I like fish 🐟, *especially dolphins* 🐬:* Addressing Contradictions in Dialogue Modeling

Yixin Nie[1], **Mary Williamson**[2], **Mohit Bansal**[1], **Douwe Kiela**[2], **Jason Weston**[2]
[1]UNC Chapel Hill
[2]Facebook AI Research

Abstract

To quantify how well natural language understanding models can capture consistency in a general conversation, we introduce the DialoguE COntradiction DEtection task (DECODE) and a new conversational dataset containing both human-human and human-bot contradictory dialogues. We show that: (i) our newly collected dataset is notably more effective at providing supervision for the dialogue contradiction detection task than existing NLI data including those aimed to cover the dialogue domain; (ii) Transformer models that explicitly hinge on utterance structures for dialogue contradiction detection are more robust and generalize well on both analysis and out-of-distribution dialogues than standard (unstructured) Transformers. We also show that our best contradiction detection model correlates well with human judgments and further provide evidence for its usage in both automatically evaluating and improving the consistency of state-of-the-art generative chatbots.

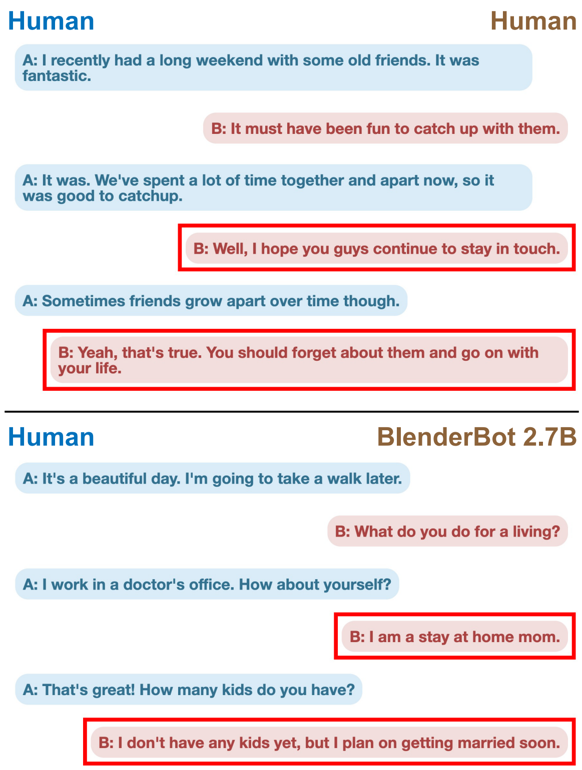

Figure 1: Contradictory dialogues contained in our new DECODE dataset. The main train, valid and test sets contain human-written dialogues with deliberate contradictions (example at top), and an out-of-domain test set consists of labeled human-bot dialogues (bottom), involving state-of-the-art bots (Roller et al., 2020).

1 Introduction

Recent progress on neural approaches to natural language processing (Devlin et al., 2019; Brown et al., 2020), and the availability of large amounts of conversational data (Lowe et al., 2015; Smith et al., 2020) have triggered a resurgent interest on building intelligent open-domain chatbots. Newly developed end-to-end neural bots (Zhang et al., 2020; Adiwardana et al., 2020; Roller et al., 2020) are claimed to be superior to their predecessors (Worsnick, 2018; Zhou et al., 2020) using various human evaluation techniques (See et al., 2019; Li et al., 2019; Adiwardana et al., 2020) that aim to give a more accurate measure of what makes a good conversation. While the success is indisputable, there is still a long way to go before we arrive at human-like open-domain chatbots. For example, it has been shown that open-domain chatbots frequently generate annoying errors (Adiwardana et al., 2020; Roller et al., 2020) and a notorious one among these is the class of contradiction, or consistency errors.

When interacting with chatbots, people carry over many of the same expectations as when interacting with humans (Nass and Moon, 2000). Self-contradictions by these bots (see Fig.1, bottom) are often jarring, immediately disrupt the conver-

* Dolphins are mammals, not fish.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1699–1713
August 1–6, 2021. ©2021 Association for Computational Linguistics

sational flow, and help support arguments about whether generative models could ever really understand what they are saying at all (Marcus, 2018). From a listener's perspective, such inconsistent bots fail to gain user trust and their long-term communication confidence. From a speaker's perspective, it violates the maxim of quality in Grice's cooperative principles (Grice, 1975) —"Do not say what you believe to be false." Hence, efforts on reducing contradicting or inconsistent conversations by open-domain chatbots are imperative.

Prior works (Welleck et al., 2019) characterized the modeling of persona-related consistency as a natural language inference (NLI) problem (Dagan et al., 2005; Bowman et al., 2015), and constructed a dialog NLI dataset based on Persona-Chat (Zhang et al., 2018), but so far state-of-the-art chatbots (Roller et al., 2020) have not been able to make use of NLI techniques in improving dialogue consistency. Overall, the challenge remains that we are still unable to answer the simple yet important question—"*how good are we at modeling consistency (including persona, logic, causality, etc.) in a general conversation?*". The inability to measure this obscures to what degree building new modules or techniques can in turn help prevent contradicting responses during generation.

Seeking to answer this question, we introduce the DialoguE COntradiction DEtection task (DECODE)[1] and collect a new conversational dataset containing human written dialogues where one of the speakers deliberately contradicts what they have previously said at a certain point during the conversation. We also collect an out-of-distribution (OOD) set of dialogues in human-bot interactive settings which contain human-labeled self-contradictions made by different chatbots.

We then compare a set of state-of-the-art systems, including a standard unstructured approach and a proposed structured approach for utilizing NLI models to detect contradictions. In the unstructured approach, a Transformer NLI model directly takes in the concatenation of all utterances of the input dialogue for prediction, following the paradigm of NLU modeling. In the structured approach, utterances are paired separately before being fed into Transformer NLI models, explicitly taking account of the natural dialogue structure.

Results reveal that: (1) our newly collected dataset is notably more effective at providing supervision for the contradiction detection task than existing NLI data including those aimed at covering the dialogue domain; (2) the structured utterance-based approach for dialogue consistency modeling is more robust in our analysis and more transferable to OOD human-bot conversation than the unstructured approach. This finding challenges the mainstream unstructured approach of simply applying pre-trained Transformer models and expecting them to learn the structure, especially for OOD scenarios which are often the case when incorporating NLU modules into NLG systems, since intermediate in-domain data are scarce.

Finally, with such improvements on the contradiction detection task, we show that our best resulting detector correlates well with human judgments and can be suitable for use as an automatic metric for checking dialogue consistency. We further provide evidence for its usage in improving the consistency of state-of-the-art generative chatbots.

2 Related Work

Several prior works on improving dialogue consistency have explored using direct modeling of the dialogue context in generation algorithms. The modeling can be implicit where the dialogue consistency-related information like style (Wang et al., 2017), topics, or personal facts are maintained in distributed embeddings (Li et al., 2016; Zhang et al., 2019a), neural long-term memories (Bang et al., 2015), hierarchical neural architecture (Serban et al., 2016), latent variables (Serban et al., 2017), topical attention (Dziri et al., 2019a), or even self-learned feature vectors (Zhang et al., 2019b). Some works have grounded generation models on explicit user input (Qian et al., 2018), or designated personas (Zhang et al., 2018). Although, improvements on automatic generation metrics were often shown on guided response generation based on the consistency modeling, the issue of contradiction has never been resolved, nor have generally applicable methods to gauge the consistency improvements been developed. Further, simply scaling models has not made the problem go away, as is evident in the largest chatbots trained such as BlenderBot with up to 9.4B parameter Transformers (Roller et al., 2020).

More similar to our work is utilizing NLI models in dialogue consistency. Dziri et al. (2019b) attempted to use entailment models trained on syn-

[1]DECODE dataset and code are publicly available at https://parl.ai/projects/contradiction.

thetic datasets for dialogue topic coherence evaluation. Particularly, Welleck et al. (2019) constructed the dialogue NLI dataset and (Li et al., 2020) utilized it to try to reduce inconsistency in generative models via unlikelihood training in a preliminary study that reports perplexity results, but did not measure actual generations or contradiction rates. We note that the dialogue NLI dataset is only semi-automatically generated, with limited coverage of only Persona-chat data (Zhang et al., 2018), whereas our DECODE is human-written and across multiple domains. Our task also involves logical and context-related reasoning beyond personal facts. We show that transfer of DECODE is subsequently more robust than dialogue NLI on both human-human and human-bot chats.

3 Task and Data

3.1 Dialogue Contradiction Detection

We formalize dialogue contradiction detection as a supervised classification task. The input of the task is a list of utterances $x = \{u_0, u_1, u_2, ..., u_n\}$ representing a dialogue or a dialogue snippet. The output is y, indicating whether the last utterance u_n contradicts any previously conversed information contained in the dialogue $\{u_0, u_1, ..., u_{n-1}\}$, where y can be 0 or 1 corresponding to the non-contradiction and the contradiction label respectively. Preferably, the output should also include a set of indices $\mathbf{I} \subseteq \{0, 1, ..., n-1\}$ representing a subset of $\{u_0, u_1, ..., u_{n-1}\}$ which contain information that is actually contradicted by the last utterance u_n. The extra indices $\mathbf{I}$ output require models to pinpoint the evidence for the contradiction, providing an extra layer of explainability.

3.2 Data Collection

Our goal is first to collect training and evaluation data for this task. We thus collect dialogues in which the last utterance contradicts some previous utterances in the dialogue history. To obtain such dialogues, we give annotators dialogue snippets from pre-selected dialogue corpora, and then ask them to continue the conversation by writing one or two utterances such that the last utterance by the last speaker contradicts the dialogue history. We also ask annotators to mark all the utterances in the dialogue history that are involved in the contradiction as supporting evidence. We ask annotators to write contradicting utterances based partly on existing dialogues rather than collecting new dialogue from scratch because the provided dialogues can often convey semantic-rich contexts from different domains and inspire annotators to write more diverse examples. We don't impose constraints on the annotation such that the annotator could have the flexibility to write more diverse contradictory responses that might not belong to pre-defined types (knowledge, emotion, persona, etc). Also note that we ask the annotator to write contradictory dialogues based on pre-selected human-human dialogue rather than collecting dialogues from human-bot interaction for the main dataset because we want the examples to be general and less bound to specific bots.[2] We crowdsource the continuation and annotation data with Amazon Mechanical Turk via ParlAI (Miller et al., 2017).

To ensure data quality, we apply three techniques: (i) an onboarding test every annotator has to pass to contribute examples; (ii) each annotator can only create up to 20 examples; and (iii) all examples in the validation and test set are verified by asking 3 additional workers. More details about annotation are provided in Appendix.

3.3 Dataset

We collected 17,713 human-written contradicting dialogues in which 4,121 are verified by 3 annotators. The pre-selected dialogue source corpora are Wizard of Wikipedia (Dinan et al., 2019), EMPATHETICDIALOGUES (Rashkin et al., 2019), Blended Skill Talk (Smith et al., 2020), and ConvAI2 (Dinan et al., 2020), covering various conversational topics. To facilitate the evaluation of consistency modeling on the dialogue contradiction detection classification task, we sample an equal number of non-contradicting dialogues according to the same dialogue length distribution as the contradicting ones from the same dialogue corpus. Then, we make the splits such that the train split contains unverified examples, and dev and test splits only contain verified examples. Each split has balanced labels between contradiction and non-contradiction. The breakdown of each of the dataset sources is shown in Appendix.

Auxiliary (Checklist) Test Sets. We further create two auxiliary checklist evaluation sets by transforming the contradiction examples in the original test in two ways such that the ground truth label is

[2]Alongside the main dataset, another portion of the examples are collected via human-bot interaction and used as out-of-domain evaluation.

	Count	Label
Main (Train)	27,184	balanced
Main (Dev)	4,026	balanced
Main (Test)	4,216	balanced
Human-Bot (Test)	764	balanced
A2T (Test)	2,079	contradiction
RCT (Test)	2,011	non-contradiction

Table 1: DECODE Dataset summary. The first column presents the different dataset types. "Main" is the collected human-written dialogues. "balanced" indicates that the contradiction and non-contradiction labels in that part of the dataset are balanced. A2T and RCT are the auxiliary test sets described in subsection 3.3.

either invariant or expected to flip. The two resultant sets serve as diagnostic tests on the behavior, generalization and transferability of our models.

The transformations are described below:

- **Add Two Turns (A2T)** We insert a pair of randomly sampled utterances into the dialogue such that the inserted utterances are between the two original contradicting utterances. This gives a new contradicting dialogue with a longer dialogue history.
- **Remove Contradicting Turns (RCT)** We remove all the turns (all pairs of utterances)[3] marked as supporting evidence for the contradiction in the dialogue except the last utterance. This results in a new non-contradiction dialogue.

Human-Bot Test Set. Our main dataset involves human-written dialogues containing contradicting utterances based on human-human dialogues from existing corpora. In practice, to evaluate the response quality of a machine rather than a human in terms of its consistent responses, we care about how well a contradiction detector can perform in human-bot interactive conversations. To that end, we further collect human-bot dialogue data by employing crowdworkers to interact with a diverse set of open-domain bots. These include Poly-encoder (Humeau et al., 2019) based retrieval models, generative models (Roller et al., 2020), unlikelihood trained models (Li et al., 2020), retrieve-and-refine models (Weston et al., 2018; Roller et al., 2020), models either pre-trained on a previously existing Reddit dataset extracted and obtained by a third party that was hosted by pushshift.io (Baumgartner et al., 2020) or fine-tuned on the Blended Skill Talk (BST) dialogue tasks (Smith et al., 2020) – that is, all the dialogue models that are compared in the study in Roller et al. (2020). During the collection, if the bot generates an utterance that contradicts itself, we ask the worker to mark the utterance. In some of the dialogues, workers are explicitly instructed to goad the bots into making contradicting utterances. The final human-bot test set we derive contains 764 dialogues, half of which end with a contradicting utterance by the bot. All the dialogues in the set, with either contradiction or non-contradiction labels, are verified by 3 additional annotators, beside the human who actually talked to the bot.

The auxiliary and human-bot test sets aim to test models' robustness and generalizability beyond the collected human-written test set (Ribeiro et al., 2020; Gardner et al., 2020), and give a more comprehensive analysis of the task. Table 1 summarizes the final overall dataset. Examples are provided for each dataset type in Fig. 1 and Appendix Table 5.

4 Models

To model the dialogue consistency task, we first employ some of the techniques used in NLI sequence-to-label modeling, where the input is a pair of textual sequences and the output is a label. The benefit of such modeling is that we can directly make use of existing NLI datasets during training. However, unlike previous work (Welleck et al., 2019) that directly utilized NLI models giving a 3-way output among "entailment", "contradiction", and "neutral", we modify the model with a 2-way output between "contradiction" and "non-contradiction" (either "entailment" or "neutral") labels, as our task is centered around the detection of inconsistency.

More formally, we denote the model as $\hat{y}_{pred} = f_\theta(\mathbf{C}, u)$, where $\hat{y}_{pred}$ is the prediction of the label y, i.e. whether the textual response u contradicts some textual context $\mathbf{C} = \{u_0, u_1, ..., u_{n-1}\}$, and θ are the parameters of the model.

4.1 Dialogue Contradiction Detectors

We explore two distinct approaches that propose differing f_θ for the detection prediction problem.

Unstructured Approach. In this approach, we simply concatenate all the previous utterances in the dialogue history to form a single textual context. Then, we apply f_θ to the context and the last

[3]The dataset dialogues involve two speakers taking turns speaking. To maintain this structure, for each marked utterance, we remove a pair of utterances that represents a turn. This also helps remove information involved in the contradiction such that the new label should be "non-contradiction".

utterance to infer the probability of contradiction:

$$\hat{y}_{pred} = f_\theta([u_0, u_1, u_2, ..., u_{n-1}], u_n) \quad (1)$$

When concatenating the utterances, we insert special tokens before each utterance to indicate the speaker of that utterance. This is aimed to provide a signal of the dialogue structure to the models. Still, this approach assumes that the model can use these features adequately to learn the underlying structure of the dialogue implicitly during training.

Structured Utterance-based Approach. Since the reasoning crucially depends on the last utterance, in this method we first choose all the utterances by the last speaker to form a set $\mathbf{S}$. We then pair every utterance in the set with the last utterance and feed them one by one into f_θ^{UB}. The final contradiction probability is the maximum over all the outputs:

$$\hat{y}_{pred} = \max\left\{f_\theta^{UB}(u_i, u_n) : u_i \in \mathbf{S}\right\} \quad (2)$$

Additionally, the utterance-based approach is able to give a set of utterances as supporting evidence for a contradiction decision by choosing the pairs having contradiction probability higher than a threshold η_e:

$$\mathbf{I} = \left\{i : f_\theta^{UB}(u_i, u_n) > \eta_e\right\} \quad (3)$$

This not only gives explanations for its prediction but can also help diagnose the model itself, e.g. we can measure metrics of the model's ability to provide these explanations by comparing them against gold supporting evidence annotations.

One downside of this modeling approach is that it will not be able to capture reasoning between speakers. A case for that would be a pronoun by one speaker might refer to something initiated by the other speaker. Nevertheless, the utterance-based approach explicitly adds an inductive structure bias to learning and inference which we will see can aid its generalization capability.

Thresholding. For both the unstructured and utterance-based approaches, the detection of contradiction is made by comparing $\hat{y}_{pred}$ with a threshold τ and by default τ is 0.5.

4.2 Experimental Setup

We study four base pre-trained models variants for f_θ: BERT (Devlin et al., 2019), Electra (Clark et al., 2020), RoBERTa (Liu et al., 2019), and BART (Lewis et al., 2020). They represent the start-of-the-art language representation models and have yielded successes in many NLU tasks. The input format of f_θ follows how these models handle sequence-pairs ($\mathbf{C}$ and u) for classification tasks with padding, separator and other special tokens such as position embeddings and segment features inserted at designated locations accordingly.

We fine-tune f_θ on different combinations of NLI training data including SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), ANLI-R3 (Nie et al., 2020a)[4], DNLI (Welleck et al., 2019), as well as our DECODE Main training set. We convert the 3-way labels of the examples in existing NLI datasets to 2-way, as described before, and θ is optimized using cross-entropy loss. When training f_θ^{UB} in the utterance-based approach using the DECODE training set, the input sequences

Model	Training Data	MT	MT (Strict)	HB	SE F1
Unstructured Approach					
RoBERTa	All	97.46	-	77.09	-
	All - DNLI	97.44	-	73.17	-
	All - ANLI-R3	**98.04**	-	73.56	-
	All - DECODE	84.42	-	61.91	-
	DNLI	57.19	-	60.34	-
	ANLI-R3	82.21	-	59.69	-
	DECODE	96.85	-	70.03	-
Utterance-based Approach					
RoBERTa	SNLI + MNLI	77.40	47.70	73.17	72.4
	All	94.19	80.08	83.64	**88.5**
	All - DNLI	94.38	**80.93**	81.68	88.4
	All - ANLI-R3	94.07	79.32	82.85	88.4
	All - DECODE	86.67	66.95	77.36	80.6
	DNLI	76.54	63.09	75.26	71.2
	ANLI-R3	81.59	69.11	70.52	74.3
	DECODE	93.19	80.86	**84.69**	87.5
BERT	DECODE	88.88	74.14	75.52	84.3
Electra	DECODE	93.17	81.19	80.76	87.5
BART	DECODE	94.47	80.10	79.19	88.2
Majority					
-	-	50.00	50.00	50.00	48.7

Table 2: Test performance on DECODE for various methods. "MT" and "HB" columns show model accuracy on the Main Human-Human Test set and the Human-Bot set, respectively. The "MT (Strict)" column indicates the percentage when both the 2-way contradiction detection and the supporting evidence retrieval exactly match with the ground truth. "SE F1" is F1 score for supporting evidence retrieval. "All" in the "Training Data" column stands for a combination of SNLI, MNLI, DNLI, ANLI-R3, DECODE. "All - DNLI" denotes all the datasets with DNLI removed.

[4]ANLI data is collected in three rounds resulting in three subsets (R1, R2, R3). We only used training data in R3 since it contains some dialogue-related examples.

are sampled utterance pairs from the DECODE dialogue. In other scenarios, f_θ or f_θ^{UB} are trained with data treated as in normal NLI training.

The models are evaluated on the test sets described in subsection 3.3. For the utterance-based approach, which provides supporting evidence utterances (Equation 3), we report F1 on evidence retrieval. We also report a stricter score which evaluates whether both 2-way contradiction detection and supporting evidence retrieval *exactly match* with the ground truth on DECODE Main test.

5 Results and Analysis

5.1 Performance on Constructed Dataset

Our main results comparing various detectors on DECODE are shown in Table 2. We now describe our key observations.

DECODE is notably more effective than other existing NLI data in providing supervision for contradiction detection in dialogue. We found that models trained on DECODE achieve higher accuracy than that of those trained on DNLI or ANLI-R3, on all evaluation sets, with large improvements, e.g. a 12-point jump from the same model training on ANLI-R3 and a 16-point jump from training on DNLI using utterance-based RoBERTa on the DECODE Main test set. Moreover, while training on "All" datasets (SNLI, MNLI, ANLI-R3, DNLI & DECODE) is effective, the removal of DECODE from the training data induces a consequential downgrade on the performance. Training on NLI data which does not cover the dialogue domain, e.g., SNLI+MNLI is even worse, only achieving 77.4% on DECODE Main (Test) vs. 93.19% for DECODE and cannot even reach the majority baseline on the "Main (Test-Strict)". Further, training on DECODE is also more helpful than DNLI or ANLI-R3 for supporting evidence retrieval. These findings indicate that existing NLI data has limited transferability to the dialogue contradiction detection task despite their coverage of the dialogue domain in addition to other domains and that our DECODE data provides a valuable resource for modeling dialogue consistency and developing data-driven approaches for contradiction detection.

Different pre-training models that perform similarly on the in-domain test set can have very different performance on OOD human-bot dialogue. The last four rows of the table show the results of utterance-based RoBERTa, BERT, Electra, and BART trained on DECODE. We can see that RoBERTa, Electra, and BART got similar in-domain accuracy on DECODE, around 93%-94%. RoBERTa stands out when comparing their performance on the human-bot test set with the highest score of 84.69% across the column and with better performance on supporting evidence retrieval as well. We speculate that this is due to the fact that RoBERTa pre-training data has a broader coverage than Electra and BART. We hope future work on dialogue contradiction detection could explore pre-training models on more dialogue-focused corpora.

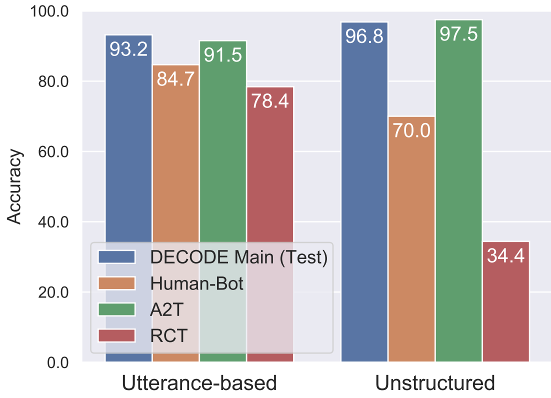

Figure 2: Comparison between utterance-based and unstructured approaches of RoBERTa pre-trained, DECODE fine-tuned models on DECODE Main (Test), Human-bot, and auxiliary test sets.

The unstructured approach gets higher accuracy on the in-domain test set. A direct comparison between unstructured RoBERTa and utterance-based RoBERTa trained on DECODE reveals that the unstructured approach more often than not gets a higher accuracy than its corresponding utterance-based approach when other experimental setups are kept identical. Noticeably, unstructured RoBERTa trained on all NLI data got a 97.46% score, whereas utterance-based yielded 94.19%. This seemingly indicates that training an unstructured model is able to yield a good representation of the consistency of the dialogue. However, analysis on the human-bot and auxiliary test sets shows that such high accuracy is an over-amplification of the model's real understanding ability, as we discuss next.

The structured utterance-based approach is more robust, and more transferable. Figure 2 gives a comparison between utterance-based and unstructured RoBERTa on each of the evaluation sets. We can see that the utterance-based model is able to maintain satisfactory performance across all the sets whereas the unstructured model underperforms at the human-bot and RCT auxiliary test sets with a 34.4% accuracy on RCT compared to 78.4% for utterance-based, in stark contrast to the high performance of the unstructured method on the in-domain DECODE Main test set. This result indicates the unstructured approach overfits on superficial patterns in the DECODE Main training data which are still present due to RCT's construction process.[5] We also provide further analysis in Appendix E, including experiments showing that simply removing speaker utterances not uttered by the last speaker does not greatly improve the unstructured method. The fact that the utterance-based approach has good transferability to the OOD human-bot test set indicates that injecting the correct inductive structure bias is beneficial for modeling dialogue consistency. We believe this is an interesting result generally for research using Transformers, where there is currently a belief amongst some practitioners that they can just use a standard Transformer and it will learn all the structure correctly on its own. In our setting that is not the case, and we provide a method that can rectify that failing.

In general, there is still much room for improvement. The results in Table 2 also demonstrate that the modeling of dialogue consistency is a demanding task. On the contradiction detection task, the best score achieved by the state-of-the-art pretrained language models on DECODE (Test-Strict) is 80.86% and the best human-bot test score is 84.69%. Considering all the examples in the test sets are verified by at least 3 annotators, humans are able to swiftly identify such contradictions. This suggests there is a large ability gap between our best automatic detectors and humans. Closing this gap is an important challenge for the community.

5.2 Performance in an Interactive Setting

Model vs. Human Judgment. To further understand the detector predictions and how well they might align with human judgments, we consider the Human-Bot data again. We first divide all the utterances into two categories based on whether they are generated by a human or a bot. Then, the bot-generated utterances that have been marked by annotators as contradicting utterances are categorized into three sets based on the number of annotators that agree on the contradiction label. By design, the more annotators that agree on the contradiction label, the more plausible that it is a contradiction. We examine detector model fire rate on the utterances in the 5 different categories and results are shown in Figure 3. The fire rate of utterance-based RoBERTa trained on DECODE on human utterances is 5.5% contrasting to the 74.3% on 3-agreed contradicting utterances, whereas the fire rates of unstructured RoBERTa on different categories are more clustered together. This finding demonstrates that our models can discriminate between utterances with a distinct nature, and the model predictions are aligned with human judgments. Moreover, a strong discriminative detector could be a useful tool to stratify utterances.

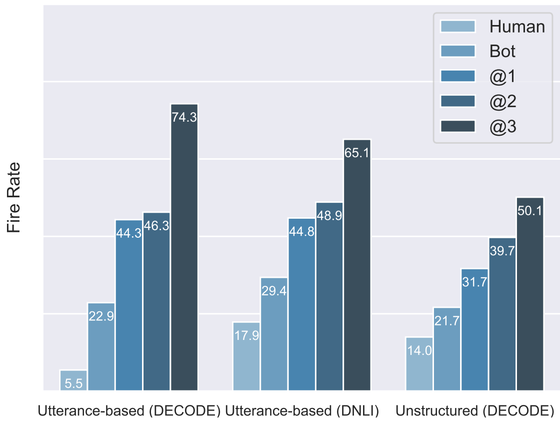

Figure 3: The fire rate (the percentage that it predicts "contradiction") of RoBERTa models with different setups on utterances belonging to different categories. "Human" and "Bot" stand for utterances by the human or the bot prospectively. "@N" indicates the category where N annotators agreed on the contradiction label. The x-axis indicates different approaches and the text in parentheses denotes the training data.

Using DECODE as an Automatic Metric. The results presented above indicate that the prediction of the detector can easily differentiate between the quality of utterances by humans and the utterances by bots. We further investigate whether it can differentiate the quality of the utterances by different bots and be used as an automatic metric checking generation consistency. We compare the average contradiction score of the detector with the contradiction rate by human judgments on the utterances generated by different classes of model (bots). The bots are the same set of models described in subsection 3.3 from which we collected our human-bot

[5] Overfitting on superficial patterns is a typical issue and open problem in NLU modeling (Nie et al., 2020a).

Figure 4: The comparison between the average contradiction score by the detector (y-axis) and the human identified contradiction rate (x-axis) on the utterances by different bots, averaged by type of bot. Each point in the plot is a bot which has conversed with humans and produced at least 180 utterances (with some identified as contradictions) in our interactive settings.

dialogue examples. The trend in Figure 4 reveals that the scores are positively correlated with human judgments, with a Pearson correlation coefficient of 0.81. We would expect that improvement on the DECODE task will directly increase the correlation between the automatically produced detection score and human judgments, where use of such an automatic metric can ease the burden on laborious human evaluation of consistency.

5.3 Generation Re-ranking

Given a contradiction detector, an obvious question other than using it as an automatic metric, is: can it be used to improve the consistency of dialogue generation models? We consider a very simple way to do that in the state-of-the-art generative model, BlenderBot (BST 2.7B) (Roller et al., 2020). During the decoding phase, for decoding methods that can output multiple hypotheses, we simply rerank the top scoring hypotheses using the contradiction detection classifier. We use our best performing classifier, our utterance-based RoBERTa model with DECODE fine-tuning, and consider three methods of decoding: beam search, top-k sampling (Fan et al., 2018) and sample-and-rank (Adiwardana et al., 2020), and compare the standard and DECODE-reranked decoding methods to each other. For beam search we use the best found parameters from (Roller et al., 2020) which are beam size 10, minimum beam length 20 and beam blocking of 3-grams. For top-k we use $k = 40$. For Sample-and-Rank we use k=40 and 20 samples. We consider the same human-bot dialogue logs as before, but only between Blenderbot BST 2.7B and humans, selecting only contradicting utterances. Table 3 presents the results.

Model + Decoding Strategy	DECODE Contradict%	Human Contradict%
Standard generation		
Beam Search	69.7%	84.2%
Top-k ($k = 40$)	42.1%	69.7%
Sample-and-Rank	39.5%	55.3%
DECODE Re-ranking		
Beam Search	46.1%	55.3%
Top-k ($k = 40$)	2.6%	39.5%

Table 3: Generation Re-ranking using DECODE vs. standard methods, reporting the contradiction % as flagged by our contradiction detection classifier (i.e., an automatic metric, "DECODE Contradict%") in addition to human judgments ("Human Contradict%").

Automatic metric using DECODE. Using our same DECODE contradiction classifier as the automatic metric, as in subsection 5.2, we observe that by re-ranking the beam of beam search (size 10) we can improve the metric. Still, 46.1% of the time the detector flags generations as contradictions (vs. 69.7% without re-ranking). Upon observation of the outputs, this seems to be due to the beam of beam decoding not being diverse enough (Vijayakumar et al., 2016): when the top scoring utterance is flagged as contradicting, many of the other utterances in the beam are similar responses with slight rephrases, and are flagged contradicting as well. Top-k sampling fares much better, where reranking in our test can very often find at least one from the $k = 40$ samples that does not flag the classifier, leaving only a 2.6% contradiction firing rate. We note we expect these numbers are over-optimistically low because the metric itself is being used to search (re-rank) and evaluate in this case.

Human Judgments. The last column of Table 3 presents human judgments of the various model generations, judged using the same approach as before with human verifiers, and reporting the percentage of contradictions. We observe similar results to the automatic metric findings. DECODE re-ranking reduces the number of contradictions, particularly for Top-k re-ranking vs. Top-k: testing for significance with a Wilcoxon signed-rank test, we get $p = 0.051$ using two human verifiers and $p = 0.023$ for three verifiers. More detailed results and analysis can be found in Appendix G.

6 Conclusion

We introduce the DialoguE COntradiction DEtection task (DECODE) and a new conversational dataset containing both human-human and human-bot contradictory dialogues. Training models on DECODE achieves better performance than other existing NLI data by a large margin. We further propose a structured utterance-based approach where utterances are paired before being fed into Transformer NLI models to tackle the dialogue contradiction detection task. We show the superiority of such an approach when transferring to out-of-distribution dialogues compared to a standard unstructured approach representative of mainstream NLU modeling. We further show that our best contradiction detector correlates with human judgments, and provide evidence for its usage in both automatic checking and improving the consistency of state-of-the-art generative chatbots.

Acknowledgement

We thank the reviewers, and Jie Lei and Hao Tan for their helpful discussions. YN interned at Facebook. YN and MB were later sponsored by NSF-CAREER Award 1846185, DARPA MCS Grant N66001-19-2-4031, and DARPA YFA17-D17AP00022.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Jeesoo Bang, Hyungjong Noh, Yonghee Kim, and Gary Geunbae Lee. 2015. Example-based chat-oriented dialogue system with personalized long-term memory. In *2015 International Conference on Big Data and Smart Computing (BigComp)*, pages 238–243. IEEE.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Andrew P Bradley. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

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

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2020. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition*, pages 187–208. Springer.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019a. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

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. *arXiv preprint arXiv:1907.11692*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Gary Marcus. 2018. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

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

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1):81–103.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020a. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020b. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial*

Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden, pages 4279–4285. ijcai.org.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

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.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301. AAAI Press.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering output style and topic in neural response generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2140–2150, Copenhagen, Denmark. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics.

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.

S Worsnick. 2018. Mitsuku wins loebner prize 2018.

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.

Wei-Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. 2019a. Neural personalized response generation as domain adaptation. *World Wide Web*, 22(4):1427–1446.

Yizhe Zhang, Xiang Gao, Sungjin Lee, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019b. Consistent dialogue generation with self-supervised feature learning. *arXiv preprint arXiv:1903.05759*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

A Annotation Interface

In the main paper, we describe the procedure of the data collection. Figure 5 shows the collection user interface.

B Annotation Quality Control

We apply the following mechanism to ensure the quality of collected data:

- **Onboarding Test**: Every annotator needs to pass an onboarding test before they can actually contribute dialogue examples. The test is the same dialogue contradiction detection task as in the actual collection procedure, including 5 dialogues where 3 of them have an ending utterance that contradicts the dialogue history. The annotator needs to select the correct label (contradiction or non-contradiction) for all five dialogues to pass the test. This mechanism tests whether an annotator understands the task.
- **Maximum Annotation Count Limit**: The maximum number of examples one annotator can create is 20. This mechanism helps further diversify the dialogue examples by reducing similar patterns that appear in one or a group of annotators (Geva et al., 2019).
- **Verification**: This subtask ensures that the dialogue examples indeed contain an ending utterance that contradicts the dialogue history. We ask 3 additional annotators to verify some of the collected examples and select the ones where all three verifiers agreed on the contradiction label, and use these for our resulting validation and tests sets. This mechanism ensures that there is a clear, agreed-upon contradiction in the dialogue, preventing the subjectivity and ambiguity issues in some NLU tasks (Nie et al., 2020b).

A pilot study with over 100 workers was conducted before the collection which then went through an internal review process and we do not collect any personal information of the workers.

C Data Statistics

Table 8 shows the breakdown of dialogue sources and data splits. For a subset of the contradicting dialogues in DECODE we asked three verifiers to determine whether the original writer indeed created a contradiction example. Table 4 shows the verification statistics. Note that we only use examples on which all three verifiers agreed for DECODE (dev) and DECODE (Test).

D Examples

As described in the main paper, DECODE consists of dialogues belonging to four categories, namely, Human-Human, Human-Bot, A2T, and RCT. Table 5 shows one example for each dataset type.

E Extra Results Analysis

Table 6 shows the performance of unstructured method when the input consists of utterances from both speakers (the default unstructured approach) and when the input consists of utterances from only the last speaker. The numbers for default two speaker unstructured approach and the utterance-based approach match with that in Table 2. The result indicates that removing speaker utterances not uttered by the last speaker does not greatly improve generalization of the unstructured method. This helps show that the out-of-domain improvement from the structured utterance-based method on human-bot data comes from the structure of the architecture.

F Performance in an Interactive Setting

The results discussed in the main paper evaluate models on constructed datasets with intentionally balanced labels. This facilitates the comparison between models following a NLU evaluation perspective. In practice, we would like to evaluate how well a model can detect contradicting utterances sampled naturally from interactive human-bot dialogue. To that end, we test our trained detection models on the raw interactive human-bot dialogue data[6] having a total number of 764 dialogues consisting of 8,933 utterances. Since the contradiction task in naturally sampled dialogue can be extremely unbalanced, the total number of contradicting utterances in the raw dialogue list is only 381[7]. We apply our contradiction detectors on every bot-generated utterance and calculate the precision, recall, and F1 on contradiction detection. Since the scores might be subjective to the threshold τ, we also evaluate the threshold-invariant Area Under the ROC Curve (AUC) (Bradley, 1997).

As shown in Table 7, model precision on the task is not satisfactory (23.94 at best). However, the best model achieves acceptable scores on both Recall and AUC. This indicates its potential usage for strict blocking of inconsistent utterances

[6] This is the same set of dialogues from which we constructed the balanced human-bot test set.

[7] The majority baseline accuracy is 95.73%.

# of Verifiers Agreed	Count	Ratio (%)
0	484	7.67%
1	497	7.87%
2	1,211	19.18%
3	6,214	65.28%

Table 4: Verification Statistics. The first column indicates the number of verifiers that agreed upon the given contradictions.

of a generative model (bot). The table also draws the same conclusion as Table 2 that the structured utterance-based RoBERTa model trained using DECODE data is the best method for contradiction detection, comparing to training on other NLI data or using an unstructured approach. In the following sections we thus use that best method as our detector for further experiments.

G Generation Re-ranking

We show in Table 9 human judgments for our generation re-reranking experiments in three settings: with at least two human verifiers, with three agreeing, or treating agreements as a fractional contradiction score. The first two, for a given utterance, assign a binary score (either contradicton or non-contradiction) depending on whether at least 2 or 3 human verifiers agree on the contradiction label. The last setting treats a given utterance as having a fractional score, either 0, 1/3, 2/3, or 3/3 depending on how many human verifiers label it as a contradiction. We then take the mean over all utterances in each setting to give the final contradiction count per setting.

In addition to the setting in the main paper (subsection 5.3), we also consider the setting where the dialogue examples we use consist of 76 examples utterances that were identified by humans as being contradictions by BlenderBot (using beam search) and 100 examples that were not. This is in contrast to Table 3 where we only considered contradicting utterances by BlenderBot only. The results are given in Table 10. We find similar results to the main paper's results but where the model's score are closer together. This should be expected as when selecting many utterances that are already non-contradicting in the original BlenderBot generations, there is not much left to improve.

Main (Human-Human)
Human1: *I recently had a long weekend with some old friends. It was fantastic.*
Human2: *It must have been fun to catch up with them.*
Human1: *it was. We've spent a lot of time together and apart now, so it was good to catchup.*
Human2: *Well I hope you guys continue to stay in touch.*
Human1: *Sometimes friends grow apart over time though.*
Human2: *Yeah, that's true. You should forget about them and go on with your life.*

Human-Bot:
Human1: *It's a beautiful day. I'm going to take a walk later.*
Bot: *What do you do for a living?*
Human1: *I work in a doctor's office. How about yourself?*
Bot: *I am a stay at home mom.*
Human1: *That's great! How many kids do you have?*
Bot: *I don't have any kids yet, but I plan on getting married soon.*

Add Two Turns (A2T):
Human1: *i hate when ash from my cigarette drops in my work pants*
Human2: *oof that sucks really bad*
Human1: *yeah, i haave to wait till i get home to get the stain off, it is really embarrassing*
Human2: *yea i can imagine it is*
Human1: *Every time I look at it I remember the good times we had together.*
Human2: *well thats nice*
Human1: *I will have to wash the stain with soap and water.*
Human2: *Ash stains on your pants is not a big deal though.*

Remove Contradicting Turns (RCT):
Human1: *I was disgusted when I noticed the food on the table*
Human2: *What kind of food?*
~~Human1: *It was brussel sprouts and Liver*~~
~~Human2: *Oh, disgusting.*~~
Human1: *I couldn't even bear to take a single bite*
Human2: *Brussel sprouts and liver sounds delicious to me!*

Table 5: Dialogue examples for different dataset types. Underline indicates that the pair of utterances is randomly added. Strikethrough text indicates that the pair of utterances is removed. Dialogue examples for Human-Human, Human-Bot, and A2T end with a contradicting utterance whereas the example for RCT has an ending utterance whereby the original contradicting pair of utterances in the dialogue history are removed.

Approach	MT (Acc.)	HB (Acc.)
Unstructured (both speaker)	96.85	70.03
Unstructured (one speaker)	96.68	73.17
Utterance-based	93.19	**84.69**

Table 6: Performance of RoBERTa trained on DECODE data with different approaches. "MT" and "HB" columns show model accuracy on the Main Human-Human Test set and the Human-Bot set, respectively.

Training Data	Precision	Recall	F1	AUC
Unstructured Approach				
All	15.89	60.11	25.14	80.47
All - DECODE	15.63	57.74	24.60	71.82
DECODE	17.05	50.13	25.45	73.40
Utterance-based Approach				
All	23.35	71.65	35.23	84.96
All - DECODE	17.17	68.50	27.46	80.09
DNLI	16.32	65.09	26.09	79.29
ANLI-R3	22.52	41.73	29.26	76.36
DECODE	**23.94**	**74.28**	**36.21**	**87.16**

Table 7: RoBERTa performance on all the bot-generated utterances from the raw interactive human-bot dialogue. The threshold τ for prediction is 0.5.

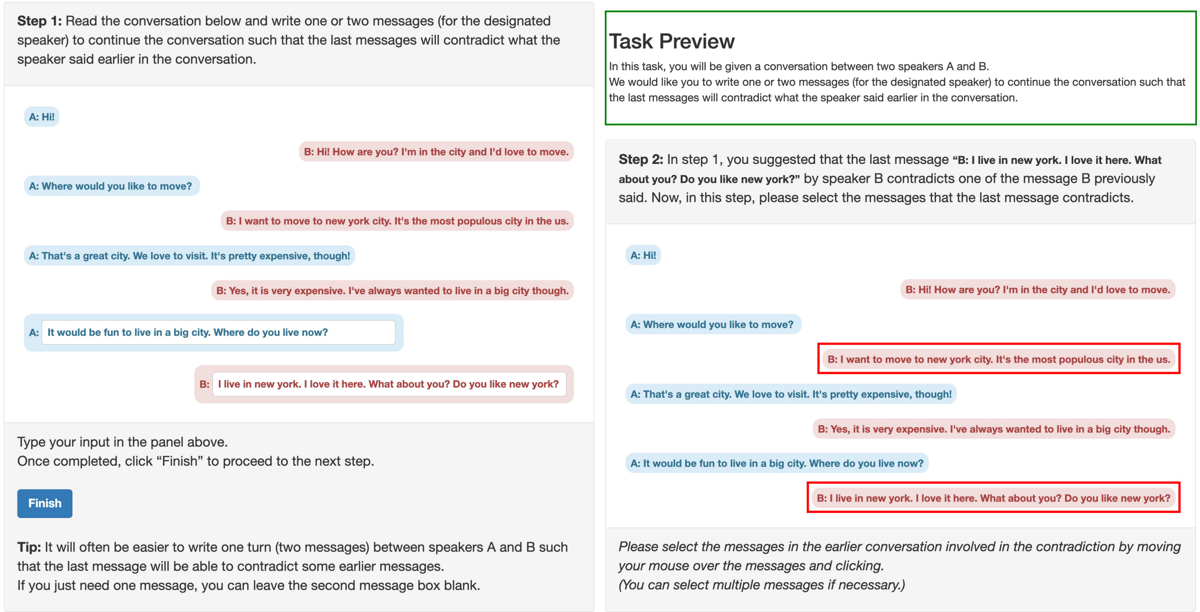

Figure 5: The collection interface. The task preview box (top right) gives a short description of the task before the annotator will work on the writing. The collection consists of two steps. In Step 1 (on the left), the annotators are asked to write one or two utterances such that the last utterance will contradict some previous utterances in the conversation. In Step 2 (on the right), the annotators are asked to pick the utterances in the conversation that are involved in the contradiction. We use a casual term "message" instead of "utterance" in the instructions.

	Train	Dev	Test
Wizard of Wikipedia	6,234	1,208	1,160
EMPATHETICDIALOGUES	6,182	1,046	1,050
Blended Skill Talk	8,554	1,200	1,310
ConvAI2	6,214	572	696
Total	**27,184**	**4,026**	**4,216**

Table 8: Our DECODE Main Dataset source statistics. The labels in each split are balanced. There are a total of 2,013+2,108 contradicting examples in the dev and test sets which are the collected 4,121 verified examples. The first column indicates the source of the dialogue.

Model + Decoding Strategy	**Human Contradict% 2-agree**	**3-agree**	**fractional**
Standard generation			
Beam Search	84.2%	42.1%	75.0%
Top-k ($k = 40$)	69.7%	44.7%	66.2%
Sample-and-Rank	55.3%	31.6%	52.2%
DECODE Re-ranking			
Beam Search	55.3%	29.0%	49.7%
Top-k ($k = 40$)	39.5%	13.2%	39.9%

Table 9: Generation Re-ranking using DECODE vs. standard methods, reporting the contradiction % as flagged by human judgments ("Human Contradict%") in three settings: with at least two human verifiers, with three agreeing, or treating agreements as a fractional contradiction score.

Model + Decoding Strategy	**DECODE Contradict%**	**Human Contradict%**
Standard generation		
Beam Search	38.1%	38.3%
Top-k ($k = 40$)	29.0%	31.8%
Sample-and-Rank	29.6%	29.0%
DECODE Re-ranking		
Beam Search	22.7%	32.0%
Top-k ($k = 40$)	1.1%	25.6%

Table 10: Generation Re-ranking using DECODE vs. standard methods, reporting the contradiction % as flagged by our contradiction detection classifier (i.e., an automatic metric, "DECODE Contradict%") in addition to human judgments ("Human Contradict%"). In this setting, the set of dialogue examples we use consists of 76 examples utterances that were identified by humans as being contradictions by BlenderBot (using beam search) and 100 examples that were not. (In contrast, Table 3 only considered contradicting utterances by BlenderBot only.) We find similar results to the main paper's results but where the model's score are closer together. This should be expected as when selecting many utterances that are already non-contradicting there is not much left to improve.