3 + \gamma_4 R_4\ , \quad (2)$$

where $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 1$. These weights can be tuned as described below to prevent biasing the policy toward a particular sub-reward.

Persona consistency sub-reward ($\mathbf{R_1}$) Recent studies (Welleck et al., 2019; Dziri et al., 2019) show that consistency with factual information, such as persona facts, can be characterized as a natural language inference (NLI) problem, where entailment labels can be taken as consistent labels and contradiction labels as inconsistent labels. Building on this, we use an NLI model to design this sub-reward. We define our NLI model using BERT as a bidirectional contextualized encoder:

$$\begin{aligned} h^{\texttt{[cls]}}, - &= \texttt{BERT}(\texttt{[cls]}\, f_i\, \texttt{[SEP]}\, r)\ , \\ [s_e, s_c, s_n] &= \texttt{MLP}(h^{\texttt{[cls]}})\ , \\ \left[\mathcal{P}_e^{\text{NLI}}, \mathcal{P}_c^{\text{NLI}}, \mathcal{P}_n^{\text{NLI}}\right] &= \texttt{Softmax}([s_e, s_c, s_n])\ , \end{aligned} \quad (3)$$

where f_i is a fact in the given persona, r is the generated response, [SEP] is the separator token, and $h^{\texttt{[cls]}}$ is provided by BERT to classify semantic relationships between input sentences (Devlin et al., 2019). MLP is a linear layer that maps $h^{\texttt{[cls]}}$ to the scores s_e, s_c and s_n, for the *entailment*, *contradiction*, and *neutral* classes, respectively. $\mathcal{P}_e^{\text{NLI}}$, $\mathcal{P}_c^{\text{NLI}}$ and $\mathcal{P}_n^{\text{NLI}}$ denote the respective class probabilities.

We train our NLI model to predict the NLI classes of pairs of utterances and persona facts (§3.3). We then use this trained model as R_1 to penalize the agent if its generated response contradicts one of the facts in the persona, and encourages the agent if its response entails a fact:

$$R_1 = \frac{1}{|p|}\sum_{f_i\in p}\mathcal{P}_e^{\text{NLI}}(f_i,r) - \frac{\beta}{|p|}\sum_{f_i\in p}\mathcal{P}_c^{\text{NLI}}(f_i,r)\ , \quad (4)$$

where $\mathcal{P}_e^{\text{NLI}}$ and $\mathcal{P}_c^{\text{NLI}}$ are the entailment and contradiction probabilities of the relationship between f_i and r. Scalar $\beta \geq 1$ is a marginal penalty for contradiction over entailment: responses that lack entailment may acceptably be neutral, while contradictory responses are a serious consistency error.

The sub-reward for the factual consistency with persona facts is not sufficient to generate a semantically plausible response. The agent can maximize

this sub-reward merely by repeating the persona's facts and ignoring topical coherence (for an example, see Appendix A). To prevent such behavior, we assess the topical coherence and grammatical fluency of a response by the following sub-rewards.

Topical coherence sub-reward ($\mathbf{R_2}$) Topical coherence is a crucial property of high-quality dialogues (See et al., 2019; Mesgar et al., 2020). We capture the topical coherence of response r to the last utterance u_{T-1} in dialogue history by representing them using an average pooling layer over their token representations obtained by BERT. Inspired by Baheti et al. (2018) and See et al. (2019), we use *cosine similarity* between $\vec{r}$ and $\vec{u}_{T-1}$ as a proxy for topical coherence:

$$R_2 = \cos(\vec{r}, \vec{u}_{T-1}) \ . \tag{5}$$

Fluency sub-rewards ($\mathbf{R_3}$ and $\mathbf{R_4}$) The above sub-rewards do not assess if the response content expressed is linguistically fluent. As also suggested in prior work (Yarats and Lewis, 2018; Zhao et al., 2019; Bao et al., 2019), applying RL for specific metrics might bring in adverse impacts on linguistic quality. As such, we add sub-rewards R_3 and R_4 to promote linguistic quality. R_3 employs a language model (LM) fine-tuned on a set of utterances (§3.4) to evaluate the language quality of response. To do so, we use the Negative Log-Likelihood (NLL) loss obtained by this LM:

$$R_3 = \frac{\alpha - \text{NLL}(r)}{\alpha}, \tag{6}$$

where parameter α is used to map any value of NLL that is greater than α to α so that the output of R_3 will be between 0 and 1. To retain the language quality of responses similar to those of TransferTransfo, we set α to the maximum NLL value that this LM returns for responses generated by the TransferTransfo model on a development set. R_3 is not biased to the length of a response as NLL is already normalized by response length.

Repeated tokens in a response *significantly and negatively* influence the quality of the response (See et al., 2019). R_4 specifically discourages the generation of 1-gram tokens that appear in a response more than one time in a row:

$$R_4 = 1 - \frac{\#\text{repeated-tokens-in-response}}{\#\text{tokens-in-response}}. \tag{7}$$

Weight optimization In combination, these sub-rewards reinforce factual consistency with persona facts, topical coherence, and language fluency. We use their linear combination as a reward R to prevent our policy from becoming overly biased towards any of the sub-rewards. For instance, while generic responses, such as *"I don't know"*, have high fluency, they are discouraged by the persona-consistency sub-reward as they cannot be entailed from any persona fact. However, the weights must be tuned to ensure a suitable balance between the sub-rewards. We apply grid search over the weights and choose the values that yield a policy with the best performance on a validation set (§3.2).

2.3 Training

The goal of RL is to learn a policy, $\mathcal{P}_\theta$, for generating a response that maximizes the expected reward:

$$\mathcal{L} = \mathbb{E}_{\substack{s\in D \\ r\sim\mathcal{P}_\theta(.|s)}} [R(r,s)] \ , \tag{8}$$

where R is the reward function (Equation 2) and $s = (p, d)$ is the given persona and dialogue history that our policy has generated response r for. Function $\mathcal{L}$ is optimized by a stochastic gradient method, where its gradient is (Mnih et al., 2016):

$$\frac{\partial \mathcal{L}}{\partial \theta} = \mathbb{E}_{\substack{s\in D \\ r\sim\mathcal{P}_\theta(.|s)}} \left[R(r,s) \frac{\partial \log \mathcal{P}_\theta(r|s)}{\partial \theta} \right]. \tag{9}$$

To avoid the high-variance issue, we adopt the *actor-critic* method (Mnih et al., 2016) to fine-tune the policy function directly for our quality goals. This approach reduces the variance in the estimated gradient by sampling a single response $r \sim \mathcal{P}_\theta(.|s)$ and computing the difference between its reward $R(r, s)$ and the reward predicted by a critic, $\eta(t_{1..k}, s)$, for the tokens up to position k in response r. The gradient in Equation 9 is then approximated as follows:

$$\frac{\partial \mathcal{L}}{\partial \theta} \approx \sum_k (R(r,s) - \eta(t_{1..k}, s)) \frac{\partial}{\partial \theta} \log \mathcal{P}_\theta(t_k|s) \ . \tag{10}$$

The critic function is $\eta = w^T h_k$, where w is its trainable parameters and h_k is the vector returned by the TransferTransfo model (our agent) at position k. We update the critic's parameters after each update of the policy's parameters by minimizing the squared error between its estimated rewards and the value our reward model assigns to the response:

$$\mathcal{L}_\eta = \mathbb{E}_{\substack{s\in D \\ r\sim P_\theta(.|s)}} \sum_k [\eta(t_{1..k}, s) - R(r,s)]^2. \tag{11}$$

	Train	Validation
Num. of dialogues	17,878	1,000
Num. of utterances	262,876	15,602
Num. of personas	955	200

Table 2: Statistics of PersonaChat as used in the ConvAI2 competition and our experiments.

3 Experiments

We measure to what extent our RL-based fine-tuning (§2) improves the factual consistency of generated responses while retaining their semantic plausibility. We first introduce the corpus used in our experiments (§3.1). We then evaluate the TransferTransfo-SL and TransferTransfo-RL systems by automatic and human evaluations (§3.2). We finally analyze the models we employ to estimate the factual consistency (§3.3) and language fluency (§3.4) sub-rewards.

3.1 PersonaChat Corpus

We use datasets built on the PersonaChat corpus (Zhang et al., 2018), which consists of dialogues, in English, with 6 to 8 turns between randomly paired human crowd-workers. The workers were assigned short text facts representing personas and instructed to talk to their dialogue partner *naturally to discover each other's persona*. We chose this corpus because of its focus on promoting natural conversations while grounding conversations in the persona facts. Each persona consists of 4 or 5 facts, and on average is assigned to 8.3 unique dialogues. We train and evaluate the aforementioned systems on the standard splits of the version of this corpus made available in ParlAI[2] for the ConvAI2 challenge (Dinan et al., 2019) (Table 2). As the test set is hidden, we evaluate the systems on the validation set. To create a training and evaluation sample consisting of a persona and a dialogue history (Table 1), each dialogue is split at each dialogue turn.

3.2 Response Generation

We study to what extent our RL approach generates a response that is factually consistent with given persona facts and semantically plausible. We use TranserTransfo, which performed best in automatic evaluation and second-best in human evaluation among 26 participants in the ConvAI2 competition, as a response generation model.

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

Settings Following the training setup used by Wolf et al. (2019), we fine-tune TransferTransfo on all training samples in PersonaChat and stop the fine-tuning after three epochs. We refer to this fine-tuned model as TransferTransfo-SL. For TransferTransfo-RL, we continue to fine-tune the TransferTransfo model with our RL approach on 90% of the training set for one epoch, where after each policy update, the critic's parameters are updated for 5 times. For R_1, we use the BERT model trained on Dialogue NLI (§3.3) with $\beta = 2$ and for R_3 we use Dialogue LM (§3.4) with $\alpha = 4$. The maximum response length is 20. The input texts are tokenized according to the GPT byte pair encoding (BPE) but the reward is computed on a completely decoded response text. We use the remaining 10% of the training set to choose the sub-reward weights (Equation 2) based on token-level F1-score, which indicates how well the system's responses match the content of human-generated responses (examined weights and their F1-scores are in Appendix B), resulting in $\gamma_1 = 0.4$, $\gamma_2 = 0.16$, $\gamma_3 = 0.22$ and $\gamma_4 = 0.22$. The high weight of the persona consistency sub-reward (γ_1) is compatible with the goal of dialogues in PersonaChat, which is to reveal the persona of dialogue partners. The weights are also consistent with See et al. (2019): fluency factors (γ_3 and γ_4) are more crucial than cosine-relatedness (γ_2) for responses in this corpus.

3.2.1 Automatic Evaluation

We evaluate these systems on the PersonaChat validation set as used in ConvAI2. We report PPL, F1, and BLEU to assess generated responses according to reference responses. We evaluate the factual consistency of a response and the given persona facts using our NLI model (§3.4). It assigns inference relations between a generated response and each fact in the given persona. Given N fact-response pairs in the whole evaluation set, this metric is:

$$\text{PC} = 100\frac{N_e - N_c}{N}\ , \qquad (12)$$

where N_e and N_c are the numbers of entailment and contradiction labels, respectively.

Results The TransferTransfo-RL outperforms its supervised counterparts on all metrics except PPL (Table 3). The improvements on F1 and BLEU indicate that responses generated by TransferTransfo-RL are more similar to reference responses generated by humans and are not biased

Method	PPL	F1	BLEU	PC
TransferTransfo-SL	21.31	17.06	0.065	09.32
TransferTransfo-RL	22.64	**17.78**	**0.067**	**13.06**

Table 3: Automatic evaluations of responses generated by TransferTransfo-SL and TransferTransfo-RL.

toward simply repeating persona facts or previous utterances. It also shows that responses are as informative as human-provided ones. Our RL method decreases the average word repetition rate (Equation 7) from 9% with TransferTransfo-SL to 7%, increasing the language fluency of responses. So far, we observe that the RL method could retain and even improve the semantic plausibility of a response.

Regarding the factual consistency between a response and given persona facts, TransferTransfo-RL scores significantly higher for the PC metric. This indicates that the number of evaluation samples for which TransferTransfo-RL generates a response consistent with given persona facts is significantly higher than what TransferTransfo-SL does. Looking at PC in detail (Table 4, top), TransferTransfo-RL increases the frequency of cases whose generated responses are entailed from (or consistent with) persona facts by 3.41% over TransferTransfo-SL, while reducing contradictions (or inconsistency) by 0.07% and neutral by 3.61%; showing that fine-tuning with RL improves the policy for generating a response that is factually consistent with persona facts. While our combined reward function achieves good all-round performance, ablation experiments (Appendix A and B) show that each sub-reward is effective and necessary to capture consistency with persona facts, topical coherence, and language fluency.

3.2.2 Human Evaluation

We also conduct a human evaluation between TransferTransfo-RL and TransferTransfo-SL. We randomly select *100 samples*, each of which consists of a dialogue history, a persona, and the responses generated by the examined systems. We ask *seven human judges* (two native and five fluent English speakers) to assign a *consistency label* from {*consistent, neutral, contradicting*} to the response concerning the facts in the persona (instructions in Appendix D). We also ask the human judges to rate the *semantic plausibility* of each response with an ordinal score ranging from 1 (worst) to 5 (best), encompassing *coherence*, *grammatical correctness*, and *low repetitiveness*.

	Consistent	Contradicting	Neutral
Automatic Evaluation			
TransferTransfo-SL	11.14	01.82	87.04
TransferTransfo-RL	**14.81**	**01.75**	**83.43**
Δ	3.41 ↑	0.07 ↓	3.61 ↓
Human Evaluation			
TransferTransfo-SL	43.71	17.71	38.58
TransferTransfo-RL	**52.71**	**14.00**	**33.29**
Δ	9.00 ↑	3.71 ↓	5.29 ↓

Table 4: Frequencies (%) of consistency labels. Automatic evaluation uses our NLI model to assign labels for the whole evaluation set. For human evaluation, judges labeled 100 samples. Δ is the improvement of TransferTransfo-RL over TransferTransfo-SL.

Method	Average Semantic Plausibility
TransferTransfo-SL	3.33
TransferTransfo-RL	**3.50**

Table 5: Human evaluation: semantic plausibility.

Results Table 4 (bottom) shows the average percentage of consistency labels human judges assign to responses generated by TransferTransfo-RL and TransferTransfo-SL. The number of samples for which TransferTransfo-SL generates a consistent response increases by 9% using our RL fine-tuning approach while contradictions (or inconsistencies) decrease by 3.71%, confirming that human judges more frequently find responses generated by TransferTransfo-RL factually consistent with persona facts than those of TransferTransfo-SL. The number of neutral responses also decreases, suggesting fewer generic responses, as neutral responses tend to be generic (Welleck et al., 2019).

Overall, Table 4 shows a similar trend between the human and the automatic evaluations, confirming the findings of the automatic evaluation. Unlike the human evaluation, our automatic evaluation shows that the models generate a neutral response for most cases. The NLI model assesses more responses to be neutral than humans do – humans can reason about entailment relations using their common senses, while the NLI model does not identify any relation. Further analysis (Appendix E) shows that for over half of the cases for which TransferTransfo-SL generates a contradicting (inconsistent) response, our TransferTransfo-RL generates a consistent response, indicating that the

idea of using RL to fine-tune a pre-trained agent improves its capability in generating a factually consistent response with persona facts.

In terms of semantic plausibility (topically coherent and linguistically fluent), Table 5 shows that the human judges find responses generated by TransferTransfo-RL are on par with those of TransferTransfo-SL, showing the effectiveness of our topical coherence and fluency sub-rewards.

3.3 Persona-Consistency Sub-reward Validation

As discussed in §2, assessing factual consistency with persona facts can be characterized as an NLI problem. In this experiment, we investigate the choice of the NLI model for this sub-reward by comparing our BERT-based NLI model (§2) with recent NLI models on the *Dialogue NLI* dataset (Welleck et al., 2019). This dataset, which is designed for evaluating factual NLI in dialogues, consists of a set of fact-utterance, fact-fact, and utterance-utterance pairs extracted from the PersonaChat corpus. Each pair is accompanied by a *human-annotated NLI label*, i.e., entailment (or consistent), contradiction (or inconsistent), and neutral. Two examples of the fact-utterance pair from this dataset are: *"My dad is a priest."* **contradicts** *"Since my dad is a mechanic we had mostly car books."*; and *"I like playing basketball"* **entails** *"I prefer basketball. Team sports are fun."*. This dataset contains 310,110 training, 16,500 validation and 16,500 test pairs. Besides the standard test set, which was annotated by *one crowdworker*, there is *Test Gold* containing 12,376 of test pairs, which were annotated by *three crowdworkers* (Welleck et al., 2019).

We compare our BERT-based NLI model with (1) **Majority**, which returns the majority class; (2) **ESIM** Enhanced Sequential Inference Model (Chen et al., 2017), an LSTM-based model with inter-sentence attentions. ESIM is the state of the art on the Dialogue NLI dataset. We use *bert-base-uncased* (Devlin et al., 2019) to encode utterances and facts. We fine-tune the whole model during training. We set the maximum input length to 128, the learning rate to 5×10^{-5}, and the training- and evaluation-batch sizes to 32 and 8, respectively. We compare the NLI models using *accuracy* (Welleck et al., 2019).

Results Table 6 shows that the BERT-based NLI model outperforms ESIM, suggesting that our model better captures the factual relationships between an utterance and a persona fact. Welleck et al. (2019) previously demonstrated that the performance of ESIM is sufficient to check the factual consistency between a response and persona facts; as our model outperforms ESIM, we chose our NLI model for the consistency sub-reward R_1. Indeed, a more accurate NLI model reduces the noise in the reward function and consequently the errors our system makes.

Model	Validation	Test	Test Gold
Majority	33.33	34.54	34.96
ESIM	86.31	88.20	92.45
Our NLI model	**86.84**	**89.50**	**93.60**

Table 6: Accuracy of candidate NLI models for R_1 on the Dialogue NLI dataset.

Model	PPL
Non-Dialogue LM	108.29
Dialogue LM	**10.01**

Table 7: The perplexity (PPL) of the language model (Dialogue LM) used for the fluency sub-reward $R_{3.1}$.

3.4 Response Fluency Sub-reward Validation

Sub-reward R_3 requires a language model to measure the language quality of a response. In this experiment, we investigate if fine-tuning a pre-trained, non-dialogue language model on dialogue utterances makes it suitable for this goal. To do so, we compare (1) **Non-Dialogue LM**, which is the GPT language model with no fine-tuning; and (2) **Dialogue LM**, which is the GPT language model fine-tuned on utterances from PersonaChat. We fine-tune the GPT language model (Radford et al., 2018) for three epochs on 90% of utterances ($\approx$236,588) from the PersonaChat training set. We evaluate the language model on the remaining 10% ($\approx$26,288) of utterances, so the PersonaChat validation dialogues remain unseen for evaluating our dialogue systems. Training- and validation-batch sizes are 8 and 16, respectively. Learning rate is 6.25×10^{-5}, and *perplexity (PPL)* is the evaluation metric.

Results Dialogue LM substantially improves perplexity over Non-Dialogue LM (Table 7). This shows that the fine-tuned language model better captures the linguistic properties of dialogue utterances, yielding a more suitable language model for the fluency sub-reward R_3. See et al. (2019) validated the benefits of cosine similarity for es-

timating the coherence (R_2) and word repetition (R_4) for language quality.

4 Discussions

Case analysis We presented one example of an evaluation sample in Table 1, in which the inconsistent response is generated by TransferTransfo-SL and the consistent one by TransferTransfo-RL. Since TransferTransfo-SL is fine-tuned only with reference responses and does not have any training signal for factual consistency, we speculate that variants of *"I'm 50 years old"* occur in the training set leading the agent to produce a response that is inconsistent with the persona fact *"I'm 40 years old"*. In contrast, TransferTransfo-RL generates a consistent response which is also topically coherent with the given question and linguistically fluent. The above sample is an example of "attribute" consistency, where the response should express an attribute of the speaker. Table 8 shows some other evaluation samples. The top sample shows that TransferTransfo-RL can deal with "have" consistency. Our system correctly recognizes the number of dogs the speaker has and grounds its response on this fact. The evaluation sample in the middle row of Table 8 shows that our RL-based model can also deal with "like-to-do" consistency.

Although TransferTransfo-RL outperforms TransferTransfo-SL in generating different types of consistent responses (such as 'attribute", "have", and "like-to-do"), they both struggle with generating consistent responses for evaluation samples in which understanding of persona facts and dialogue history requires common sense knowledge. As an example, consider the second evaluation sample shown in Table 8. TransferTransfo-SL generates the response "*I'm not married yet*" which contradicts the first fact of the given persona "*My husband is adopted.*"; it seems the model does not have enough knowledge to capture the semantic relationship between "*my husband*" and "*marriage*". The bottom evaluation sample in Table 8 demonstrates the lack of common sense knowledge for TransferTransfo-RL as well. The response "*I like to go to church to sing with wife*" contradicts the fact "*My wife left me and took my children*" in the given dialogue history.

Limitations One limitation of our work is to narrow a speaker's persona to a set of facts expressed as short sentences. Persona has other aspects, such as speaking styles, which need a separate study. Nevertheless, the research question and experiments presented in this work demonstrate the benefits of RL methods for fine-tuning transformer-based models, which are already pretrained, to obtain a policy more aligned with target quality factors. Other aspects of the persona can also be involved in the reward function, given that our method potentially reduces the need for the high-quality demonstration responses generated by humans for supervised fine-tuning.

Future directions In this paper, we demonstrate the effectiveness of RL over SL for fine-tuning pretrained neural models (like GPT) for generating responses that fulfill quality goals such as factual consistency with given persona facts and semantic plausibility in a single round of dialogue. Therefore, the next step might be adopting our reward function to generate factually-consistent responses while retaining the diversity of responses through multiple rounds of dialogue.

5 Related Work

There are two types of approach to persona consistency. The first category includes systems that learn speaker-level embeddings from responses produced by a particular speaker (Li et al., 2016a; Madotto et al., 2019). These systems depend on the availability of suitable responses performed by the speaker whose persona we wish to imitate. If those responses do not reveal the persona information, dialogue systems cannot learn the persona. Moreover, these systems cannot be adapted to new personas at deployment time since the persona embeddings must be learned from training data. So our approach is complementary to them and not directly comparable.

The second category includes systems that rely on a set of facts about a persona. For example, Zhang et al. (2018) propose a key-value memory neural model for this task. This model is outperformed by TransferTransfo (Wolf et al., 2019), which is used in our experiments. Welleck et al. (2019) rank a given set of utterances using an NLI model to select a persona-consistent response. In contrast, we use NLI to train a generative model. Song et al. (2020) propose an NLI-based reward for persona consistency that calculates a score using only the persona facts with the highest entailment and contradiction probabilities, rather than the whole persona. Their approach does not reward topical coherence, which we found crucial

Persona	Dialogue History		TransferTransfo-SL	TransferTransfo-RL
fact 1: i go to the gym regularly . *fact 2*: i like doing the treadmill and rowing machine . *fact 3*: i ' ve short hair . *fact 4*: two dogs live with me . *fact 5*: eating is something i do when i am bored .	*message*: *response*: *message*:	hello , how are you ? hello , i am feeling fit since i visited the gym today ! how are you ? that ' s absolutely great just fantastic after rowing and running on the treadmill , i ' m pretty hungry . do you work out ? yes after i cook all my chef food	i do a lot of cardio . i ' m really short , so i am always looking to	i love to cook , i have 2 dogs .
fact 1: my husband is adopted . *fact 2*: i like to watch nhl hockey . *fact 3*: i live is a rural farming community . *fact 4*: i ' ve a german shepherd dog .	*message*: *response*: *message*: response: *message*:	the farm animals and a dog . his name is lucky . do you go to church ? we have a german shepard named bacon . i love bacon we cook it for breakfast every sunday after church we are christain . we don ' t go to church often . our religion is hockey . my sisters boyfriend plays hockey , she works as a pole dancer and met him there .	i ' m not married yet , i ' m not married yet .	i love to watch nhl games .
fact 1: i am overweight and unhappy *fact 2*: i do not believe in god *fact 3*: i spend most of my time on facebook when i am not working *fact 4*: i work at a nursing home *fact 5*: my wife left me and took my children	*message*: *response*: *message*:	hi i am a custodian named darlene , how are you ? i am good i work in a nursing home and just at home on facebook my hobbies include playing guitar and singing , do you have any?	i do i am a drummer in a band and i work as a nurse at a nursing home	i like to go to church to sing with my wife

Table 8: Some example personas, dialogue histories and the responses generated by our examined agents from the evaluation set.

for relieving effects of the persona-consistency sub-reward on the quality of response.

Persona consistency was also a quality target in the ConvAI2 dialogue generation competition (Dinan et al., 2019). The winner of the human evaluation part of ConvAI2 is the "Lost in Conversation" system (Dinan et al., 2019), which is also a transformer-based model trained by SL on two extra datasets besides PersonaChat. In our paper, we used TransferTransfo trained only on PersonaChat. Our experiments showed that our idea of using RL for fine-tuning neural agents improves factual consistency between a response and persona facts by accounting for it in its reward function.

RL has been extensively used for training task-oriented dialogue systems (e.g., Nogueira and Cho (2017); Liu et al. (2018)). Unlike task-oriented scenarios, where a reward can measure if a task is fulfilled or not, incorporating persona facts lacks a straight-forward measurable outcome. Li et al. (2016b) use RL for generating open-domain dialogue using REINFORCE (instead of Actor-Critic) and an RNN-based model. This agent has no notion of factual consistency with facts about a persona, so is not comparable with our system.

6 Conclusions

We proposed to fine-tune response generation models by RL to improve on the quality goals that matter, e.g., factual consistency between a response and persona facts while retaining semantic plausibility. We adopted the actor-critic method for fine-tuning a pre-trained transformer-based model by defining an efficient and effective reward function measuring persona consistency, topical coherence, and language fluency. Automatic and human evaluations on PersonaChat demonstrate that compared to just using supervised learning, further fine-tuning with RL yields responses that are more frequently factually consistent with persona facts while still semantically plausible.

Acknowledgments

This work was supported by the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1 and grant GU 798/17-1). We thank Kevin Stowe and Leonardo Filipe Rodrigues Ribeiro for valuable feedback on earlier drafts of this paper. We also thank anonymous reviewers for their constructive suggestions.

References

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 2018. Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* Brussels, Belgium, 31 October – 4 November 2018, pages 3970–3980.

Siqi Bao, Huang He, Fan Wang, Rongzhong Lian, and Hua Wu. 2019. Know more about each other: Evolving dialogue strategy via compound assessment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Florence, Italy, 28 July – 2 August 2019, pages 5382–5391.

Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. 1977. GUS, a frame-driven dialog system. *Artificial intelligence*, 8(2):155–173.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Vancouver, Canada, 30 July – 4 August 2017, pages 1657–1668.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , Volume 1 (Long and Short Papers) ,* Minneapolis, Minnesota., 2–7 June 2019, pages 4171–4186.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander H. Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, Shrimai Prabhumoye, Alan W. Black, Alexander I. Rudnicky, Jason Williams, Joelle Pineau, Mikhail Burtsev, and Jason Weston. 2019. The second conversational intelligence challenge (ConvAI2). *CoRR*, abs/1902.00098.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , Volume 1 (Long and Short Papers) ,* Minneapolis, Minnesota., 2–7 June 2019, pages 3806–3812.

Yang Gao, Christian M. Meyer, Mohsen Mesgar, and Iryna Gurevych. 2019. Reward learning for efficient reinforcement learning in extractive document summarisation. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2350–2356. International Joint Conferences on Artificial Intelligence Organization.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Berlin, Germany, 7–12 August 2016, pages 994–1003.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* Austin, Texas, 1–5 November 2016, pages 1192–1202.

Bing Liu, Gokhan Tür, Dilek Hakkani-Tür, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Melbourne, Australia, 15–20 July 2018, pages 2060–2069.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Florence, Italy, 28 July – 2 August 2019, pages 5454–5459.

Mohsen Mesgar, Sebastian Bücker, and Iryna Gurevych. 2020. Dialogue coherence assessment without explicit dialogue act labels. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1439–1450, Online. Association for Computational Linguistics.

Volodymyr Mnih, Adria Puigdomenech, Badia Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of The 33rd International Conference on Machine Learning,* New York, New York, 20–22 Jun 2016, pages 1928–1937.

Rodrigo Nogueira and Kyunghyun Cho. 2017. Task-oriented query reformulation with reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* Copenhagen, Denmark, 7—11 September 2017, pages 574–583.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), Florence, Italy, 28 July – 2 August 2019, pages 1702–1723.

Haoyu Song, Wei-Nan Zhang, Jingwen Hu, and Ting Liu. 2020. Generating persona consistent dialogues by exploiting natural language inference. In *Proceedings of the 34rd Conference on the Advancement of Artificial Intelligence,* New York, New York, 7–12 February 2020, page [TO APPEAR].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing (NeurIPS 2017),* Long Beach, California, 4–9 December 2017, pages 5998–6008.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Florence, Italy, 28 July – 2 August 2019, pages 3731–3741.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, abs/1901.08149.

Denis Yarats and Mike Lewis. 2018. Hierarchical text generation and planning for strategic dialogues. In *Proceedings of the 35th International Conference on Machine Learning*, page 5591–5599.

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? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Melbourne, Australia, 15–20 July 2018, pages 2204–2213.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies , Volume 1 (Long and Short Papers)* , Minneapolis, Minnesota., 2–7 June 2019, pages 1208–1218.

A Using Only the Persona-Consistency Sub-reward R_1

By using the persona-consistency sub-reward as the only reward function, the policy learns only to generate responses that state information about the persona facts and ignores the topical coherence with its former utterance.

Persona
fact 1: i prefer vinyl records to any other music format. *fact 2*: i fix airplanes for a living. *fact 3*: i drive junk cars that no one else wants. *fact 4*: i think if i work hard enough i can fix the world.
Dialogue
message: hi how are you doing? *response*: i am fine. i am a mechanic, so i fix airplanes. i fix them. *message*: do you have any hobbies? *response*: i enjoy listening to vinyl records. i also fix planes. i also fix them. i also fix *message*: sounds cool. *response*: i also fix airplanes. i also fix airplanes. i also fix them. i also fix them.

Table 9: Response generated by the TransferTransfo model where it is fine-tuned by only the persona-consistency sub-reward ($R = R_1$).

Table 9 illustrates an example dialogue that is conducted with an agent trained with only the persona consistency sub-reward ($R = R_1$). The agent always repeats, *"i fix airplanes. i fix them."*, no matter what the input message is about. This problem not only produces topically irrelevant responses but also makes the agent look nagging and self-centered in a conversation.

Persona
fact 1: i like hunting. ...
Dialogue
message: hi how are you doing? *response*: hunting hunting hunting hunting hunting hunting *message*: do you have any hobbies? *response*: hunting hunting hunting hunting hunting hunting

Table 10: Another example dialogue with an agent that is trained by only the persona-consistency sub-reward ($R = R_1$).

Table 10 illustrates another example dialogue with the agent where it is trained only by persona-consistency sub-reward. The agent keeps repeating *"hunting"* from the persona to maximize its reward. The NLI model used for R_1 evaluates the inference relation between a response and a persona and does not capture the topical coherence of the response with its former utterance and language fluency of the response. It is therefore necessary to use R_1 in combination with topical coherence (R_2) and language fluency sub-rewards (R_3 and R_4), as we propose in our reward function.

B Weight Optimization and Reward Ablation

We examine various weight sets $(\gamma_1, \gamma_2, \gamma_3, \gamma_4)$ to balance the contribution of sub-rewards in the complete reward function on the held out set (10% of the PersonaChat training set). Table 11 shows those weights. The balanced weights give the highest

γ_1	γ_2	γ_3	γ_4	$\mathbf{F_1}$
0.00	0.00	0.00	0.00	02.04
1.00	0.00	0.00	0.00	16.34
0.00	1.00	0.00	0.00	16.12
0.00	0.00	1.00	0.00	12.77
0.00	0.00	0.00	1.00	17.91
0.70	0.00	0.30	0.00	16.37
0.65	0.00	0.35	0.00	19.90
0.60	0.00	0.40	0.00	16.98
0.50	0.00	0.50	0.00	16.07
0.25	0.25	0.25	0.25	18.27
0.60	0.20	0.00	0.20	15.54
0.40	0.20	0.20	0.20	20.57
0.40	**0.16**	**0.22**	**0.22**	**20.75**
0.45	0.13	0.17	0.20	19.98
0.47	0.12	0.17	0.20	19.95
0.47	0.10	0.17	0.21	20.33
0.47	0.10	0.19	0.19	19.73
0.50	0.10	0.16	0.20	19.10
0.40	0.20	0.15	0.20	19.34
0.43	0.20	0.12	0.20	20.22
0.45	0.20	0.12	0.20	20.13
0.45	0.25	0.00	0.25	17.40
0.40	0.10	0.25	0.25	18.93
0.40	0.15	0.20	0.20	19.80
0.40	0.20	0.20	0.15	20.44
0.45	0.17	0.21	0.17	18.85
0.50	0.15	0.15	0.15	20.25
0.47	0.13	0.20	0.15	19.97
0.50	0.15	0.20	0.15	17.64
0.55	0.15	0.15	0.10	19.15

Table 11: The examined sub-reward weights and their corresponding F1 on our validation set, i.e., 10% of the PersonaChat training set.

F1 score, suggesting that a combination of sub-rewards leads to responses that are more similar to the human responses.

We also evaluate the use of each sub-reward in isolation, and show the results in Table 12, in comparison with our chosen balanced weights in the bottom line. For the other metrics, we can see that $\gamma_1 = 1$ maximizes the number of entailments from

γ_1	γ_2	γ_3	γ_4	F_1	PPL	Repetition(%)	Consistent (%)	Neutral (%)	Contradiction (%)	PC
1.00	0.00	0.00	0.00	16.34	25.35	36.20	**34.02**	64.62	01.36	**66.33**
0.00	1.00	0.00	0.00	16.12	21.44	13.36	12.24	86.34	**01.42**	55.50
0.00	0.00	1.00	0.00	12.77	**04.76**	12.21	00.46	**99.35**	00.20	51.49
0.00	0.00	0.00	1.00	17.91	21.35	**01.62**	03.53	95.90	00.57	50.13
0.40	0.16	0.22	0.22	**20.75**	12.36	09.61	14.14	84.54	01.32	56.50

Table 12: Performance metrics on our validation set (10% of the PersonaChat training set) when training is performed with each individual sub-reward and our chosen weighted sum of sub-rewards.

the persona facts, $\gamma_3 = 1$ minimizes perplexity, and $\gamma_4 = 1$ gives lowest repetition. Besides F1 score, the balanced weights give good performance across perplexity, repetition, and persona consistency. The setups with fewer neutral responses also tend to have more responses that contradict the persona facts, e.g., for $\gamma_1 = 1$. Neutral responses are a trivial way to avoid contradictory responses and the setup with the least contradictions, $\gamma_3 = 1$, has almost no responses that are consistent with the persona facts. The better overall persona consistency is reflected in the highest PC score for $\gamma_1 = 1$ and next highest for the balanced weights, which trades of PC for less repetition, lower perplexity and a higher F1 score.

C REINFORCE vs Actor-Critic

Figures 2 and 3 show the trend of changes in our reward function during training by REINFORCE and Actor-Critic, respectively. All parameters are the same for the two experiments. We observe that the actor-critic approach converges faster and also is less noisy (has a lower variance) than REINFORCE.

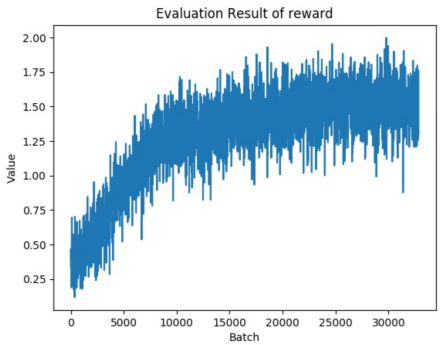

Figure 2: The reward curve during training by REINFORCE.

D Human Evaluation

For each sample, we show to each participant a set of persona facts, a dialogue history, and the response generated by one of TransferTransfo-SL and TransferTransfo-RL. We instruct our participants to assess semantic plausibility according to the following objective definition: *"grammatical correctness, lowest repetitiveness, and coherence"*. The plausibility rates are integer values between 1 and 5, where 5 is most plausible.

To measure persona consistency, we instruct participants as follows:

An answer is considered consistent if:

- It contradicts with neither the dialogue history nor the persona facts;
- It is relevant to any of the given persona facts.

An answer is considered neutral if:

- It contradicts with neither the dialogue history nor the persona facts;
- It is **not** relevant to any of the given persona facts.

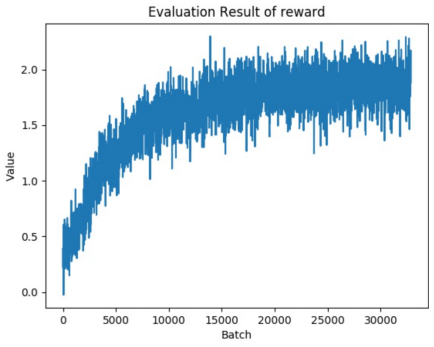

Figure 3: The reward curve during training by Actor-Critic.

E Human Evaluation: Confusion Matrix

Table 13 presents the distributions of consistency labels for TransferTransfo-RL's responses given the consistency labels for TransferTransfo-SL's responses. For the majority of cases whose

		TransferTransfo-RL label		
		Consistent	Neutral	Contradicting
Transfer	Consistent	57.46	27.73	14.82
Transfo	Neutral	46.87	44.99	08.14
-SL	Contradicting	52.48	22.05	25.47

Table 13: Each row corresponds to the cases in the human evaluation for which TransferTransfo-SL received a particular consistency label. The values in each row show the percentages of consistency labels for TransferTransfo-RL for the same data points.

TransferTransfo-SL's responses are contradictory or neutral, TransferTransfo-RL generates consistent responses, showing improved factual consistency with persona facts. However, TransferTransfo-RL generates contradictory responses for some cases whose TransferTransfo-SL responses are consistent with their personas. This may be due to errors in the NLI model's predictions of entailment, hence a more accurate NLI model may improve the quality of the reward function and consequently the consistency of responses. Alternatively, these contradictory responses may receive high rewards from the topic consistency and fluency sub-rewards, which could override R_1.