Explaining Dialogue Evaluation Metrics using Adversarial Behavioral Analysis

Baber Khalid
Rutgers University
110 Frelinghuysen Rd
Piscataway, NJ 08854
baber.khalid@rutgers.edu

Sungjin Lee
Amazon Alexa
300 Pine Street
Seattle, WA 98103
sungjinl@amazon.com

Abstract

There is an increasing trend in using neural methods for dialogue model evaluation. Lack of a framework to investigate these metrics can cause dialogue models to reflect their biases and cause unforeseen problems during interactions. In this work, we propose an adversarial test-suite which generates problematic variations of various dialogue aspects, *e.g.* logical entailment, using automatic heuristics. We show that dialogue metrics for both open-domain and task-oriented settings are biased in their assessments of different conversation behaviors and fail to properly penalize problematic conversations, by analyzing their assessments of these problematic examples. We conclude that variability in training methodologies and data-induced biases are some of the main causes of these problems. We also conduct an investigation into the metric behaviors using a black-box interpretability model which corroborates our findings and provides evidence that metrics pay attention to the problematic conversational constructs signaling a misunderstanding of different conversation semantics.

1 Introduction

Automatic evaluation of natural language models in general and dialogue models in specific has been a focus of ongoing research. The gold standard for evaluation of dialogues is human judgement (Meena et al., 2014; Ultes et al., 2013; Jang et al., 2020; Shim et al., 2021; Khalid et al., 2020b; Panfili et al., 2021) but human judgements are hard to obtain. Other than human judgements, dialogue simulations are used to judge different aspects of a model behavior (Jung et al., 2009; Eckert et al., 1997; Cuayáhuitl et al., 2010; Khalid et al., 2020a; Kreyssig et al., 2018; Sun et al., 2021a). Neural models of dialogue rely on text similarity metrics like BLEU, ROUGE or METEOR (Papineni et al., 2002; Lin, 2004; Banerjee and Lavie, 2005), however these do not correlate well with the human judgement (Lowe et al., 2017).

A: Do you like singing?
B: Yes, I do.
A: Let's go to a ETV tonight.
B: But I can't sing it well.
A: It doesn't matter. No one will laugh at you.
B: OK, I'Il go with you. When shall we go?
Actual Response: What about six?
Adversarial Response: OK. Let's make it.
Human Score: 0.05; **Adversarial Score:** 0.43

Table 1: This example presents a problematic judgement by the DialogRPT metric. This metric scores an irrelevant response higher than a relevant and natural response. This example highlights the volatile nature of neural evaluation metrics.

Recent research focuses on the use of neural networks to tackle this problem (Kreyssig et al., 2018; Sun et al., 2021a; Lowe et al., 2017; Jang et al., 2020; Shim et al., 2021; Mehri and Eskenazi, 2020a; Pang et al., 2020; Gao et al., 2020; Kachuee et al., 2021). However, neural methods are known to be 1) poor at dealing with out-of-distribution data, 2) very hard to explain and 3) hard to train because of data-availability issues. An example of poor behavior by a neural evaluation metric, DialogRPT (Gao et al., 2020), is shown in the table 1. As these metrics are used in conjunction with different training methods to improve the quality of intelligent conversation models (Kachuee et al., 2022; Park et al., 2021), these problems and biases can trickle down into the trained models. Therefore, it is important to formulate techniques which can fix the existing problems in the evaluation metrics.

In this work, we present an adversarial test-suite which uses automatic heuristics to generate adversarial examples targeting specific aspects of dialogues *e.g.* logical entailment or natural vocabulary. Performance analysis of metrics on these adversarial examples provides insights into their assessments of different problematic behaviors and lets us deduce problems they face while judging these behaviors. We also use a black-box interpretability

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5871 - 5883

technique, a modification of Kernel SHAP from Lundberg and Lee (2017), to find important language features for the metrics to provide additional insights into the behavior of evaluation metrics. This test-suite is meant to provide a novel benchmark for community which can be used to analyze the performance of proposed metrics and can be improved with further research.

2 Related Work

2.1 Language Model Evaluation Test-Suites

There have been several test-suites which compare the performance of the language models on contrasting examples. Both Gauthier et al. (2020); Warstadt et al. (2020) compute *surprisal* for sentence pairs, where one of the pair has a syntactical mistake, to test if a language model finds the syntactically wrong sentence more *surprising*. Beyer et al. (2021a); Pishdad et al. (2020) both rely on this concept of calculating surprisal but use pairs of coherent and in-coherent language uses. However, they don't release heuristics for automatic generation of adversarial cases and just focus on coherence-based manipulations while we go one step further to see the effect of other manipulations on various core aspects of dialogue in an automated way. Ribeiro et al. (2020) presents a tool which evaluates language models with their performance on pre-determined tests and their outcomes.

2.2 Adversarial Evaluation Techniques

Previous works have tried to use adversarial evaluations to judge the performance of dialogue models. Cheng et al. (2019) successfully trains a RL agent using adversarial rewards against a negotiation dialogue agent and reduces its effectiveness. Jia and Liang (2017) proposes an adversarial attack where adding extra sentences in the comprehension text reduces the performance of comprehension models significantly. The closest work to ours is Sai et al. (2019) which evaluates a neural metric (Lowe et al., 2017). It shows effectiveness of simple syntactical manipulations, like reversing a sentence, in fooling the metric. We, however, rely on attacking more complex semantics using simple heuristics, like breaking co-reference chains, to pinpoint metric performance fluctuations.

2.3 Interpretability Techniques

There have been several works proposing algorithms to explain the predictions of neural networks. A measure of co-operative game theory used to measure marginal contribution of players in a game is called Shapley value (Shapley, 1952). It has been a focus of attention for ML community to explain predictions of neural models. LIME (Ribeiro et al., 2016) and trumbelj and Kononenko (2013) present methods which rely on feature pertubrations to generate explanations. Bach et al. (2015) presents a method which provides explanation in form of pixel-based heatmaps. Lundberg and Lee (2017) modifies the methods mentioned earlier to approximate Shapley values. Li et al. (2017a) presents the variations in model outputs by erasing several input features as a measure of importance. We use a modification of Kernel SHAP from (Li et al., 2017a) to approximate feature importance in this work.

3 Conversation Properties of Interest

Dialogue systems– specially those based on neural architectures, show poor performance in generating consistent responses by keeping track of information in a dialogue context and are prone to generating repetitive, unnatural and bland responses (Yang et al., 2021; Khandelwal et al., 2018; Gao et al., 2018). These problems also guide the research in the area of dialogue system evaluation with many metrics focusing on evaluating coherence and consistency of dialogue models (Lai and Tetreault, 2018; Beyer et al., 2021b; Gao et al., 2020; Pang et al., 2020; Sun et al., 2021b). These findings help us choose relevant dialogue attributes to manipulate so the insights from the performance of different metrics can be used to address different problems in dialogue modeling.

To generate adversarial examples we focus on the following aspects of conversations: 1) *coherence* 2) *naturalness* 3) *interestingness*. We manipulate each of these attributes by targeting specific aspects which contribute to these: **Coherence:** i) *entailment* ii) *pronoun use* iii) *named-entity use* iv) *co-reference* v) *contradictions* vi) *speaker sensitiveness* **Naturalness:** vii) *unnatural repetitions* viii) *vocabulary diversity* ix) *unnatural paraphrasing* x) *natural paraphrasing* xi) *entrainment* **Interestingness:** xii) *dullness*. One example for manipulation of each of the major attributes is shown in the table 2 and other examples are presented in the appendix A.

Coherence	Naturalness	Interestingness
Contradiction	Unnatural Paraphrasing	Dullness
A: Could you explain what you saw? **B**: I was in the bank at the time of the robbery. **A**: What did you see? **B**: I saw a man come in with a gun. **A**: Did you see his face? **Actual**: No . He had a mask on. --- **Adversarial**: I did not see a man come in with a gun.	**A**: You saved my life yesterday, Rachel. I can't believe I forgot to bring my wallet when we went to lunch with those clients. **B**: It was a good thing I had enough on me. **Actual**: Let me buy you lunch today to pay you back. --- **Adversarial**: allow me purchase you luncheon today to give you back.	**A**: What are you doing tonight? **B**: I have to run to the grocery store. **A**: Don't you hate fighting the crowds on the weekends? **Actual**: Yes, but I am out of food and milk. --- **Adversarial**: I don't have a good answer to that.

Table 2: This table showcases one adversarial example for each dialogue attribute we target: coherence, naturalness and interestingness.

4 Adversarial Techniques

We specifically focus on four conversational datasets *Daily Dialogue (DDial)* (Li et al., 2017b), *Persona Chat (PerCh)* (Zhang et al., 2018), *Reddit* (Gao et al., 2020) and *MultiWOz* (Budzianowski et al., 2018) to test our heuristics. We test an implementation of our heuristics using these datasets and release it with this work[1].

4.1 Problem Definition

In accordance with the given datasets, we consider a conversation D as a series of utterances between two alternating speakers $(x_1, y_1, x_2, y_2, ..., x_n, y_n)$. We formulate the adversarial conversation generation problem as follows: Given a conversation snippet $C_i = (x_1, y_1, ..., x_i, y_i, x_{i+1})$, from a randomly sampled conversation $D = (x_1, y_1, ..., x_n, y_n)$, we generate an adversarial conversation snippet $C'_i = (x_1, y_1, ..., x_i, y_i, x'_{i+1})$. This helps us learn the impact of different responses to the same context on neural conversational metrics. x'_{i+1} is generated using adversarial heuristics which are explained as follows:

4.2 Coherence

The purpose of these attacks is to disturb the logcial flow of human conversations and investigate how different evaluation metrics react to these disturbances.

4.2.1 Entailment

In this case, the adversarial response x'_{i+1} is either a randomly sampled utterance from a random conversation in the dataset or a random response from the following: $(y_{i+1}, x_{i+2}, ..., x_n, y_n)$ from D. (See Appendix A for example.)

[1]https://github.com/baber-sos/Explaining-Dialogue-Evaluation-Metrics-using-Adversarial-Behavioral-Analysis

4.2.2 Pronoun Use

Given $x_{i+1} = (w_1, w_2, ..., w_k)$, where w_j is an utterance token, we manipulate $w_j \subset P$, where P is the set of pronouns, in the following manner: 1) We convert gender specific pronouns to object pronouns e.g. he/she to it/they and vice versa. 2) We convert singular pronouns *it/this/I* to the plural ones and vice versa. 3) We convert 1st person pronouns like *I/we/our* to a random 2nd/3rd person pronoun or vice versa. (See Appendix A for example.)

4.2.3 Named Entities

We replace named entities in x_{i+1} with 1) another random entity of same type *e.g.* person with person 2) a named entity of different type *e.g.* person with company to generate x'_{i+1}. (See Appendix A for example.)

4.2.4 Co-Reference

Human conversations have extensive use of references to different entities being discussed. We use a pre-trained reference resolution model to detect co-reference clusters in the conversation snippet and replace references in x_{i+1} to generate x'_{i+1}. This results in adversarial x'_{i+1} which means the same thing as x_{i+1} and should be assigned a similar score by an ideal metric. (See Appendix A for example.)

4.2.5 Speaker Sensitiveness

Given a conversation snippet C which ends at a response x_{i+1} to a question-answer pair (x_i, y_i), we augment the response x_{i+1} by concatenating it to the answer y_i to the question x_i to generate the adversarial $x'_{i+1} = [y_i, x_{i+1}]$. This creates situations which contain unnatural redundancies because a speaker does not acknowledge the answer

to a given question. A simple and effective heuristic to detect question answer pairs is to find an utterance with the question-mark ?. (See Appendix A for example.)

4.2.6 Contradictions

Here we focus on generating adversarial x'_{i+1} which contradicts the contextual information in a given conversation C. We generate x'_{i+1} by augmenting x_{i+1} in the following manner: 1) by negating the verbs in x_{i+1} using simple heuristics, 2) by replacing named entities in x_{i+1} occurring at least twice in C with a random entity of same type or 3) by replacing references in x_{i+1} to different named entities with the random entities of same type. For the verb negation, we either add a *not* after an auxiliary or add *did not* and *does not* depending on plurality of the subject and tense of the sentence while changing the verb form if needed. (See the table 2 for example.)

4.3 Naturalness

Humans prefer some vocabulary or style choices over others and a good evaluation metric should respect this behavior. We disturb these natural choices in the following manner:

4.3.1 Unnatural Repetitions

We augment x_{i+1} to generate x'_{i+1} by randomly sampling and repeating k times 1) non-stop English words in x_{i+1} or 2) multi-word noun phrases to generate unnatural repetitions in the adversarial response. (See Appendix A for example.)

4.3.2 Vocabulary Diversity

We replace randomly sampled words w_j, which have a synonym in the conversation C, from x_{i+1} with one of their synonyms in C. This results in x'_{i+1} with vocabulary choices which may not be as natural. An ideal metric would assign these examples a score no better than their human generated counter-parts. As a simple heuristic, we consider two words (w_j, w_k) as synonyms if they are in synset of each other as specified in the Wordnet(Miller, 1995). (See Appendix A for example.)

4.3.3 Unnatural Paraphrasing

We generate unnatural paraphrase x'_{i+1} by replacing randomly sampled non-stop English words from x_{i+1} with their synonyms from Wordnet. For synonym sampling, we rank the Wordnet synonyms according to word2vec embedding (Mikolov et al., 2013) similarity and sample from the top k synonyms. To make paraphrasing unnatural, $1/4^{th}$ of words are replaced with synonyms which are least likely to be used by humans while keeping it grammatically correct. (See the table 2 for example.)

4.3.4 Natural Paraphrasing

We generate natural paraphrases using a T5 model fine-tuned on the *PAWS* (Zhang et al., 2019) dataset. T5 model makes minor structural changes to x_{i+1} like removing punctuation or re-organizing utterances while using same vocabulary. An ideal metric should assign both x_{i+1} and x'_{i+1} similar scores. (See Appendix A for example.)

4.3.5 Entrainment

To disturb entrainment, we replace non-stop English words $(w_j \in x_{i+1})$ used by both parties involved in the conversation with a synonym sampled from the top k synonyms in the Wordnet. (See Appendix A for example.)

4.4 Interestingness

Interestingness is a subjective measure but humans prefer interesting conversations over the dull ones. We consider those responses *interesting* which progress the conversation in a natural way than those which do not add anything meaningful to it.

4.4.1 Dullness

To generate a dull x'_{i+1} we either 1) replace x_{i+1} with an utterance from a set of generic responses, 2) replace an answer in a QA pair with a generic answer or 3) replace x_{i+1} with one of following: $(x_1, ..., x_i)$ which results in the repetition of a speaker contribution. (See the table 2 for example.)

5 Experiment Results and Analysis

We use three different evaluation metrics *DialogRPT* (Gao et al., 2020), *Pang-Evaluation* (Pang et al., 2020) and *User Satisfaction* (Sun et al., 2021b) to test our adversarial test-suite. The first two metrics we test are for open-domain conversations while the last one is trained to judge task-oriented conversations. We focus on both open-domain dialogue and task-oriented dialogue metrics to highlight the diversity of our test-suite and show how neural metrics in both of these settings contain fundamental problems. Both DialogRPT and User Satisfaction metrics use human feedback

Metrics	Datasets	Entailment	Pronoun	NER	Co-reference	Speaker Sensitive-ness	Contradic-tion	Dullness
DialogRPT	DDial	0.3	0.48	0.6	0.56	0.98	0.78	0.48
	Reddit	0.13	0.66	0.64	0.32	0.91	0.72	0.53
	PerCh	0.4	0.55	0.74	0.45	0.89	0.87	0.47
Pang-C	DDial	0.46	0.08	0.49	0.03	0.01	0.26	0.78
	Reddit	0.5	0.49	0.41	0.58	0.04	0.47	0.9
	PerCh	0.47	0.2	0.53	0.74	0.00	0.51	0.86
Pang-L	DDial	0.45	0.75	0.72	0.23	0.74	0.32	0.67
	Reddit	0.49	0.63	0.57	0.35	0.59	0.29	0.81
	PerCh	0.54	0.76	0.73	0.35	0.69	0.33	0.81
UWBERT		0.98	1.00	1.00	0.01	1.00	0.97	0.98
WHiGRU	MultiWOz	0.93	0.98	0.94	0.07	0.96	0.96	0.95
WBERT		0.98	0.97	0.97	0.01	0.96	0.98	0.94

Table 3: Error rates of different evaluation metrics on different datasets for coherence and interestingness attacks.

Metrics	Datasets	Unnatural Repetitions	Vocabulary Diversity	Unnatural Paraphrases	Natural Paraphrases	Entrainment
DialogRPT	DDial	0.86	0.14	0.42	0.44	0.62
	Reddit	0.90	0.08	0.42	0.35	0.59
	PerCh	0.91	0.08	0.53	0.41	0.67
Pang-C	DDial	0.02	0.00	0.00	0.87	0.46
	Reddit	0.05	0.02	0.15	0.33	0.59
	PerCh	0.01	0.06	0.00	0.89	0.23
Pang-F	DDial	0.99	0.88	1.00	0.97	1.00
	Reddit	0.80	0.64	0.88	0.44	0.93
	PerCh	0.87	0.83	0.98	0.87	0.94
UWBERT		1	0.01	0.98	0.04	1
WHiGRU	MultiWOz	0.96	0.02	0.98	0.1	0.95
WBERT		0.99	0	0.98	0.03	1

Table 4: Error rates of different evaluation metrics on different datasets for naturalness attacks.

during the training while components of Pang-Evaluation rely on pre-trained neural backbones. We postulate that this variability in training methods might help provide specific insights for different methods.

From our analysis of metric behavior on adversarial conversations, we have the following key takeaways: *1)* neural metrics are not suited to judge overall quality of the conversations; *2)* they are unable to correctly understand conversation semantics; and 3) they are prone to data and training-induced biases.

5.1 Noise in Adversarial Heuristics

Proposed heuristics are expected to fail in some cases while generating adversarial conversations. To estimate this failure, we manually analyze the 250 generated conversations across all attacks and find that heuristics work successfully *84%* of the time. This signifies that heuristics are succeeding most of the time but requires further investigation for a more detailed estimate of their noise.

5.2 Analysis Methodology

An ideal metric would grade adversarial conversations worse than the original conversations in most cases, as most of the attacks are focused on generating bad behaviors. However, natural paraphrasing and attacks on reference use are expected to be graded similarly to the original conversations as they generate similar conversations as the original ones. Similarly, conversations generated by vocabulary diversity and entertainment attacks are expected to be scored *at most* as well as the original conversations as both of these manipulate vocabulary choices and sometimes these manipulations result in natural outcomes. To capture this variability, we define *error rate* as the proportion of times a conversation metric does not conform to the expected behavior.

We sample 100 adversarial conversations for each attribute per dataset and analyze the metric performance on those. We compute the proportion of times the metrics rate original conversations higher, equal and lower than the adversarial ones and use these statistics to compute the error rates. To make sure, we don't categorize insignificant

changes as greater or lesser we compute a minimum score threshold by computing the minimum change from the mean of metrics scores assigned to human conversations required to get a p-value less than *0.05* ($p < 0.05$) in a t-test. We use pre-trained NER model in the spacy python package (Honnibal and Montani, 2017) for named-entity detection and reference resolution models in (Clark and Manning, 2016a,b) to detect co-reference clusters.

5.3 DialogRPT

This is an evaluation metric which was trained on Reddit threads using human feedback in the form of up/down votes, number of replies, and the depth of a conversation thread. It also has a component focused on separating human from random utterances. DialogRPT requires significant improvements when dealing with the most adversarial cases as shown by error rates calculated in tables 3 and 4. It performs the best on the entailment task on Reddit data because it was trained to pick relative human responses from random ones. Additionally, it performs better on Reddit data on average than other datasets which provides evidence for better performance on in-distribution data.

It is evident that DialogRPT favors repetitions when we analyze adversarial cases which are generated using repetitions e.g. speaker sensitive cases in table-3 and unnatural repetitions cases in table-4. DialogRPT has a 93% error rate out of which 17% of the adversarial conversations are rated similar to the original for the speaker sensitivity. This proportion of similar ratings increases to 57% in case of unnatural repetitions. The fact that conversations in speaker sensitive case are rated higher more often provides evidence for a bias towards rating conversations with similar responses to context.

The performance of DialogRPT on attacks other than entailment is not good with error rates greater than 40%, which provides evidence that DialogRPT does not have a correct understanding of different dialogue aspects especially contradictions and individual speaker contributions.

5.4 Pang-Evaluation Metric

Pang-Evaluation presents four evaluation metrics which try to evaluate specific properties of dialogues. 3 of these depend on a pre-trained neural backbone: i) context-coherence (Pang-C) ii) fluency (Pang-F) iii) logical consistency (Pang-L). Pang-C and Pang-F rely on a GPT2 (Radford et al., 2018) backbone fine-tuned on Daily Dialogue dataset while the Pang-L relies on a Roberta model (Liu et al., 2019) fine-tuned on MNLI task (Williams et al., 2018) to detect contradictions.

We test Pang-C on all of the adversarial cases while we test the Pang-L metric on coherence and interestingness attacks and Pang-F metric on the naturalness because Pang-C is compared with other overall evaluation metrics in the paper while Pang-F and Pang-L are presented to judge fluency and logical consistency of a dialogue response. Pang-C seems to be robust to attacks which induce unnatural sentence structure. This is evident by looking at the results for 1) pronoun, speaker sensitiveness and co-reference attacks in the table 3; and 2) unnatural repetitions, paraphrasing attacks and vocabulary diversity attacks in the table 4. This highlights again that metrics perform best on the tasks they are designed for. Since GPT2 predicts the likelihood of next token given history, the metric is sensitive to unnatural manipulations. However, its performance varies across other attacks e.g. the metric fails to reliably penalize contradictions or entailments. Similar to DialogRPT, Pang-C also performs worse on out-of-distribution data as visible from the performance drop on datasets other than Dialy Dialogue. GPT2 could be fine-tuned on datasets to retain performance, but not every dataset has a large number of dialogues and fine-tuning large neural models is not a trivial task. This highlights that pre-trained models like GPT2 may retain performance for attacks which generate unnatural examples but still show a significant room for improvement when dealing with complex conversation semantics or out-of-distribution data.

Pang-C metric performs the best out of the three metrics presented by the Pang-Evaluation. Their Pang-F metric, results shown in the table 4, differs from Pang-C metric because it does not take context into account while assigning scores to the utterance under consideration. However, this causes it to under-perform in comparison to Pang-C metric while judging the adversarial cases. This also highlights a weakness in the training methodology itself. Since the metric has only seen full conversations during training, it fails to reliably penalize adversarial vocabulary choices when it evaluates individual utterances.

Pang-L metric is proposed to score the logical consistency of the conversations. The metric performs better on average than Pang-C at detecting contradictions (the task it was trained for), compa-

rable while detecting broken entailments but worse in other cases as visible from the table 3. Its performance is also consistent across datasets most-likely because it was trained on the MNLI task and not on any of the conversation datasets it was being evaluated on. This again highlights the limited applicability of the metric in-line with the other metric we have examined.

5.5 User Satisfaction Metric

User satisfaction simulation metric is trained using human-feedback ratings to predict a score from 1-5, given a conversation. The authors pose this task as a classification problem and train a classifier to predict a score. One of the major problems with this metric, lies in the data it was trained on. Since the data they use has an highly skewed rating distribution, the metric performance reflects this. We test the classifier version of the metric using fine-tuned BERT model (UWBERT). This version assigned the same score to almost all of the adversarial and original conversations. We also computed the score as the weighted average of ratings, using both their hierarchical GRU (WHiGRU) and BERT (WBERT) models, to test if minor variation in assigned scores could provide some insights but the results remained the same. As shown in the tables 3 and 4, this bias in the data results in high error rates closer to 1.0. When error rate relies on scores being similar e.g. in case of co-reference attack, it drops to 0.0 because the model assessment never changes. This highlights a case of bias in model assessments because of the problems in training data.

5.6 Performance Highlights and Conclusion

All metrics that we test perform better on the data they were trained on with the best performance on the trained down-stream task(s). However, none of them are suitable to judge the overall quality of conversations. To be more specific the results show that variability in training methods can cause metrics to assess conversation behaviors differently e.g. DialogRPT vs Pang-C metric performance, metrics are dependent on the training data and reflect the biases in the data and training methodologies e.g. performance of User Satisfaction metric. These insights help us draw two conclusions: 1) a series of metrics specific to individual dialogue behaviors might judge overall quality better than a single evaluation metric and 2) failure of metrics to reliably judge complex behaviors like contradictions

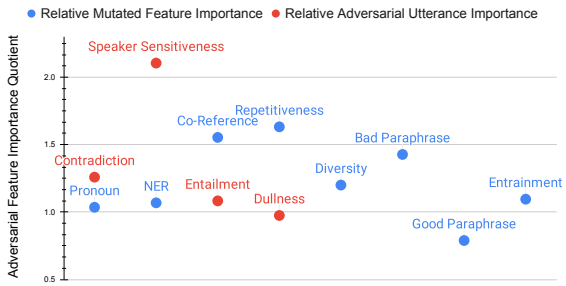


Figure 1: The graph shows relative importance of adversarial conversation features. All adversarial cases having relative importance $>= 1.0$ shows that DialogRPT is paying more importance to adversarial features.

indicate that either metrics should be trained using examples of these behaviors or should be exposed to more information about conversation flow than just surface textual features.

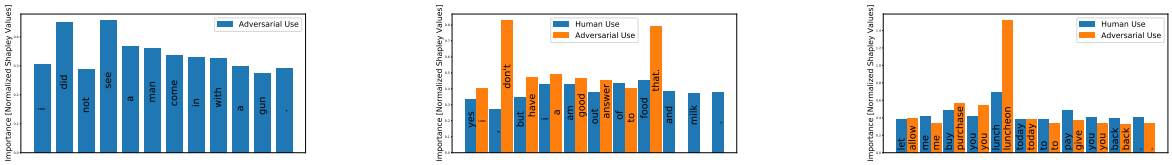
6 Interpretability Analysis

6.1 Method

In addition to the behavioral analysis described above, we use a black-box model interpretability technique to approximate the feature importance for different adversarial cases for the DialogRPT metric. This shows the relative importance of adversarial features as the metrics paying more importance to adversarial features to assign higher rating to adversarial conversations would show further evidence for misunderstanding of different dialogue semantics.

More specifically, we use Kernel SHAP algorithm presented by (Lundberg and Lee, 2017) which is a modification of LIME (Ribeiro et al., 2016). To determine the Shapley values, the algorithm masks the features (utterance tokens in this case), and replaces them with features from a pre-specified set. The dialogue metric scores are then measured using the augmented instances. These scores along with the score assigned to the original utterance are used to approximate a linear model over the utterance features. The weights of this linear model are the approximate Shapley values.

Instead of replacing the masked tokens with the vocabulary from other utterances we replace them with the masking token specific to the model under consideration. This helps measure the effect on the metric when a certain feature is missing. This was inspired by the *Partition SHAP* algorithm in SHAP python package released by the authors and erasure of feature representations in Li et al. (2017a).



(a) Contradiction Example (b) Dullness Example (c) Unnatural Paraphrases

Figure 2: Shapley values of adversarial features for each case mentioned above. Each adversarial example shows that DialogRPT metric pays high importance to the adversarial features.

6.2 Overall Analysis

We conduct an interpretability analysis of DialogRPT metric performance on adversarial conversations from the Daily Dialogue dataset. We compare the aggregate Shapley values of the human and adversarial features to judge how the DialogRPT metric pays importance to different tokens in the adversarial and human utterances. The aggregation of Shapley values is done using an addition operation. We normalize the Shapley values of features using the minimum and maximum Shapley values for all the features. This helps in conducting a fair comparison of contribution each feature makes to the result. We compare the aggregate Shapley values : 1) between features that are different in human and adversarial responses for cases generated by mutating some of the human features 2) of all features for cases in which the whole human utterance is replaced with an adversarial response. The average relative feature importance of adversarial features (*average adversarial feature importance divided by original feature importance*) is presented in the figure 1. The relative importance of $>= 1.0$ for almost all the adversarial cases shows that DialogRPT considers adversarial features more important than the original ones and points towards a misunderstanding of the role that language features play in the conversation flow.

6.3 Feature Importance Analysis

We present an analysis of feature importance for three adversarial cases below. These use the same conversations presented earlier in the table 2.

6.3.1 Contradictions

Figure 2(a) shows the importance assigned by the DialogRPT to the adversarial features in the contradiction example, shown in the table 2. Tokens such as *"did not see"* directly contradict the context and are given relatively higher importance in comparison to the other features in the utterance. This proves that DialogRPT fails to understand the role of verb negations in contradicting the context.

6.3.2 Dullness

Figure 2(b) shows the importance of human and adversarial features side by side for the dullness example in the table 2. DialogRPT metric pays more importance to the features which do not progress the conversation in comparison to the human response which contributes meaningfully to the conversation by answering the question being asked. This shows that DialogRPT does not understand which features progress the conversation meaningfully.

6.3.3 Unnatural Paraphrasing

As shown in the figure 2(c), for the unnatural paraphrasing example in the table 2, it is clear that the DialogRPT metric pays more or relatively same importance to the features which are either not used in the same sense, *give* instead of *pay*, or not used as frequently in the human conversations, like the use of *luncheon* instead of *lunch*. This shows that DialogRPT does not prefer human-like vocabulary choices.

6.4 Discussion

Our analysis on the DialogRPT suggests that it does not know the correct semantics of conversations. We hypothesize it could be because it is not exposed to bad dialogue behaviors and that leads to the metric paying high attention to erroneous constructions. To correct these, it would be a good idea to either augment the training data with adversarial dialogue behaviors, use adversarial learning to make the metrics more robust or use semantic features in addition to the surface language features during training.

7 Choice of Evaluation Metrics

The metrics evaluated in this work are some of the recently proposed with varied training methodolo-

gies e.g., training using human judgments versus some measure of text similarity. Also, our choice of these metrics is rooted in the fact that these metrics are primarily judged by a singular measure of improvement which in most cases is the correlation with human judgments. Our analysis highlights that such singular measures of improvement are not enough to capture the variability of performance exhibited by the evaluation metrics in judging complex conversation behaviors like contradictions. Our goal is not to provide an exhaustive accounting of the performance of all available neural metrics but empirically highlight the problems and performance variability which arise because of different training methodologies. Given our results, we hypothesize that other metrics like (Mehri and Eskenazi, 2020b; Lowe et al., 2017; Tao et al., 2018) trained similarly have deeper problems which need to be highlighted using a fine-grained methodology similar to ours.

8 Conclusion and Future Work

Our test-suite helps us draw various insights about performance of different metrics, and shows that all the metrics we test have room for improvement. It points at several flaws in the metrics: 1) lack of generalization ability to unseen data, 2) inability to correctly understand different conversation semantics and 3) prone to training and data-induced biases. Furthermore, our interpretability analysis further corroborates these shortcomings and helps us conclude that metrics need to be exposed to more information about conversation behaviors to make them more robust.

Adversarial behaviors from our test-suite help point to many shortcomings in the metrics we test, but many behaviors are very simplistic *e.g.* speaker sensitivity attacks. Further research can help make these attacks more human-like which may help reveal more information about the evaluation metrics. Our test-suite can also be used directly to make the metrics more robust *e.g.* by augmenting their training data with adversarial examples or by using it as a reward signal in a RL setup for the training of evaluation metrics. This shows several use-cases of our test-suite and directions for future research.

References

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021a. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021b. Is incoherence surprising? targeted evaluation of coherence prediction from language models. *CoRR*, abs/2105.03495.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

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.

Kevin Clark and Christopher D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, Texas. Association for Computational Linguistics.

Kevin Clark and Christopher D. Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany. Association for Computational Linguistics.

Heriberto Cuayáhuitl, Steve Renals, Oliver Lemon, and Hiroshi Shimodaira. 2010. Evaluation of a hierarchical reinforcement learning spoken dialogue system. *Computer Speech & Language*, 24(2):395–429.

W. Eckert, E. Levin, and R. Pieraccini. 1997. User modeling for spoken dialogue system evaluation. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 80–87.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational AI. *CoRR*, abs/1809.08267.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Youngsoo Jang, Jongmin Lee, and Kee-Eung Kim. 2020. Bayes-adaptive monte-carlo planning and learning for goal-oriented dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7994–8001.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems.

Sangkeun Jung, Cheongjae Lee, Kyungduk Kim, Minwoo Jeong, and Gary Geunbae Lee. 2009. Data-driven user simulation for automated evaluation of spoken dialog systems. *Computer Speech & Language*, 23(4):479–509.

Mohammad Kachuee, Jinseok Nam, Sarthak Ahuja, Jin-Myung Won, and Sungjin Lee. 2022. Scalable and safe self-learning for skill routing in large-scale conversational ai systems. In *NAACL*.

Mohammad Kachuee, Hao Yuan, Young-Bum Kim, and Sungjin Lee. 2021. Self-supervised contrastive learning for efficient user satisfaction prediction in conversational agents. In *NAACL*.

Baber Khalid, Malihe Alikhani, Michael Fellner, Brian McMahan, and Matthew Stone. 2020a. Discourse coherence, reference grounding and goal oriented dialogue.

Baber Khalid, Malihe Alikhani, and Matthew Stone. 2020b. Combining cognitive modeling and reinforcement learning for clarification in dialogue. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4417–4428, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. Sharp nearby, fuzzy far away: How neural language models use context. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.

Florian Kreyssig, Iñigo Casanueva, Paweł Budzianowski, and Milica Gašić. 2018. Neural user simulation for corpus-based policy optimisation of spoken dialogue systems. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 60–69, Melbourne, Australia. Association for Computational Linguistics.

Alice Lai and Joel Tetreault. 2018. Discourse coherence in the wild: A dataset, evaluation and methods. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 214–223, Melbourne, Australia. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017a. Understanding neural networks through representation erasure.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

Raveesh Meena, Gabriel Skantze, and Joakim Gustafson. 2014. Data-driven models for timing feedback responses in a map task dialogue system. *Computer Speech & Language*, 28(4):903–922.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*,

pages 225–235, 1st virtual meeting. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Laura Panfili, Steve Duman, Andrew Nave, Katherine Phelps Ridgeway, Nathan Eversole, and Ruhi Sarikaya. 2021. Human-ai interactions through A gricean lens. *CoRR*, abs/2106.09140.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, and Kewei Tu. 2020. Towards holistic and automatic evaluation of open-domain dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3619–3629, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sunghyun Park, Han Li, Ameen Patel, Sidharth Mudgal, Sungjin Lee, Young-Bum Kim, Spyros Matsoukas, and Ruhi Sarikaya. 2021. A scalable framework for learning from implicit user feedback to improve natural language understanding in large-scale conversational AI systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6054–6063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leila Pishdad, Federico Fancellu, Ran Zhang, and Afsaneh Fazly. 2020. How coherent are neural models of coherence? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6126–6138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Ananya B. Sai, Mithun Das Gupta, Mitesh M. Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses.

Lloyd S. Shapley. 1952. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA.

Heereen Shim, Dietwig Lowet, Stijn Luca, and Bart Vanrumste. 2021. Building blocks of a task-oriented dialogue system in the healthcare domain. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 47–57, Online. Association for Computational Linguistics.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021a. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21. ACM.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021b. *Simulating User Satisfaction for the Evaluation of Task-Oriented Dialogue Systems*, page 2499–2506. Association for Computing Machinery, New York, NY, USA.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 722–729. AAAI Press.

E. trumbelj and I. Kononenko. 2013. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665.

Stefan Ultes, Alexander Schmitt, and Wolfgang Minker. 2013. On quality ratings for spoken dialogue systems – experts vs. users. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 569–578, Atlanta, Georgia. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

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, New Orleans, Louisiana. Association for Computational Linguistics.

Jie Yang, Hirofumi Kikuchi, Takatsugu Uegaki, and Hideaki Kikuchi. 2021. The effect of the repetitive utterances complexity on user's perceived empathy and desire to continue dialogue by a chat-oriented dialogue system. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, HAI '21, page 241–244, New York, NY, USA. Association for Computing Machinery.

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)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Adversarial Examples

We present one adversarial example for all of the dialogue attributes (not present in the table 2) in the tables 5 and 6. These examples are generated in an automatic way using our test-suite and testify to its effectiveness in generating different adversarial behaviors.

Coherence		
Entailment	Pronoun	Named Entity Attacks
A: It's a Taiwanese puppet doll. **B**: It's huge! **A**: Yeah. They're usually this big. **Actual**: The craftsmanship is excellent. --- **Adversarial**: I'm Rose Teller. I think I've seen you somewhere before?	**A**: Harry, come here immediately! **B**: What? **Actual**: Don't take that tone with me! I saw you hit your brother. --- **Adversarial**: Don't take that tone with you! we saw me hit your brother.	**A**: Can you tell me what bus to catch from Altadena to downtown LA? **B**: You can catch the 486. **Actual**: That bus goes all the way to LA? --- **Adversarial**: That bus goes all the way to Chilin?

Coherence		Naturalness
Co-reference Attacks	Speaker Sensitiveness	Unnatural Repetitions
A: i am only five feet and five inches, so i am short too. are you married? **B**: i am not. i just have my dog pedro. he is my family **A**: i do not like dogs. i was attacked when i was a little girl. **Actual**: i am so sorry to hear that . i bet you would like pedro he is sweet --- **Adversarial**: i am so sorry to hear that. i bet you would like pedro pedro is sweet	**A**: i hate families. i prefer to be alone. **B**: i am short at 5 ft tall. how about you? **A**: i am 5ft and 6in tall.i weigh 220 pounds and its all muscle. **Actual**: that sounds very nice, yes. --- **Adversarial**: i am 5ft and 6in tall. i weigh 220 pounds and its all muscle. that sounds very nice, yes.	**A**: i just ate a mango and now i need to go to the hospital. **Actual**: are you allergic? dogs give me bad allergies. --- **Adversarial**: are you allergic? dogs give me bad allergies bad allergies bad allergies.

Table 5: Examples for the manipulation of coherence based attacks.

Naturalness		
Entrainment	Vocabulary Diversity	Natural Paraphrasing
A: what music do you like? **B**: i dont really prefer any kind of music **Actual**: well, does your mom play any music at her restaurant? --- **Adversarial**: well, does your mom play any euphony at her restaurant?	**A**: i for sure read an speak english **B**: that is helpful. i do as well and just graduated college. **A**: i love pork, especially bacon. **Actual**: bacon is good. i do not eat much meat though. do you have pets? --- **Adversarial**: bacon is well. i do not eat much meat though. do you have pets?	**A**: my hobbies are fashion an clothes! **B**: fashion is cool. i am an avid gamer playing second life. **Actual**: awesome! what is second life? never heard of it --- **Adversarial**: What is Second Life? Never heard of it!

Table 6: Examples of different adversarial responses for naturalness cases generated by our heuristics.