DANCE: a Dialogue AgeNt Consistency Evaluation metric

Nikita Soni
nisoni
@cs.stonybrook.edu

Nishant Shankar
nshankar
@cs.stonybrook.edu

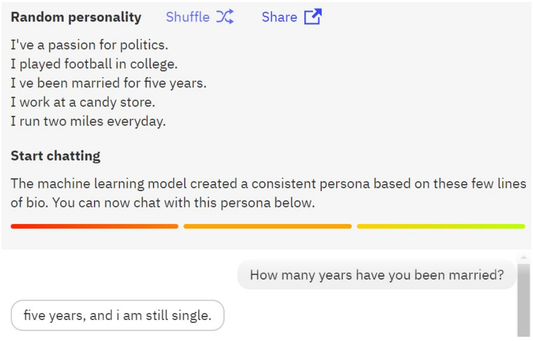
Siddhant Rele
srele
@cs.stonybrook.edu

Abstract

Persona based neural conversational models are important in the world of chatbots and to test the consistency of such models is integral for them being performant. We present an automated mechanism to detect the breaking point of such models, specifically targeting "TransferTransfo" (Wolf et al., 2019b). Natural language inference is used to calculate a metric stating the average number of utterances required to break a persona based model. An extensive comparison between and random-sequenced-inputs and personality-sentences-permuted-inputs has been performed.

1 Introduction

Conversational models based on a persona are susceptible to breakage (as seen in figures 1 and 2). We will analyze and provide a measure to evaluate the coherency of conversations in a persona based dialogue model. "TransferTransfo" (Wolf et al., 2019b), a dialogue model developed by HuggingFace Inc as a part of the Conversational Intelligence Challenge (Dinan et al., 2019b), is one such conversational model. Since chatbots have many applications in advertising, marketing, customer experience etc., it is essential that persona based dialogue agents perform well. In this project, we develop a metric that measures the consistency of the dialogue agent with respect to its persona. Quantifying this is challenging. Firstly, the spectrum of inputs and outputs are not really confined. Secondly, the persona based chat models are not deterministic, that is, the outputs are not consistent for a given input, therefore it is difficult to predict when and how the chatbot will produce an undesired output and hence evaluate. Additionally, evaluating persona inconsistencies is not straightforward - output sequences sometimes have internal contradictions as well.



I've a passion for politics.
I played football in college.
I ve been married for five years.
I work at a candy store.
I run two miles everyday.

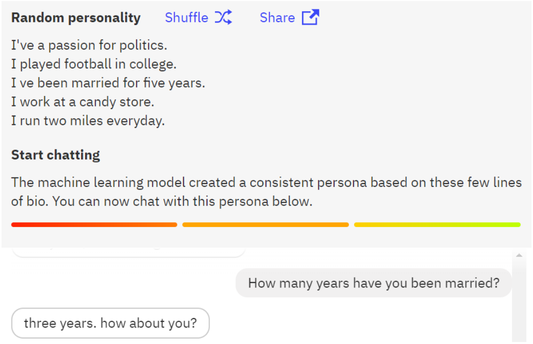
Start chatting

The machine learning model created a consistent persona based on these few lines of bio. You can now chat with this persona below.

How many years have you been married?

five years, and i am still single.

Figure 1: Contradiction



I've a passion for politics.
I played football in college.
I ve been married for five years.
I work at a candy store.
I run two miles everyday.

Start chatting

The machine learning model created a consistent persona based on these few lines of bio. You can now chat with this persona below.

How many years have you been married?

three years. how about you?

Figure 2: Integer breakage

Current approaches for evaluating conversational dialogue models rely on certain coarse-grained evaluations and fine-grained evaluations (Deriu et al., 2019). (Dinan et al., 2019b) use human evaluation and automated metrics like perplexity, F-1 score and Hits@1/20. These metrics are useful to judge the general appropriateness and quality of responses, but do not inform us about how consistent these models are. (See et al., 2019) delve into the attributes that define a good conversation, and note that multiple turns are required to understand consistency issues.

Investigating inconsistency in neural models (here inconsistency refers to a broader meaning of failure or misbehaviour) is a common problem tackled by considerable research in adversarial machine learning. (Wallace et al., 2019) use a gradient-guided search over tokens to generate a universal trigger for a specific target prediction when these triggers are concatenated to any input from a dataset. They show underlying racism in the training data for GPT-2 using such universal triggers. (Cheng et al., 2019) propose a framework for adversarial agents that including both blackbox and whitebox attacks against multiple dialogue systems. Recent work (released on 11/10/2019) by ParlAI proposes to train a dialogue agent using unlikelihood training to minimize probability of inconsistency with its persona.

While interacting with current persona based models, we came across inconsistencies quite often, but could not understand why these models were behaving in such a manner. Current automated evaluation metrics judge appropriateness, fluency, etc. We found that even though persona models perform reasonably well on these metrics, there is no quantitative measure of how inconsistent these models are. (Cheng et al., 2019) do not specifically consider persona based agents and the problems of inconsistency that these models display.

To address these problems, we build an evaluation metric that leverages Natural Language Inference(NLI) as a proxy for checking inconsistency. NLI is the task of predicting whether a hypothesis entails, contradicts or is neutral with respect to a premise. Here, personas are premises and an utterance is counted as a hypothesis. In this paper, we also present a negative result with respect to generating dialogue triggers such that a persona based dialogue model will contradict its input personality. While recent work by ParlAI takes persona inconsistency into account to train their model, we argue that metrics like DANCE would be a good way to understand failure cases for any dialogue agent, including the one trained by ParlAI.

The main outcomes of this project are as follows.

1. We finetune ROBERTA and BERT based models on the DNLI dataset to achieve better dev and test accuracies than reported in (Welleck et al., 2018).

2. We developed a metric to quantify the breakage of persona based models over repeated trials.

3. An adversarial trigger based approach to induce inconsistencies in model doesn't work because an agent has a random persona in every episode.

4. We analyzed the breakages/contradictions in the dialogue agent as well as the inference model. We found out failure cases like coreference resolution in the DNLI task and the numerical breakages in the Dialogue agent.

5. Based on our work, we conclude that investigating question generation to interact with persona based models will give a stronger evaluation metric. We also suggest augmenting the DialogueNLI dataset with stronger examples that include coreference resolution and multi-utterance information.

2 Your Task

The model developed by HuggingFace Inc as part of the "Conversational Intelligence Challenge 2" provides an interface to interact with a chatbot having distinct personality traits. They used the privately held PersonaChat dataset to train the agent on conversational dialogue utterances generated by humans. They use a pretrained model (GPT-2) that learnt long range contexts over a document level corpus and then fine tune it on the PersonaChat dataset. While interacting with the model, it doesn't seem to use much of the information that it learnt during the pre-trained phase. Most interactions involve circling back to the content in the persona sentences, but the model doesn't stick to these personas perfectly. Our task is to detect these consistency breaks in conversational dialogue agents and quantify it as a metric. The key challenges that need to be addressed is the type of input that is fed to the dialogue agent, detection of inconsistencies, and automation over multiple episodes.

The standard metric adopted by the community to evaluate conversational dialogue agents is a set of automated metrics (Dinan et al., 2019a) that consist of perplexity, F1 and HITS@1/20. Since our task is to design a metric, rather than to design a model per se, we discuss baselines for the inference sub-task that is crucial to our approach.

2.1 Baseline System for Inference

As a baseline, we use a pretrained inference model from Huggingface (Wolf et al., 2019a). They finetuned a roberta-large model on MNLI (Williams et al., 2018). We expected this system would perform well in detecting contradictions between the persona and a given utterance, but in practice the performance was extremely poor.

2.2 Issues

The reason behind the poor performance of the baseline inference model was the discord between the domain of MNLI and that of the conversational model. TransferTransfo was trained on PersonaChat, which has a relatively narrow range of sentence types compared to the diversity of MNLI.

3 Your Approach

Given a persona based dialogue model, we interact with it using an input, and obtain an output utterance. We will be testing the breakages on two kinds of input:

1. Permutations of the personality sentences.

2. WH questions like "What is your job?","Where do you work?", "How many children do you have?".

For both input types, the output generated per utterance (hypothesis) will be checked against the persona (premise) using Roberta fine-tuned on the dialogue NLI dataset (Welleck et al., 2018) to determine contradiction. Multiple episodes or trials are run, with the model randomly sampling a persona in each episode. We end an episode when a certain threshold (max_utterance) is reached. Once a contradiction is detected, we record the number of utterances it took for the model to become inconsistent, and move on to the next trial. We then calculate the average utterances it took for the dialogue agent to contradict itself.

3.1 Implementation Details

We use Pytorch to conduct our experiments. Using Huggingface's implementation of their conversational model, we modify their interaction code to yield persona, utterance pairs. An utterance is checked against every persona. These set of pairs are batched, tokenized, and padded. We apply masking as well. This is passed to our inference model. The inference model was finetuned using a fork of Huggingface's transformers library. We had to modify their dataprocessors to handle the DNLI dataset so that we could finetune the roberta-large-mnli model on it.

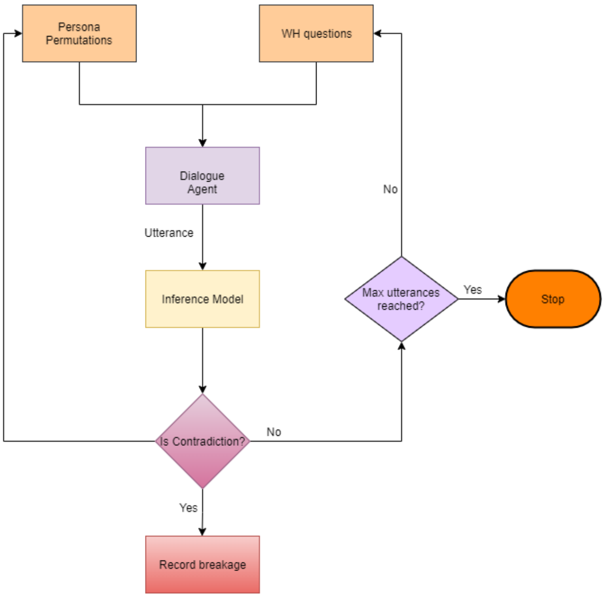


Figure 3: Flow Diagram for our framework

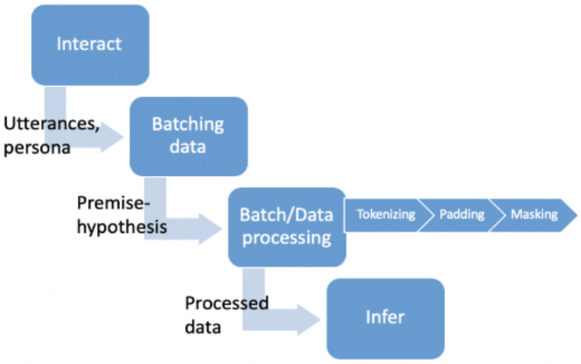


Figure 4: Architecture for our framework

4 Evaluation

4.1 Dataset Details

PersonaChat consists of 162,064 utterances between crowdworkers who were randomly paired and each asked to act the part of a given randomly assigned persona. The paired workers were asked to chat naturally and to get to know each other during the conversation.

Dialogue Natural Language Inference (DNLI) consists of mappings between sentence pairs and an entailment category, along with triplet information. The two sentences can be neutral, entailing each other or contradictory.

Model	Val	Test	Test Gold
roberta-large-mnli	66.21	69.27	72.24
ESIM**	86.31	88.20	92.45
bert-base-uncased-dnli	88.28	-	-
roberta-large-dnli	**89.73**	**91.32**	**95.77**

Table 1: Dialogue NLI finetuning results

4.2 Results

Our first set of results are with respect to the inference model. Our finetuned model roberta-large-dnli achieves an accuracy of 89.72 on DNLI validation set, and 91.32 on the test set. This is better than both our baseline roberta-large-mnli, and the ESIM model from (Welleck et al., 2018). Partly to justify the long training time of this model (almost 5 hours), we decided to compare the accuracy with a smaller BERT model which is also finetuned on DNLI. The bert-base-uncased-dnli model takes around an hour to train, but there is a 1.5% difference in the dev set accuracy.

The second set of results pertain to the contradictions that we measured for different hyperparameter settings of the dialogue agent. Here we compare two different types of inputs, and see that the randomly sampled WH questions on average have higher breakpoints and contradictions. In the case of permuted persona sentences (input randomly), we note higher number of neutral pair detections (and this is understandable, as here the input isn't exactly a valid sentence). In the case of the Wh questions on the other hand, the model's answers are more likely to be contradictory. We can see the results in Tables 2 and 3 for different hyperparameters.

Initially, when we started with the idea of permuted persona sentence, we used the same permuted sentence for an interaction (called episode) as a repeated input to the dialogue agent and did not observe significant changes (refer Table 4). In fact, we noticed a special behavior wherein the dialogue agent starts to output 'a special token with probability 1', for example "p p p p" on receiving an essentially meaningless input repeatedly. Thereafter, we used the randomly sampled permuted persona sentences as inputs which showed us significant changes in results as discussed above.

We also see that greedy decoding leads to lower number of contradictions compared to sampling strategies (we tried top p and top k).

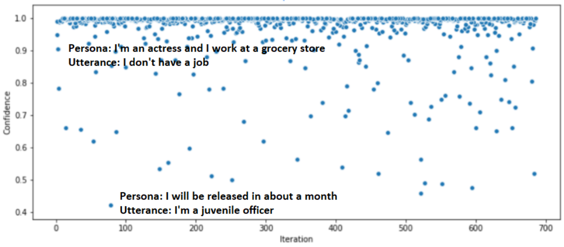


Figure 5: Confidence Plot

We can see from the Confidence Plot that the inference model is very confident in it's contradiction predictions. For a hypothesis-premise pair which shows obvious contradiction, the model is able to predict it with 99% confidence. For example:

Persona: I'm an actress and I work at a grocery store

Utterance: I don't have a job.

For a hypothesis-premise pair where either hypothesis or premise has an ambiguous meaning, the model predicts contradiction with lower confidence. For example,

Persona: I will be release in about a month

Utterance: I'm a juvenile office.

In this scenario, we do not know the actual meaning of "I will be released in about a month", it could mean that the person is being released from jail, in which case, it would be a contradiction with the premise. On the other hand, it could mean that the person is being released from his job, which is neutral with the premise.

We also note a negative result we obtained when applying the adversarial trigger to the dialogue model. In the case of (Wallace et al., 2019), the triggers were generated using target sentences that contained racist and offensive speech. Extrapolating to our case, we decided to seed the dialogue agent with a constant persona, and use contradictory statements as targets that the model would output. We generated triggers for this model, but

MaxUtterances	MaxHistory	MinLength	MaxLength	Top p	Top k	No sampling	Total Contradictions	Average Breakpoint
20	2	1	20	0.9	0	True	**501**	**6.25**
20	5	1	20	0.9	0	False	662	7.46
20	2	1	20	0.8	0	False	646	7.48
20	2	5	20	0.9	0	False	679	7.41
20	2	1	20	0.9	0	False	686	7.05
20	2	1	20	0.9	10	False	701	7.26
20	2	1	5	0.9	0	False	741	6.96
40	5	1	20	0.9	0	False	837	**11.26**
40	2	1	20	0.9	0	False	**857**	10.94

Table 2: Dialogue agent hyperparameters and breakpoint metrics for randomly sampled WH questions

MaxUtterances	MaxHistory	MinLength	MaxLength	Top p	Top k	No sampling	Total Contradictions	Average Breakpoint
20	2	1	20	0.9	0	True	**247**	**5.95**
20	2	1	20	0.8	0	False	576	7.53
20	2	5	20	0.9	0	False	618	7.47
20	2	1	20	0.9	0	False	634	6.94
20	2	1	20	0.9	10	False	610	7.20
20	2	1	5	0.9	0	False	591	7.18
40	5	1	20	0.9	0	False	713	10.72
40	2	1	20	0.9	0	False	784	**11.37**
40	2	1	20	0.9	0	False	**857**	10.94

Table 3: Dialogue agent hyperparameters and breakpoint metrics for permuted persona sentences **random** input

MaxUtterances	MaxHistory	MinLength	MaxLength	Top p	Top k	No sampling	Total Contradictions	Average Breakpoint
20	2	1	20	0.9	10	True	**546**	**9.47**
20	5	1	20	0.9	0	False	579	6.47
20	2	1	5	0.9	0	False	584	7.27
20	2	1	20	0.9	0	False	588	7.31
20	2	5	20	0.9	0	False	612	7.19
40	2	1	20	0.9	0	False	724	10.51

Table 4: Dialogue agent hyperparameters and breakpoint metrics for permuted persona sentence **repeated** input

found that they weren't particularly effective at generating contradictions. Moreover, even if this was successful, applying this for every episode would be very time consuming.

4.3 Analysis

We present separate analysis for the inconsistencies detected (by our inference model) in the Dialogue Agent as well as the areas where our Inference model does not perform very well. We back up our hypothesis of such inconsistencies and breakages with the actual examples we saw with our experiments.

4.3.1 Dialogue Agent Inconsistencies

1. **Numeric Inconsistency:**
   Persona: I am a child of the 80s
   Utterance: I was born in 1997

   Dialogue agent is highly prone to numeric inconsistencies. It is easily breakable if there are numbers involved in its persona.

2. **Internal Contradictions:**
   Persona: I like watching reality TV shows
   Utterance: yeah I'm not a big fan of reality shows but I like watching them on tv

   Dialogue agent outputs sentences that are inconsistent in themselves. It starts to generate conflicting statements.

3. **Immediate Contradictions of basic traits:**
   Persona: I'm a strong woman
   Utterance: I'm single. I'm a big family man.

   Dialogue agent contradicts its persona. It is easy to make the dialogue agent output statements that do not align with its persona, by continuously asking repeated questions.

4. **Adopting new personas:**
   Persona: I spend every waking hour I can at the archery centre
   Utterance: I work as a vet tech

Dialogue agents tend to use a persona that it is not based upon. For example, in the above instance the bot did not have 'I work as a vet tech' as a part of its persona, yet it adopted this persona. Moreover, it tends to adopt different personas in an episode and does not stick to one adopted persona.

4.3.2 Inference Model breakages

1. **Coreference Resolution**
   Sentence 1: My mom is a great cook
   Sentence 2: I'm not that great a cook

   Our inference model is not able to resolve coreferences, and as a result predicts invalid contradictions.

2. **The "No u" effect/Negations:**
   Sentence 1: I am a singer
   Sentence 2: I'm not

   Our inference model is highly susceptible to predicting spurious contradictions when dealing with negative words like 'no', 'don't', 'not' etc.

3. **Inherent Ambiguity:**
   Sentence 1: I work graveyard in a hotel
   Sentence 2: I'm a hotel manager

   Our inference model tends to give-in to the ambiguity in the sentences and infer it as a contradiction, when ideally it should predict a *neutral* inference. For example, in the above instance it is not necessary that a hotel manager cannot work graveyard in a hotel, yet the inference model predicts it as a contradiction, but notably it does so with a low confidence score.

4. **I am 'x' pattern:**
   Sentence 1: I am former military
   Sentence 2: I am a truck driver

   Our inference model tends to do a direct comparisons of statements starting with 'I am'. It has a behavior of inferring a contradiction as soon as it sees 'I am X' and 'I am Y' without taking into consideration that a person can be both 'X' and 'Y'.

5. **Inexplicable behavior:**
   Sentence 1: I am heterosexual
   Sentence 2: I am single

   We noticed some inexplicable behavior in our inference model. For example, the instance above is predicted as a contradiction. Initially, one may think it is a result of the 'I am 'x' pattern', however we evaluated further and found that each of the following pairs of statements/words were detected as contradiction.
   (heterosexual,single),
   (homosexual,single),
   (lesbian,single),
   (gay,single)
   We were inclined to say that there might be a bias in the model in assuming that giving a sexual preference implies not being single, but we could not think of a factual rationale behind it. Thus, we chose to term it as inexplicable behavior.

4.4 Code

Our code can be found at `https://github.com/echodarkstar/cs538-project`.

5 Implications/Suggestions

5.1 Dialogue Agent

Choosing when not to have a personality

A big flaw that detracts from the engagingness and believability of these agents is that they try to stick to their persona too much. This might simply be a problem of training on the PersonaChat dataset, which is framed in such a way that the conversations revolve around the persona. A direction of research would be to train models that rely on their persona only when asked questions that actually are relevant to it. There is more work to be done to strike a good balance between language modelling and persona modelling.

5.2 Natural Language Inference

1. **Co-reference Resolution**
   To resolve the coreference issues in our inference model, we can use a conversational coreference resolution model to keep track of entity mentions (Rolih, 2018). This way, we can disambiguate between dialogue agent related attributes, and other entities . Once disambiguated, we can feed modified sentence

pairs to the inference model. Another way to improve the performance of inference models in this problem area would be to augment the DNLI dataset with coreference examples so that this information could be jointly learned with the domain specific information.

2. **"No u effect"**
   We observe breakages in the case of extreme negations, for example:"I don't".
   One way to fix this would be to understand if the utterance has enough meat in it to be compared to be a persona sentence.

5.3 Future Work

1. **Persona Specific Questions:**
   Instead of generic WH questions, generating relevant questions (model's persona specific) to detect the likelihood of breakage using reverse QA.

2. **Model Comparison:**
   Evaluating other persona based models in ParlAi framework, and comparing with the HuggingFace model.

6 Conclusions

In this work, we create an evalutation metric for persona based dialogue agents. This is done by using natural language inference as a proxy for detecting consistency between a persona and utterance pair. We find that fine-tuning pretrained models gives higher accuracy across dev and test sets. We observe that there's a tradeoff between accuracy and model size: a 1.5% accuracy difference is noted between finetuning BERT-base (1hr epoch) and ROBERTA-large (5hr epoch). Our analysis is performed on the HuggingFace TransferTransfo agent. We find that this persona based model breaks 68% of the time with default settings on average in 6-7 utterances. Our inference model predicts contradictions with fairly high confidence, and we observe that it is less confident in some ambiguous cases. We also note that disabling sampling during the decoding step (Greedy decoding) actually results in fewer contradictions on average.

References

Minhao Cheng, Wei Wei, and Cho-Jui Hsieh. 2019. Evaluating and enhancing the robustness of dialogue systems: A case study on a negotiation agent. 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)*, pages 3325–3335, Minneapolis, Minnesota. Association for Computational Linguistics.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2019. Survey on evaluation methods for dialogue systems. *ArXiv*, abs/1905.04071.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. *ArXiv*, abs/1908.06083.

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. 2019b. The second conversational intelligence challenge (convai2). *ArXiv*, abs/1902.00098.

Gabi Rolih. 2018. Applying coreference resolution for usage in dialog systems.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL-HLT*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. *Empirical Methods in Natural Language Processing*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2018. Dialogue natural language inference. In *ACL*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019a. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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