

# Improving Bot Response Contradiction Detection via Utterance Rewriting

Di Jin, Sijia Liu, Yang Liu, Dilek Hakkani-Tur

Amazon Alexa AI

{djinamzn, sijial, yangliud, hakkaniit}@amazon.com

## Abstract

Though chatbots based on large neural models can often produce fluent responses in open domain conversations, one salient error type is contradiction or inconsistency with the preceding conversation turns. Previous work has treated contradiction detection in bot responses as a task similar to natural language inference, e.g., detect the contradiction between a pair of bot utterances. However, utterances in conversations may contain co-references or ellipsis, and using these utterances as is may not always be sufficient for identifying contradictions. This work aims to improve the contradiction detection via rewriting all bot utterances to restore antecedents and ellipsis. We curated a new dataset for utterance rewriting and built a rewriting model on it. We empirically demonstrate that this model can produce satisfactory rewrites to make bot utterances more complete. Furthermore, using rewritten utterances improves contradiction detection performance significantly, e.g., the AUPR and joint accuracy scores (detecting contradiction along with evidence) increase by 6.5% and 4.5% (absolute increase), respectively.

## 1 Introduction

Latest chatbots powered by large pre-trained neural models have shown decent capabilities to maintain fluent and interesting conversations with human users (Paranjape et al., 2020; Roller et al., 2021; Bao et al., 2021; Konrád et al., 2021). However, they are still prone to various kinds of annoying mistakes (Xu et al., 2020; See and Manning, 2021; Xu et al., 2021). One such error is contradiction or inconsistency, as illustrated in Table 1.

In order to reduce contradiction errors, one approach is to develop a detection model to identify such problems after a system produces response candidates. To this end, Welleck et al. (2019) characterized the modeling of persona-related consistency as a natural language inference (NLI) problem and constructed a dialog NLI

dataset based on Persona-Chat. To cover a broader range of consistency types (e.g., persona, logic, causality, etc), Nie et al. (2021) collected DECODE, a dataset containing human written dialogues with self-contradictory utterances. Besides the in-distribution human-human dialogues test set, they collected an out-of-distribution set containing dialogues between human and different chatbots. This human-bot test set can better evaluate models' performance in detecting contradiction in conversations between human and chatbots, which is the focus of this work.

We find one failure of the state-of-the-art (SOTA) contradiction detection model is due to the frequent anaphora and ellipses in chatbot utterances. One typical example is shown in Table 1, where the first bot utterance has an anaphor, "mine", and the last bot utterance misses an important entity, "Johnny Cash's concert". Such incomplete utterances would prevent detection models from fully understanding the bot utterances in the dialog, thus leading to detection errors. Therefore, we propose to first rewrite the bot utterances to recover all the missing information and then perform the contradiction detection task. To support this goal, we first collect a new dataset for incomplete utterance rewriting, which is a widely studied task (Pan et al., 2019; Su et al., 2019; Hao et al., 2021) but still lacks supporting datasets for open-domain conversations in English (Quan et al., 2019). Then we propose a rewriting model trained on this data to rewrite the anaphors to their corresponding entities and restore any missing content. We conduct experiments on the DECODE dataset (Nie et al., 2021), and demonstrate substantial performance improvement in contradiction detection when the utterance rewrite module is applied. Overall, we have made the following contributions in this work:

- We have collected a new English dataset for incomplete utterance rewriting for general open-domain conversations, and developed a rewrite-

Speakers	Original Dialogues	Rewritten Dialogues	Rewriting Type
Human:	Hi, what's your favorite singer?	—	
Bot:	<b>M</b> ine is johnny cash of course.	<b>My favorite singer</b> is Johnny Cash, of course.	Co-reference
Human:	He's amazing, I love his songs.	—	
Bot:	<b>I never got to see johnny cash play but I wish I did.</b>	I never got to see Johnny Cash play but I wish I did.	None
Human:	Same, I wish I could go to one of his concerts.	—	
Bot:	<b>I have not been since last year though. I like sports.</b>	I have not been <b>to Johnny Cash's concert</b> since last year though. I like sports.	Ellipsis

Table 1: Examples of human-bot conversations with contradictory bot utterances marked by red color. We rewrite every bot utterance to restore co-references and ellipsis (the restored parts are highlighted by bold font).

ing model for utterance restoration.

- With bot utterance rewriting, we can improve the previous best contradiction detection model by 6.5% in AUPR and 4.5% in joint accuracy that considers both contradiction and evidence labels.
- We relabeled the human-bot test set of the benchmark DECODE dataset and corrected some annotations.<sup>1</sup>

## 2 Contradiction Detection Method

### 2.1 Task Definition

We formalize dialogue contradiction detection as an NLI task. Given a list of utterances  $x = \{u_1^H, u_1^B, \dots, u_n^H, u_n^B\}$  representing a dialogue, the task is to determine if the last bot utterance  $u_n^B$  contradicts any previously conveyed information contained in the past bot utterances  $\{u_1^B, \dots, u_{n-1}^B\}$ . Note that we are using human and bot alternating turns here (referred to as H and B), but they can be human-human conversations too. In addition to the binary label  $y$ , with 0 or 1 corresponding to the non-contradiction and the contradiction labels, respectively, we also output a set of indices  $I \in \{1, \dots, n-1\}$  representing the utterances in  $\{u_1^B, \dots, u_{n-1}^B\}$  that is actually contradicted by the last utterance  $u_n^B$ .

### 2.2 Detection Models

Based on the benchmark DECODE dataset, Nie et al. (2021) proposed two approaches for contradiction detection: an unstructured approach and a structured utterance-based (SUB) approach. The former one concatenates all the previous utterances in the dialogue history to form a single textual context. Then a classification model  $f_\theta$  is applied to the context and the last utterance to infer the probability of contradiction. The latter SUB approach pairs every past bot utterance with the last one, and then

<sup>1</sup>Code and data are released at: <https://github.com/jind11/utterance-rewriting>

feeds each pair to the classification model  $f_\theta^{SUB}$ . The final contradiction probability is the maximum over all the outputs:  $\hat{y} = \max\{f_\theta^{SUB}(u_i^B, u_n^B) : i \in \{1, \dots, n-1\}\}$ . The supporting evidence (SE) for a contradiction decision contains the pairs having contradiction probability higher than a threshold  $\eta$ , i.e.,  $I = \{i : f_\theta^{SUB}(u_i^B, u_n^B) > \eta\}$ . Nie et al. (2021) demonstrated that the latter SUB approach significantly outperforms the former one on the human-bot test set (more than 10% in accuracy). This SUB method is **the current SOTA model for contradiction detection**, which we adopted as one baseline.

### 2.3 Utterance Rewriting for Contradiction Detection

As discussed earlier, we noticed that many bot utterances contain co-references and ellipses and thus the baseline model fails to capture the semantic meaning or contradiction in the sentence pair. Therefore, we propose to first rewrite the bot utterances to restore co-references and ellipsis, and then feed the rewritten utterances (e.g., the dialogues on the right in Table 1) to the model. To this end, we first collect a new dataset specially for utterance rewriting and then develop a rewriting model.

**Rewriting Data Collection** To get parallel training data for utterance rewriting for open-domain conversations, we sub-sampled 6,000 and 4,000 dialogues from the DailyDialog (Li et al., 2017) and BST (Smith et al., 2020) datasets, respectively, as the training set. Besides, we sub-sampled 400 and 400 dialogues from DailyDialog and BST, respectively, as the test set. We only use the first six utterances in each dialog. Specifically, we use the first two utterances (from both speakers) as leading context and ask annotators to check the remaining four utterances, following Pan et al. (2019).<sup>2</sup> Empirically we find that the context information

<sup>2</sup>Utterance rewriting needs context to resolve co-references and ellipsis, and thus the first two utterances are not suitable for rewriting annotation.

needed to resolve co-references and ellipsis can always be found within 1-3 turns (Pan et al., 2019; Su et al., 2019). We ask annotators to identify whether an utterance is complete and can be understood without reading the context, and if not, then rewrite it to restore any missing information.

To ensure the annotation quality, we hired three in-house professional data annotators, who have been first trained via a pilot annotation session and then proceed to the official annotation phase after passing our provided qualification set. In the official annotation phase, two of them first worked independently and then the third annotator was tasked to make the adjudication over the two annotations and pick the best one or make revisions if needed. Besides, we periodically sampled 10% of the annotations from each annotator throughout the annotation process and provided feedback. The annotation is considered valid only when the accuracy of examined results surpassed 95% (we deem those rewrites that are both correct and complete as correct rewrites, and then calculate the percentage of correct rewrites as the accuracy). Overall, we have obtained 40,000 and 3,200 samples for training and testing, respectively.

**Rewriting Model** We treat rewriting as a sequence-to-sequence (Seq2Seq) task and adopt two pre-trained Seq2Seq models, T5 (Raffel et al., 2019) and Pegasus (Zhang et al., 2019). The input is the concatenated context utterances and the original last utterance, with special tokens inserted before each utterance to indicate its speaker.

## 3 Experiments

### 3.1 Contradiction Detection Data

We use the DECODE dataset (Nie et al., 2021) in this study. However, we found some issues with its human-bot test set: (1) Around one third of non-contradiction dialogues contain only one human and one bot utterances, which makes the detection task over-simplified, since there are no previous bot utterances. (2) Not every bot utterance has been annotated for contradiction with respect to its history. (3) Evidence is not labeled to indicate which history bot utterance contradicts the last one.

To resolve the above-mentioned issues, we curate new annotation using the dialogues in the original test set. Details of annotation procedures are provided in Section A of the Appendix. Overall, we have obtained 1,889 samples (453 positive samples and 1,436 negative ones), which we call an unbal-

anced set. Besides, we sub-sampled 453 negative samples and combined them with all the positive ones to form a balanced set. Table A.1 summarizes the data statistics. We will release this new test set.

### 3.2 Baselines

We compare the contradiction detection performance with and without rewriting bot utterances, all based on the same SUB model framework, which is the current SOTA model for contradiction detection. Another baseline we introduced is SUB-CONCAT, where each bot utterance is the concatenation of the original one with the preceding human utterance such that the missing information (coreference or ellipsis) can be recovered from the included previous utterance.

For rewriting, we compare our model against four strong baselines: one is the off-the-shelf SOTA co-reference resolution model trained on OntoNotes (named as “Co-reference”) (Toshniwal et al., 2021; Wu et al., 2020), and the other three are developed based on three related datasets for rewriting, named as “CANARD” (Elgohary et al., 2019), “Gunrock” (Zhang et al., 2020), and “MuDoCo” (Tseng et al., 2021). Specifically, CANARD is a query rewriting dataset that aims to rewrite a query/question based on previous consecutive QA pairs for the conversational question answering task. The Gunrock dataset focuses on resolving ellipsis while containing a small portion of co-reference cases, and it consists of 1745 samples where all dialogues are in-house curated following the Alexa Prize competition format. The MuDoCo dataset is also for query rewriting for task-oriented dialogues covering 6 domains.

### 3.3 Evaluation Metrics

To evaluate incomplete utterance rewriting, we use both automatic and human evaluation. For human evaluation, we propose two metrics: (1) Correctness; (2) Completeness. The former one checks whether the rewriting part is correct and obeys the information in dialogue context, while the latter one checks whether the rewritten utterance is complete enough to be understood without reading the context. We have binary labels for both metrics and report the percentage of positive labels after human evaluation. For automatic evaluation, in addition to the widely used BLEU (Papineni et al., 2002), ROUGE-1 (R-1), and ROUGE-L (R-L) (Lin, 2004), we have added two more metrics specially for evaluating text editing models: exact match

(EM) accuracy, and the  $F_1$  score, which was proposed in Pan et al. (2019) and focuses on n-grams that contain at least one restored word. Specifically, the n-gram restoration precision, recall, and F-score can be calculated as:

$$P_n = \frac{\{\text{restored n-grams}\} \cap \{\text{n-grams in ref}\}}{\{\text{restored n-grams}\}}$$

$$R_n = \frac{\{\text{restored n-grams}\} \cap \{\text{n-grams in ref}\}}{\{\text{n-grams in ref}\}}$$

$$F_n = 2 \cdot \frac{P_n * R_n}{P_n + R_n}$$

where “restored n-grams” refer to the n-grams in restored utterance that contain at least one restored words, and “n-grams in ref” refer to the n-grams in reference that contain at least one restored words.

For contradiction detection, we first set the threshold  $\eta$  to be 0.5, and report Precision/Recall/F1 for both the binary contradiction label and the support evidence labels, following Thorne et al. (2018).<sup>3</sup> Besides, we report Joint Accuracy, which indicates the performance when both the 2-way contradiction detection and the supporting evidence retrieval are correct. Considering that these scores are sensitive to  $\eta$ , we also report Area-under-Precision-Recall-Curve (AUPR) as a threshold-independent score.

### 3.4 Experimental Setup

For utterance rewriting, we have used three kinds of pre-trained models: T5-Base, T5-Large, and Pegasus-Large, whose parameter sizes are 220 M, 770 M, and 568 M. Each model is trained for 4 epochs with a learning rate of  $5e^{-5}$ , and beam search (beam size of 5) is used for generation.

For contradiction detection, following Nie et al. (2021), we used the RoBERTa-Large model whose parameter number is 330 M, which is trained for 3 epochs with a learning rate of  $1e^{-5}$ . We have used the Huggingface Transformer code base<sup>4</sup> and all experiments were run on Nvidia V100 GPUs.

## 4 Results and Discussion

### 4.1 Utterance Rewriting

We performed both automatic and human evaluation for utterance rewriting (please refer to Section 3.3 for evaluation details). Table 3 summarizes the automatic evaluation results. As can be seen, the three models perform similarly overall, with T5-Large slightly outperforming the other two. We

<sup>3</sup><https://github.com/sheffieldnlp/fever-scorer>

<sup>4</sup><https://github.com/huggingface/transformers/tree/master>

thus adopt it as the main rewriting model in later experiments.

We also sub-sample 100 rewritten utterances by T5-Large for human evaluation. As shown in Table 4, the correctness and completeness scores for both test sets are above 85%, validating the high-quality of the rewriting model. We also report the change rate in the table that defines the percentage of the rewritten utterances that are different from the original ones (only differences in punctuation and upper/lower-case are not considered). The bottom block of Table 4 shows the percentage of utterances containing co-reference or ellipsis, or either, i.e, incomplete utterances. We see that **co-reference and ellipsis occur almost equally frequently** in incomplete utterances. Considering all the numbers together, we demonstrate that the rewriting model has covered most of those incomplete utterances.

### 4.2 Contradiction Detection

Table 2 compares the contradiction detection performance without rewriting and with rewriting by different rewriting models. First of all, the SUB-Concat method without rewriting does not yield any performance gain although it has included the context utterances. More importantly, after rewriting all bot utterances for both training and test sets, only our rewriting model can lead to significant improvements for all the evaluation metrics, while those baseline rewriting models either maintain or deteriorate the performance (we provided the rewriting performance of these baselines in Section B of Appendix for reference). We see that the AUPR metric has been improved by around 2.8% and 3.2% absolutely for the balanced and unbalanced sets by our model, respectively. We also implemented model ensemble where we rewrite bot utterances using our three rewriting models (T5-Base/Large and Pegasus-Large), run contradiction detection using each, and average their contradiction scores to obtain the final prediction. This further improves the detection performance over single models. Overall, we have achieved a substantial increase of 4.2% and 6.5% for AUPR and 4.5% and 3.4% for Joint-Acc. for the balanced and unbalanced sets, respectively.

### 4.3 Error Analysis

We conducted additional error analysis to understand the performance gains and remaining errors. We first obtained 95 false negative samples by the

Detection Method	Rewriting Model	Balanced Set				Unbalanced Set		
		P/R/F1	AUPR	SE (P/R/F1)	Joint-Acc.	P/R/F1	AUPR	Joint-Acc.
SUB-Bot only	None	89.4/70.6/78.9	89.0	90.4/62.9/74.2	69.9	73.2/70.6/71.9	75.4	81.4
SUB-Concat		88.1/66.9/76.0	88.1	90.0/60.0/72.0	68.1	66.3/66.9/66.6	71.6	78.7
SUB-Bot only	Co-reference	89.0/71.3/79.2	89.1	90.0/64.5/75.2	69.7	73.1/71.3/72.2	75.8	81.3
	CANARD	79.3/37.1/50.5	73.9	89.6/26.3/40.7	54.6	60.0/37.8/46.3	52.8	74.8
	Gunrock	79.2/59.8/68.2	74.0	88.5/50.2/64.0	60.0	53.0/59.8/56.2	49.0	71.9
	MuDoCo	88.1/65.3/75.0	87.4	91.6/59.5/72.1	67.9	70.6/65.3/67.9	71.2	80.2
SUB-Bot only	Ours (single)	<u>90.9/72.9/80.9</u>	91.8	<u>93.0/67.6/78.3</u>	73.6	<u>73.5/72.9/73.2</u>	78.6	82.8
	Ours (ensemble)	<b>92.9/71.7/81.0</b>	<b>93.2</b>	<b>93.9/66.1/77.6</b>	<b>74.4</b>	<b>80.1/71.7/75.7</b>	<b>81.9</b>	<b>84.8</b>

Table 2: Contradiction detection performance (%) on new human-bot test set. Best results for single model (T5-Large) are marked by underlines while overall best results are marked bold. ‘SUB-Bot only’ means feeding only bot utterances to the SUB model while SUB-Concat uses the concatenated bot utterance and the preceding human turn.

Models	BLEU	R-1	R-L	EM	F <sub>1</sub>
T5-Base	0.653	0.822	0.801	0.213	0.402
T5-Large	0.653	0.820	0.798	0.199	0.422
Pegasus-Large	0.649	0.822	0.801	0.212	0.391
Agreement	0.714	0.840	0.837	0.323	0.309

Table 3: Automatic evaluation results for the rewriting model on the rewriting test set. Agreement is the inter-annotator agreement between two rewrites in test set.

Test Set	Correctness	Completeness	Change Rate
Rewriting	92.0	85.0	59.0
Contradiction	98.0	93.1	62.4
Test Set	Co-reference	Ellipsis	Incomplete
Rewriting	39.0	42.0	68.0
Contradiction	42.6	27.7	58.4

Table 4: Upper block: human evaluation of rewriting for both the rewriting and contradiction detection test sets (%); bottom block: percentage of utterances containing co-reference or ellipsis, or either (incomplete).

“SUB-Bot only” model without applying rewriting, and then manually identified 28 samples whose last bot utterances are incomplete. We then manually rewrote those incomplete bot utterances. With such manual rewriting, we are able to correctly classify 18 out of 28 samples to be positive (64.3% in accuracy), whereas, with the T5-Large rewriting model, 15 samples can be correctly predicted (53.6% in accuracy). This comparison indicates that our automatic rewriting has pushed the performance improvement close to the upper bound achieved by manual rewriting. More error analysis is provided in Section C of Appendix.

#### 4.4 Why Utterance Rewriting Helps?

As illustrated by Table 1, in order to infer the entailment relationship between the premise (i.e. “Mine is johnny cash of course.”) and hypothesis (i.e. “I have not been since last year though.”), we need to resolve the anaphora and ellipses so that some key information can be restored, e.g., “Mine” is replaced by “My favorite singer” in the premise and

the missing phrase of “to Johnny Cash’s concert” is restored in the hypothesis. Without restoring such key information from the dialogue context, the contradiction detection model cannot fully understand the premise and hypothesis sentences, thus not being able to accurately detect contradictory cases. One could argue that we can simply concatenate the context with both premise and hypothesis respectively so that the detection model could grab the missing information itself from the context, however, the baseline method “SUB-Concat”, which follows this setting, still under-performs the baseline without concatenating the context (i.e. SUB-Bot only). This indicates that when the premise and hypothesis are organized in a dialogue structure with multiple turns rather than as single-turn sentences, the NLI based detection model is not good at inferring their relationship anymore. Therefore, we need to use the utterance rewriting model to grasp the most necessary information from context and insert into the bot utterances so that we can still use the single-turn format while making up the missing information for entailment inference.

#### 4.5 Future Work

We will keep improving the utterance rewriting model. Besides, we will showcase that utterance rewriting can also help improve other dialogue related tasks, such as task-orientated dialogue state tracking and response generation, open-domain dialogue response selection and generation, etc.

### 5 Conclusion

In this work, we aim to improve contradiction detection in chatbot utterances via rewriting to restore anaphora and ellipsis. To develop such an utterance rewriting model, we curated a dataset by crowd-sourcing and demonstrated that the rewriting quality is satisfactory. With such a rewriting technique, we are able to significantly improve the contradiction detection performance.

## References

- Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2021. [PLATO-2: Towards building an open-domain chatbot via curriculum learning](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2513–2525, Online. Association for Computational Linguistics.
- Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. [Can you unpack that? learning to rewrite questions-in-context](#). 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 5918–5924, Hong Kong, China. Association for Computational Linguistics.
- Jie Hao, Linfeng Song, Liwei Wang, Kun Xu, Zhaopeng Tu, and Dong Yu. 2021. [RAST: Domain-robust dialogue rewriting as sequence tagging](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4913–4924, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jakub Konrád, Jan Pichl, Petr Marek, Petr Lorenc, Van Duy Ta, Ondřej Kobza, Lenka Hýlová, and Jan Šedivý. 2021. [Alquist 4.0: Towards social intelligence using generative models and dialogue personalization](#). *arXiv preprint arXiv:2109.07968*.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. [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.
- Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. [I like fish, especially dolphins: Addressing contradictions in dialogue modeling](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713, Online. Association for Computational Linguistics.
- Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu. 2019. [Improving open-domain dialogue systems via multi-turn incomplete utterance restoration](#). 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 1824–1833, Hong Kong, China. 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.
- Ashwin Paranjape, Abigail See, Kathleen Kenealy, Haojun Li, Amelia Hardy, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, and Christopher D Manning. 2020. [Neural generation meets real people: Towards emotionally engaging mixed-initiative conversations](#). *arXiv preprint arXiv:2008.12348*.
- Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. [GECOR: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue](#). 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 4547–4557, Hong Kong, China. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *arXiv preprint arXiv:1910.10683*.
- Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. [Recipes for building an open-domain chatbot](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.
- Abigail See and Christopher Manning. 2021. [Understanding and predicting user dissatisfaction in a neural generative chatbot](#). In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 1–12, Singapore and Online. Association for Computational Linguistics.
- 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.
- Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. [Improving multi-turn dialogue modelling with utterance ReWriter](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 22–31, Florence, Italy. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. [The fact extraction and VERification \(FEVER\)](#)

- shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.
- Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. [On generalization in coreference resolution](#). In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Bo-Hsiang Tseng, Shruti Bhargava, Jiarui Lu, Joel Ruben Antony Moniz, Dhivya Piraviperumal, Lin Li, and Hong Yu. 2021. [CREAD: Combined resolution of ellipses and anaphora in dialogues](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3390–3406, Online. 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.
- Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. [CorefQA: Coreference resolution as query-based span prediction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for safety in open-domain chatbots. *arXiv preprint arXiv:2010.07079*.
- Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. [Bot-adversarial dialogue for safe conversational agents](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online. Association for Computational Linguistics.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. [Pegasus: Pre-training with extracted gap-sentences for abstractive summarization](#).
- Xiyuan Zhang, Chengxi Li, Dian Yu, Samuel Davidson, and Zhou Yu. 2020. Filling conversation ellipsis for better social dialog understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9587–9595.

## A Contradiction Detection Data Collection

Considering that the original human-bot test set of the benchmark DECODE dataset is problematic, we specially curate new annotation based on those dialogues of the original test set via the following steps: (1) We first obtained 507 unique and full dialogues from the original human-bot test set<sup>1</sup> by merging dialogues with overlaps and removing dialogues of only one turn. We then obtained 1,889 partial dialogues for annotation by cutting each full dialogue from the beginning to each bot utterance so that we can annotate whether each bot utterance contradicts against its context. (2) In the first round of annotation, we ask three Amazon Mechanical Turk workers (from English-speaking countries, including USA, England, and Canada) to annotate both the binary label of contradiction and evidence indices that indicate which history bot utterance contradicts the last one. When setting-up the annotation interface, we have provided one line of guidance to warn annotators not to reveal any personal information during annotation. We keep those samples with three full votes as finalized samples and pass those without three equal votes to the second round. (3) In the second round, we provide the maximum set of evidence indices to another three AMT workers and let them verify and write down new annotation if they do not agree. Again, samples with three agreements are selected as finalized ones and those without are passed to authors of this work for final adjudication. Finally, among all the 1,889 samples, we have obtained 453 positive samples and 1,436 negative ones, which we call an unbalanced set. Besides, we have also sub-sampled 453 negative samples and combine them with all positive ones to form a balanced set. Table A.1 summarizes the data statistics.

Dataset	Positive	Negative	Type
Train	13,592	13,592	Human-Human
Balanced Test	453	453	Human-Bot
Unbalanced Test	453	1,436	Human-Bot

Table A.1: Statistics of the contradiction detection dataset.

## B Rewriting Quality of Baselines

Table B.2 compares our rewriting model with baselines that are developed on three related datasets for

<sup>1</sup><https://github.com/facebookresearch/ParlAI/blob/main/projects/contradiction>

utterance rewriting (CANARD, Gunrock, and MuDoCo) and we report performance on our rewriting test set. As expected, our rewriting model that is trained on our own rewriting dataset performs the best. And by combining Table 2 and Table B.2, we can see that better rewriting performance on our rewriting test set can lead to better contradiction detection performance.

Models	BLEU	R-1	R-L	EM	$F_1$
CANARD	0.363	0.631	0.623	0.105	0.215
Gunrock	0.352	0.626	0.612	0.008	0.223
MuDoCo	0.576	0.783	0.781	0.214	0.133
Ours	0.653	0.820	0.798	0.199	0.422

Table B.2: Automatic evaluation results for the rewriting model on the rewriting test set. Agreement is the inter-annotator agreement between two rewrites in test set.

## C Qualitative Error Analysis

Among all 95 false negative samples predicted by the baseline, we find that the last bot utterances of 28 samples are incomplete and need rewriting for restoration. After automatic rewriting of all bot utterances, we can get 15 samples correct but still have 13 samples being false negatives. Now we would like to analyze the error pattern of these 13 false negatives after rewriting and we can categorize the errors into four types: numerical reasoning, logical reasoning, common sense reasoning, and hard to judge. Table C.3 provides examples for each type. Here are detailed definitions for these four types: (1) Numerical reasoning: Models need to do some numerical calculation or comparison to make a decision; (2) Logical reasoning: It requires logical reasoning for prediction; (3) Common sense reasoning: Some common sense knowledge needs to be equipped for reasoning; (4) Hard to judge: It is even hard for human to judge whether contradiction really exists or not.

Table C.4 provides several examples that are originally false negatives before rewriting bot utterances but later get correctly predicted after rewriting. As can be seen, the rewriting process can make up those critical information needed for detecting contradiction cases. Taking the first sample in Table C.4 as an example, there is ellipsis in the last two bot utterances in the original dialogue, which would lead to models' misunderstanding. After rewriting, the last two bot utterances get complete, which makes it much easier for models' decision making.



Original Dialogue	Rewritten Dialogue	Error Type
<p><b>human:</b> Hi!</p> <p><b>bot:</b> What do you do for a living? I'm retired, so I just sew all day.</p> <p><b>human:</b> I'm a homemaker. What kind of things do you sew?</p> <p><b>bot:</b> I sew a lot of quilts for my grandchildren. Do you have any kids?</p> <p><b>human:</b> Aww that's sweet! Yes, I have one son.</p> <p><b>bot:</b> How old is he? I have two daughters. They're a handful!</p> <p><b>human:</b> He's four and yes they are a handful! How old are your daughters?</p> <p><b>bot:</b> 4, 5, and 8. I love them to death, but they're exhausting!</p>	<p><b>human:</b> Hi!</p> <p><b>bot:</b> What do you do for a living? I'm retired, so I just sew all day.</p> <p><b>human:</b> I'm a homemaker. What kind of things do you sew?</p> <p><b>bot:</b> I sew a lot of quilts for my grandchildren. Do you have any kids?</p> <p><b>human:</b> Aww that's sweet! Yes, I have one son.</p> <p><b>bot:</b> How old is your son? I have two daughters. Your daughters are a handful!</p> <p><b>human:</b> He's four and yes they are a handful! How old are your daughters?</p> <p><b>bot:</b> My daughters are 4, 5 and 8. I love my daughters to death, but they're exhausting!</p>	Numerical Reasoning
<p><b>human:</b> Hi!</p> <p><b>bot:</b> Hi there! How are you? What do you do for a living? I'm a college student.</p> <p><b>human:</b> I'm good, thankyou! I'm an electrician. What do you study?</p> <p><b>bot:</b> It is in education. I just moved to an apartment near my job. What are you studying?</p>	<p><b>human:</b> Hi!</p> <p><b>bot:</b> Hi there! Are you okay? What do you do for a living? I'm a college student.</p> <p><b>human:</b> I'm good, thankyou! I'm an electrician. What do you study?</p> <p><b>bot:</b> I study education. I just moved to an apartment near my job. What are you studying?</p>	Logical Reasoning
<p><b>human:</b> I've never left the US, I think I'm afraid to travel internationally though.</p> <p><b>bot:</b> I can understand that. I have never been to the united states. I would love to though.</p> <p><b>human:</b> Oh I thought you lived in the South United States. Where do you live?</p> <p><b>bot:</b> I live on the south side of the state. It is very cold in the winter, though.</p>	<p><b>human:</b> I've never left the US, I think I'm afraid to travel internationally though.</p> <p><b>bot:</b> I can understand being afraid to travel internationally. I have never been to the united states. I would love to travel internationally though.</p> <p><b>human:</b> Oh I thought you lived in the South United States. Where do you live?</p> <p><b>bot:</b> I live on the south side of the state. It is very cold in the winter though.</p>	Common Sense Reasoning
<p><b>human:</b> What did you do today? I have an adult son.</p> <p><b>bot:</b> I went to the mall with my mom. She's my best friend. How old is your son?</p> <p><b>human:</b> My son is 29. How many children do you have?</p> <p><b>bot:</b> I have three kids. They're all grown up now, but they're still my best friends.</p> <p><b>human:</b> Do your kids live with you?</p> <p><b>bot:</b> Yes, they live with me and my husband. What do you like to do in your spare time?</p> <p><b>human:</b> Have you been shopping lately? I love shopping.</p> <p><b>bot:</b> No, I don't have much time for shopping. I spend most of my time with my kids.</p>	<p><b>human:</b> What did you do today? I have an adult son.</p> <p><b>bot:</b> I went to the mall with my mom. My mom is my best friend. How old is your son?</p> <p><b>human:</b> My son is 29. How many children do you have?</p> <p><b>bot:</b> I have three kids. My kids are all grown up now, but my kids are still my best friends.</p> <p><b>human:</b> Do your kids live with you?</p> <p><b>bot:</b> Yes, my kids live with me and my husband. What do you like to do in your spare time?</p> <p><b>human:</b> Have you been shopping lately? I love shopping.</p> <p><b>bot:</b> No, I don't have much time for shopping. I spend most of my time with my kids.</p>	Hard to Judge

Table C.3: Examples that are wrongly predicted as negatives before and after automatic rewriting.

Original Dialogue	Rewritten Dialogue
<p><b>human:</b> Hi!  <b>bot:</b> Hey!  <b>human:</b> How are you today?  <b>bot:</b> Great! And you?  <b>human:</b> I am doing great as well! What do you do for a living?  <b>bot:</b> I injured in an accident.  <b>human:</b> When and what happened?  <b>bot:</b> I had a sports injury.  <b>human:</b> What kind of accident?  <b>bot:</b> Broke an ankle. Very painful.  <b>human:</b> What sport were you in?  <b>bot:</b> I compete in rowing.  <b>human:</b> How did you break an ankle from rowing?  <b>bot:</b> A four wheeler accident.</p>	<p><b>human:</b> Hi!  <b>bot:</b> Hey!  <b>human:</b> How are you today?  <b>bot:</b> I am great today! And how are you?  <b>human:</b> I am doing great as well! What do you do for a living?  <b>bot:</b> I was injured in an accident.  <b>human:</b> When and what happened?  <b>bot:</b> I had a sports injury that caused me to be injured.  <b>human:</b> What kind of accident?  <b>bot:</b> I broke an ankle. My injury was very painful.  <b>human:</b> What sport were you in?  <b>bot:</b> I compete in rowing when I broke my ankle.  <b>human:</b> How did you break an ankle from rowing?  <b>bot:</b> I broke my ankle in a four wheeler accident.</p>
<p><b>human:</b> Hi!  <b>bot:</b> Hello, how are you? I'm a huge fan of star wars. How about you?  <b>human:</b> I have never seen a Star Wars movie.  <b>bot:</b> Neither have I. What do you like to do in your spare time? I like coffees.</p>	<p><b>human:</b> Hi!  <b>bot:</b> Hello, how are you? I'm a huge fan of star wars. How about you?  <b>human:</b> I have never seen a Star Wars movie.  <b>bot:</b> I have never seen a Star Wars movie. What do you like to do in your spare time? I like coffees.</p>
<p><b>human:</b> Hi!  <b>bot:</b> What do you like to do in your spare time? I like to run and play soccer.  <b>human:</b> I like to work on cars in my spare time. Do you play any sports?  <b>bot:</b> No, I don't. I'm not very athletic. What kind of cars do you work on?</p>	<p><b>human:</b> Hi!  <b>bot:</b> What do you like to do in your spare time? I like to run and play soccer in my spare time.  <b>human:</b> I like to work on cars in my spare time. Do you play any sports?  <b>bot:</b> No, I don't play any sports. I'm not very athletic. What kind of cars do you work on?</p>

Table C.4: Samples that are wrongly predicted as negatives before rewriting but get corrected after rewriting.