Personalized Dialog Agent Without Persona Descriptions

Charis Chan cchan13@berkeley.edu
Joyce Ching joycelc@berkeley.edu
Inderpal Kaur ikaur@berkeley.edu

Abstract

Our goal was to create a personalized dialog agent that does not require generated personality descriptions as inputs. This would allow personalized dialog agents to be applied to broader contexts by taking in data such as movie/television scripts, interviews, and podcasts that only require indicators for who is speaking at a given time. We built our model off of HuggingFace's winning entry for ConvAI2 by taking their TransferTransfo model and adding a third personality classifying head that tries to learn the personalities associated with each conversation in place of the generated personality descriptions. Our approach of adding another head ended up performing worse than if the persona descriptions were simply removed from the model. We hypothesize that this may be because our added head ended up competing with the existing Next Sentence Prediction head and that our data set was not properly suited for our task.

1 Introduction

Personalized dialog agents today have been able to display a consistent personality and engage in personal topics. This has been done by utilizing the extensive PERSONA-CHAT data set which consists of 162,064 conversations between Amazon Mechanical Turk (MTurk) crowd-workers who were each randomly assigned a persona, instead of typical training data sets that contain many dialogs each with different speakers. However, models currently using the PERSONA-CHAT data set rely on the multi-sentence descriptions that define the profile of a given persona. For PERSONA-CHAT, these sentences were created by another set of MTurk crowd-workers. This requirement of a manually-generated persona description limits the applicability of personalized dialog agents for new data sets where persona descriptions have not already been created.

Our goal is to create a personalized dialog agent that does not require a persona description as an input. Instead, the model takes in unique persona tokens and learns the unique attributes of the personas through the training dialogs. Removing the persona description and focusing more heavily on the available training dialogs associated with a given persona enables us to apply our model to a broader range of existing dialog data sets that only need to be labeled with the speaker, such as movie and TV show scripts, interviews, and podcasts, and makes our model less reliant on the specific details in the persona description to inform the chatbot's dialog.

2 Background

2.1 PERSONA-CHAT

The PERSONA-CHAT data set was used in the original "Personalizing Dialogue Agents: I have a dog, do you have pets too?" (Zhang et al., 2018). Its creation is motivated by desire to facilitate a more personal and engaging chit-chat dialog. The data comes from three crowd-sourced MTurk tasks:

1. 1155 possible personas were created, each with at least 5 profile sentences that contain typical topics of human interest that can be brought up in conversation.

2. Additional revised sets of the 1155 personas were added to avoid modeling that takes advantage of trivial word overlap. The related sentences consisted of rephrases, generalizations, or specializations.

3. Two Turkers were paired and each assigned a random persona. They were asked to "chat with the other person naturally and try to get to know each other" while playing the role of their given personas.

This resulted in 162,064 dialogs where each turn was a maximum of 15 words per message and dialogs lengths were randomly defined between 6 and 8 turns each.

2.2 HuggingFace's TranserTransfo approach

HuggingFace's model was initially created as an entry to The Second Conversational Intelligence Challenge (ConvAI2) (Dinan et al., 2019) where participants were asked to create models that can both ask and answer questions about personal topics and use the resulting dialog to build a persona of the speaking partner. HuggingFace's model won the automatic evaluation track and has since been offered implementations of large-scale models like OpenAI GPT and GPT-2 (Wolf et al., 2019).

It was trained and tested on the CONVAI2 dataset which is a more extensive version of the original PERSONA-CHAT data set. However, the computing power we had available could not handle such a large data set.

Their model takes a large-scale pre-trained Transformer language model, OpenAI GPT, and applies Transfer Learning fine-tuning to adapt it to the dialog end-task. OpenAI GPT is a Transformer-based language model pre-trained on the BooksCorpus data set. To adapt the model to utilize several types of contexts (persona, history, and reply) to generate its output instead of the single sequence of words OpenAI GPT is trained with, they concatenate the context segments into a single sequence and add word, position, and segment embeddings as parallel input sequences.

To compute their losses, they implement a "DoubleHead" model where one head computes the language modeling predictions while the other predicts next-sentence classification labels. This is done by calculating the negative log-likelihood loss on the portion of the target corresponding to the reply and the cross-entropy loss on classifying the correct reply among distractors respectively.

The decoders used are top-k and nucleus/top-p sampling which succeed beam-search and greedy decoding in reproducing the distributions of words in human-generated texts.

2.3 ConvAI2 Evaluation Guidelines

Designers of the competition argue for human evaluation metrics in addition to automated metrics in order to account for aspects of multi-turn conversations that humans consider important but are not fully taken into account if only automated metrics are used. These aspects include repetition, consistency and balance of dialog acts throughout the conversation (e.g. the amount of questions asked versus answered).

The competition first evaluated its submissions using automatic metrics such as perplexity and Hits@1. They then used additional human evaluations through MTurk by having Turkers chat with a given model and scoring its performance. Performance was based on how much the Turker enjoyed talking to the model and their ability to distinguish the persona used by the model from a random one.

3 Methods

3.1 Our Models

3.1.1 Baseline: With Personality Descriptions

The Baseline model that we use for comparison is HuggingFace's TransferTransfo approach from ConvAI2, which utilizes Transfer Learning to build a dialog agent based on the OpenAI GPT Transformer model. This model is a multi-task learner with two linear layer heads that sit on top of the transformer outputs: a language model (LM) head, and a multiple choice response (MC) head.

The transformer model is a 12-layer decoder-only transformer with 768 dimensional hidden states and 12 masked self-attention heads that only attend to the context of the already seen words on the left. The original pre-trained weights for the model are the result of training on the BooksCorpus data set (Zhu et al., 2015), which includes documents from roughly 7,000 unpublished books. This pre-training task enables it to learn from the context of long sequences of words to generate coherent sentences in the long-run.

The HuggingFace team fine-tuned this model to generate personality-aligned dialog for the ConvAI2 challenge by training on the PERSONA-CHAT data set containing personality descriptions (usually 4 to 6 sentences) as well as utterances from conversations between personalities. The inputs to this model for every utterance include the concatenated personality sentences of the current speaker, the last few utterances in the conversation history, as well as the reply that the current speaker generates in response to the history. After being tokenized (using the OpenAI GPT tokenizer), input embeddings to the transformer are generated for each input token by combining the pre-trained word embedding and positional embedding with a fine-tuned dialog state embedding that indicates whether the token is part of the personality description, an utterance from the current speaker, or an utterance from the speaker's partner.

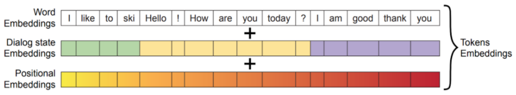


Figure 1: TransferTransfo's input representation. Each token embedding is the sum of a word embedding, a dialog state embedding and a positional embedding.

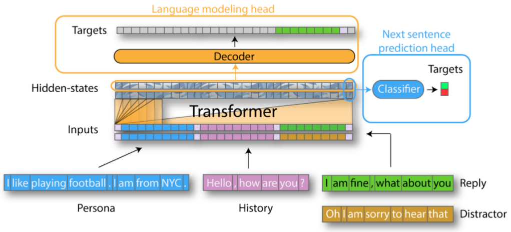


Figure 2: Multi-task training objective – the model is provided with two heads for language modeling prediction (orange) and next-sentence classification (blue)

The LM head takes the last hidden state outputted by the transformer as its input and returns the probabilities for the next token in the reply sequence. The LM loss is the negative log-likelihood loss comparing the next token probabilities to the gold next tokens. From the NLL, we calculate the perplexity of the gold reply as our baseline metric of comparison across the models.

In addition to the gold reply, each utterance also contains a list of distraction responses to the conversation history that are randomly sampled from the dataset. The MC head takes as input the hidden state that corresponds to the CLS token appended at the end of each reply and uses the state to classify the correct reply for that particular personality and conversation history. The MC loss is the cross-entropy loss comparing the reply probabilities to the gold reply. We use the classification accuracy as our baseline metric of comparison across the models.

The overall loss of the model is computed as a function of both the LM loss and the MC loss.

Our Baseline model implements the structure described above with minor modifications for compatibility with Transformers 4.4.2.

3.1.2 Baseline Named: Personality Description Removed

For our first experiment, we modified the original model to replace personality sentences with unique tokens or "names" for each personality. This version of the model (referred to as Baseline Named) learns the personality token embeddings during fine-tuning using only the context of the conversations associated with each personality.

3.1.3 TripleHead: Added Persona Head

Similar to Baseline Named, our TripleHead model replaced personality sentences with unique tokens for each personality. However, this final version of the model added a second multiple choice head for classifying the correct personality associated with a response. To use a similar approach to the MC head, we modified our data to also include a list of several distraction personalities randomly sampled from the other personalities as well as a gold personality associated with each utterance.

The Persona head takes as input the hidden state that corresponds to the unique personality token inputted at the beginning of each utterance and uses the state to classify the correct personality for that conversation history and gold reply. The Persona loss is the cross-entropy loss comparing the personality probabilities to the gold personality. We use the classification accuracy as our baseline metric of comparison across the models.

The overall loss of the TripleHead model is then computed as a function of the LM loss, MC loss, and Persona loss.

3.2 Data Sets

3.2.1 PERSONA-CHAT

We started with the full data set of persona chat which had 17878 different dialogs in the training and 1000 dialogs for validation. Overall we found 6113 unique persona sets which counts personas that differ slightly such as an additional sentence for a description as different personas. The dialogs are broken down into personality and utterances. The personality contains a list of 3-5 lines of persona descriptions like "i am a farmer", "i love comic books", or "i like long walks on the beach ." This was the description which the mTurk worker was given for the artificial persona to play as for the conversation. The utterances are broken down further into a list of dictionaries containing candidates and history. History is the past responses in the conversation so far which always ends with the other speaker's response. The candidates is a list of potential responses of the persona with the last in the list the actual reply of the persona.

3.2.2 Consolidated 10

The full PERSONA-CHAT data was too large to train on fully for our model so we subsetted the data by taking the dialogs of 10 personas. While trying to grab unique personas, we noticed that many of the personas were similar, with some just differing by one line in their persona description. Therefore we consolidated similar personas by preprocessing to correct small differences in the descriptions like "i m" to "i am" and grouping subsets of the same persona into the larger persona description. For example, the descriptions:

(a) ["i like to go hunting .", "i like to remodel homes .", "i like to shoot a bow .", "my favorite holiday is halloween ."]

(b) ["i like to go hunting .", "i like to go shopping with my daughters .", "i like to shoot a bow .", "my favorite holiday is halloween ."]

will both be included in the larger persona description as:

(c) ["i like to go hunting .", "i like to go shopping with my daughters .", "i like to remodel homes .", "i like to shoot a bow .", "my favorite holiday is halloween ."]

This resulted with 1420 unique personas out of the 6113 personas. With the consolidated personas we are able to have more training data for a given persona instead of smaller sets of data for a wider range of personas that only differ marginally. By sampling ten unique personas, we ended up with a training set of 107 dialogs, a validation set of 23 dialogs and a test set of 24 dialogs.

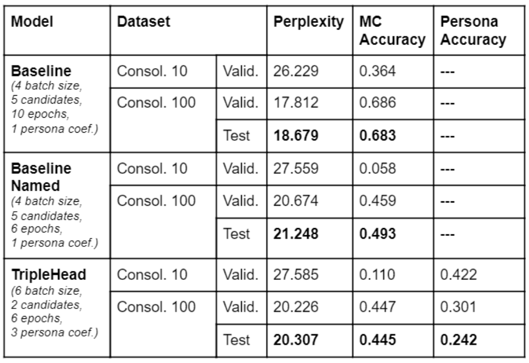
There are three different iterations of our data set with changes to our personas to fit our different models. First is the Baseline data set with the original data with only the slight change of using our consolidated persona description list as the personality. The second data set is the Baseline Named data which takes out the persona description entirely and replaces it with a persona token. Each persona token is mapped to a specific persona description. The third data set is the TripleHead data set with the personality being a list of four persona tokens where the last token is the gold persona. This is similar to how we constructed the candidates list where the last candidate in the list is the gold reply of the persona. All three data sets have the same dialogs and are connected to the same persona with different ways of expressing the personality for the model.

3.2.3 Consolidated 100

We repeated the process above to get three different data sets (baseline, baseline named, and triplehead) for a sample of 100 unique personas. With more personas and dialogs we hope the model will be able to learn and produce better results. For our sample of 100 unique consolidated personas we have 1067 dialogs in the training set with 229 different dialogs in both validation and test set.

4 Results and Discussion

To train and tune the hyperparameters of each model given the resources we had available, we ran the models on the Consolidated 10 data set as well as the Consolidated 100 data set. Across all models, the performance was significantly better on the Consolidated 100 data set than the Consolidated 10, which indicates that a sample of 10 unique personas likely did not have enough information for the model to learn from given that each persona had roughly 15 unique dialogs.



Model	Dataset		Perplexity	MC Accuracy	Persona Accuracy
Baseline *(4 batch size,* *5 candidates,* *10 epochs,* *1 persona coef.)*	Consol. 10	Valid.	26.229	0.364	---
	Consol. 100	Valid.	17.812	0.686	---
		Test	**18.679**	**0.683**	---
Baseline Named *(4 batch size,* *5 candidates,* *6 epochs,* *1 persona coef.)*	Consol. 10	Valid.	27.559	0.058	---
	Consol. 100	Valid.	20.674	0.459	---
		Test	**21.248**	**0.493**	---
TripleHead *(6 batch size,* *2 candidates,* *6 epochs,* *3 persona coef.)*	Consol. 10	Valid.	27.585	0.110	0.422
	Consol. 100	Valid.	20.226	0.447	0.301
		Test	**20.307**	**0.445**	**0.242**

Table 1: Model Results

4.1 Baseline: With Personality Descriptions

The Baseline model with personality descriptions performed relatively well (test metrics indicated in bold font), with a perplexity score of 18.679 and a gold reply classification accuracy (denoted as MC Accuracy in the table) of 0.683. The classifier chooses the correct reply out of 20 candidates, which means the classification accuracy reported is significantly above the statistical baseline of 0.05 (classification at random).

4.2 Baseline Named: Without Personality Descriptions

The Baseline Named model with personality tokens instead of descriptions does not perform as well as the Baseline with a significantly lower gold reply accuracy of 0.493. However, this model's perplexity score of 21.248 is relatively close to the original. These results seem to indicate that the language model head (responsible for next word prediction) is not strongly impacted by the removal of the personality description, however the next sentence prediction task is more directly impacted. The MC head is likely incorporating information about both the personality and the conversation history to choose a reply that fits in the context (which makes sense given our exploration of the PERSONA-CHAT data set showed a fair amount of conversations in which the speaker responded to the last utterance with a sentence directly from their persona description). Despite losing information about the personality, the Baseline Named model is able to learn enough from the inputs containing the fine-tuned personality token embeddings to achieve a classification accuracy above the statistical baseline of 0.05.

4.3 TripleHead: Added Persona Head

The TripleHead model with the additional layer for gold persona classification achieved a slightly lower test perplexity (20.307) compared to the Baseline Named model, but did not perform as well in terms of gold reply accuracy (0.445). This model included an additional persona accuracy metric to evaluate the new Persona head, and it appears to be the only metric of 3 that produces better results using the Consolidated 10 data set compared to the Consolidated 100. This may be due to the significant difference in the number of candidate distraction personas in each dataset; the Persona head for Consol. 10 has 9 distraction candidates to sample from, while Consol. 100 has 99 candidates, each of which have a small chance of appearing in the data because of the small personality-to-dialog ratio. Due to resource constraints, we kept the total number of personalities for the Persona head to choose from at 4, which means that our test accuracy for the gold persona classification was not significantly different from the statistical baseline of 0.25.

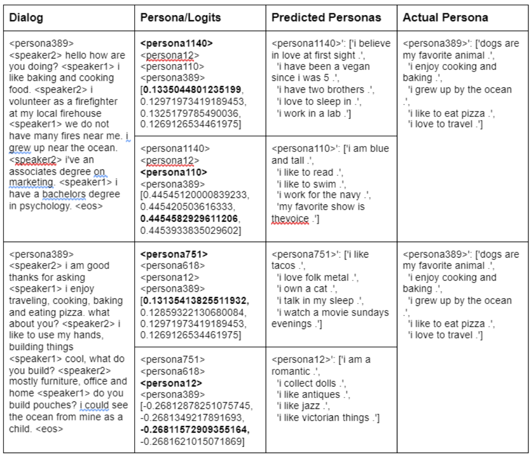
To check this metric against human performance, we also completed two manual evaluation tasks on a sample of the dialogs from our training data. Task 1 required the human evaluator to match the indicated (but unlabeled) speaker in a conversation with the appropriate personality sentences for that speaker (out of a selection of 4 sets of personality sentences). Task 2 required the human evaluator to match a conversation to one of 4 possible conversations based on whether or not the same personality is found in both. The second task did not include any personality descriptions and was solely based on matching complete dialogs (which was closer to the task that our Named model had to complete without personality sentences). Our human evaluation accuracy scores were 93% and 90% on average for each respective task, which means that humans perform significantly better on the persona classification task in the context of the given data.

Additionally, in some runs of the model we noticed an inverse relationship between the gold reply classification accuracy and the gold persona classification accuracy. This seems to suggest that the third head in the model may be competing with the second and causing the two to split their learning about the personalities unevenly. The authors of the original TransferTransfo paper also noted that increasing the importance of the gold reply classification score "results in utterances that stick more closely to the provided personality sentences but...reduce the diversity of the dialog," which supports our hypothesis that in the context of the PERSONA-CHAT data set (where replies were often derived directly from the persona descriptions), our ThirdHead approach may be competing rather than supporting the goal of the original MC head.

In summary, our results conclude that the Named approach (replacing personality sentences with unique tags) ultimately performed on par with or slightly better than the TripleHead approach in the context of the PERSONA-CHAT dialog task.

4.4 Error Analysis

To better understand our errors, we took random samples of misclassified dialogs and their errors while hypertuning the Consolidated 10 data. As we compared the dialogs with the correct persona and the predicted persona, we noticed that many of the same dialogs were frequently misclassified, a few of which are shown in the table below. It was interesting to note that the predicted personas did not remain consistent for the same misclassified dialogs. Upon closer examination of the personas that frequently got mixed up, their descriptions were fairly generic and likely resulted in vague indicators of the supposedly distinct personas. It would be difficult for the persona distinctions to be demonstrated prominently in the short PERSONA-CHAT conversations. Additionally, we noticed there were also specific personas out of the 10 in the Consolidated 10 data set that were frequently misclassified. Out of 25 randomly selected misclassified cases, 9 of them were dialogs spoken by Persona 389 and there were 7 times in which Persona 751 was incorrectly classified as having spoken. In contrast, our results from the Consolidated 100 data set did not face the same repeated misclassification issue as the larger data set provided a wider variety of personas and conversations to train on. However, it faced its own set of concerns where unique personas were difficult to train on because their unique attributes are difficult to even bring up in a short conversation. This was especially prominent in cases where the other speaker dominated the conversation, making it difficult for the individual to bring up unusual facts about themselves into the conversation. As a result, many of the defining attributes mentioned in the original descriptions never make it into the conversations for the model to train on.



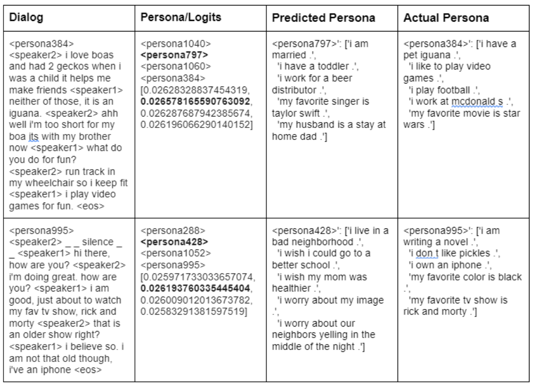
Dialog	Persona/Logits	Predicted Personas	Actual Persona
<persona389> <speaker2> hello how are you doing? <speaker1> i like baking and cooking food. <speaker2> i volunteer as a firefighter at my local firehouse <speaker1> we do not have many fires near me. i grew up near the ocean. <speaker2> i've an associates degree on marketing. <speaker1> i have a bachelors degree in psychology. <eos>	**<persona1140>** <persona12> <persona110> <persona389> [**0.1335044801235199**, 0.12971973419189453, 0.1325179785490036, 0.1269126534461975]	<persona1140>': ['i believe in love at first sight .', 'i have been a vegan since i was 5 .', 'i have two brothers .', 'i love to sleep in .', 'i work in a lab .']	<persona389>': ['dogs are my favorite animal .', 'i enjoy cooking and baking .', 'i grew up by the ocean .', 'i like to eat pizza .', 'i love to travel .']
	<persona1140> <persona12> **<persona110>** <persona389> [0.44545120000839233, 0.445420503616333, **0.4454582929611206**, 0.4453933835029602]	<persona110>': ['i am blue and tall .', 'i like to read .', 'i like to swim .', 'i work for the navy .', 'my favorite show is thevoice .']	
<persona389> <speaker2> i am good thanks for asking <speaker1> i enjoy traveling, cooking, baking and eating pizza. what about you? <speaker2> i like to use my hands, building things <speaker1> cool, what do you build? <speaker2> mostly furniture, office and home <speaker1> do you build pouches? i could see the ocean from mine as a child. <eos>	**<persona751>** <persona618> <persona12> <persona389> [**0.13135413825511932,** 0.12859322130680084, 0.12971973419189453, 0.1269126534461975]	<persona751>': ['i like tacos .', 'i love folk metal .', 'i own a cat .', 'i talk in my sleep .', 'i watch a movie sundays evenings .']	<persona389>': ['dogs are my favorite animal .', 'i enjoy cooking and baking .', 'i grew up by the ocean .', 'i like to eat pizza .', 'i love to travel .']
	<persona751> <persona618> **<persona12>** <persona389> [-0.26812878251075745, -0.2681349217891693, **-0.26811572909355164,** -0.2681621015071869]	<persona12>': ['i am a romantic .', 'i collect dolls .', 'i like antiques .', 'i like jazz .', 'i like victorian things .']	

Table 2: Examples of Misclassified Dialog from Consolidated 10 data set

5 Conclusion

The results from our models show that the Named approach (replacing personality sentences with unique tags) ultimately performed on par with or slightly better than the TripleHead approach in the context of the PERSONA-CHAT dialog task.

While our resource constraints prevented us from being able to run our model with greater variations in the parameters and a larger data set, the main area we would like to explore more in the future is how our model would perform with data actually designed for our particular use case. The PERSONA-CHAT data was originally created to capture distinct personas that could be learned by the model, but the reliance on persona descriptions resulted in the actual dialogs being short and often not fully encompassing the details included in the description. Our decision to consolidate similar personas allowed us to have a data set with more dialogs per persona, but it also made our remaining personas more broad as they needed to cover a wider range of dialogs. Additionally, MTurkers were prone to mentioning additional details that were not explicitly mentioned in the original persona description and these details vary by person. As our model does not have the persona description to reference, these new attributes are also taken into account by the model without distinction from the supposed main characteristics of the persona. There is also the possibility of MTurkers interpreting the same personalities differently as they try to connect the details in conversation or develop a backstory to justify bringing them up. Overall, we would be interested in trying out our model with larger amounts of data that allow our personas to be more fully developed and are better suited for our task.



Dialog	Persona/Logits	Predicted Persona	Actual Persona
<persona384> <speaker2> i love boas and had 2 geckos when i was a child it helps me make friends <speaker1> neither of those, it is an iguana. <speaker2> ahh well i'm too short for my boa its with my brother now <speaker1> what do you do for fun? <speaker2> run track in my wheelchair so i keep fit <speaker1> i play video games for fun. <eos>	<persona1040> **<persona797>** <persona1060> <persona384> [0.02628328837454319, **0.026578165590763092**, 0.026287687942385674, 0.026196066290140152]	<persona797>': ['i am married .', 'i have a toddler .', 'i work for a beer distributor .', 'my favorite singer is taylor swift .', 'my husband is a stay at home dad .']	<persona384>': ['i have a pet iguana .', 'i like to play video games .', 'i play football .', 'i work at mcdonald s .', 'my favorite movie is star wars .']
<persona995> <speaker2> _ _ silence _ _ <speaker1> hi there, how are you? <speaker2> i'm doing great. how are you? <speaker1> i am good, just about to watch my fav tv show, rick and morty <speaker2> that is an older show right? <speaker1> i believe so. i am not that old though, i've an iphone <eos>	<persona288> **<persona428>** <persona1052> <persona995> [0.025971733033657074, **0.026193760335445404**, 0.026009012013673782, 0.02583291381597519]	<persona428>': ['i live in a bad neighborhood .', 'i wish i could go to a better school .', 'i wish my mom was healthier .', 'i worry about my image .', 'i worry about our neighbors yelling in the middle of the night .']	<persona995>': ['i am writing a novel .', 'i don t like pickles .', 'i own an iphone .', 'my favorite color is black .', 'my favorite tv show is rick and morty .']

Table 3: Examples of Misclassified Dialog from Consolidated 100 data set

References

Dinan, E., Logacheva, V., Malykh, V., Miller, A. H., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I., Lowe, R., Prabhumoye, S., Black, A. W., Rudnicky, A. I., Williams, J., Pineau, J., Burtsev, M. S., & Weston, J. (2019). The second conversational intelligence challenge (convai2). *CoRR*, *abs/1902.00098*arXiv 1902.00098. http://arxiv.org/abs/1902.00098

Wolf, T., Sanh, V., Chaumond, J., & Delangue, C. (2019). Transfertransfo: A transfer learning approach for neural network based conversational agents. *CoRR*, *abs/1901.08149*arXiv 1901.08149. http://arxiv.org/abs/1901.08149

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? *CoRR*, *abs/1801.07243*arXiv 1801.07243. http://arxiv.org/abs/1801.07243

Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *CoRR*, *abs/1506.06724*arXiv 1506.06724. http://arxiv.org/abs/1506.06724