# Synthesizing Adversarial Negative Responses for Robust Response Ranking and Evaluation

Prakhar Gupta[♣] Yulia Tsvetkov[♠] Jeffrey P. Bigham[♣,♡]
[♣]Language Technologies Institute, Carnegie Mellon University
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Human-Computer Interaction Institute, Carnegie Mellon University
prakharg@cs.cmu.edu, yuliats@cs.washington.edu, jbigham@cs.cmu.edu

## Abstract

Open-domain neural dialogue models have achieved high performance in response ranking and evaluation tasks. These tasks are formulated as a binary classification of responses given in a dialogue context, and models generally learn to make predictions based on context-response content similarity. However, over-reliance on content similarity makes the models less sensitive to the presence of inconsistencies, incorrect time expressions and other factors important for response appropriateness and coherence. We propose approaches for automatically creating adversarial negative training data to help ranking and evaluation models learn features beyond content similarity. We propose mask-and-fill and keyword-guided approaches that generate negative examples for training more robust dialogue systems. These generated adversarial responses have high content similarity with the contexts but are either incoherent, inappropriate or not fluent. Our approaches are fully data-driven and can be easily incorporated in existing models and datasets. Experiments on classification, ranking and evaluation tasks across multiple datasets demonstrate that our approaches outperform strong baselines in providing informative negative examples for training dialogue systems.[1]

## 1 Introduction

Due to growing availability of dialogue corpora (Li et al., 2017; Zhang et al., 2018; Smith et al., 2020) and the advancement of neural architectures (Radford et al., 2019; Brown et al., 2020; Devlin et al., 2019), dialogue systems have achieved considerable success. As typically formulated, dialogue models generate one or more candidate responses to a provided context, consisting of past dialogue turns. Dialogue ranking (Zhou et al., 2018; Wu et al., 2019) and evaluation models (Tao et al., 2018; Yi et al., 2019; Sato et al., 2020), in turn, are deployed to select and score candidate responses according to coherence and appropriateness.

Ranking and evaluation models are generally trained using true positive responses and randomly selected negative responses, which raises two issues. First, random negative candidates often have low content similarity with the context, and thus models learn to associate response coherence and appropriateness with content similarity (Yuan et al., 2019; Whang et al., 2021; Sai et al., 2020). In real systems, generated response candidates tend to be more similar in terms of content, and so other factors (e.g., time expressions, dialogue acts, inconsistencies) tend to be more important. Second, randomly selecting candidates as negative examples in an open domain context can result in false negatives, leading to misclassification of appropriate responses.

To make dialogue models more robust to the spurious pattern of content similarity, prior work proposed to leverage adversarial and counterfactual examples (Kaushik et al., 2020; Srivastava et al., 2020). A reliable method for creating counterfactual data is to collect human-written adversarial negative responses (Sai et al., 2020), but it is expensive, time-consuming, and difficult to scale. Our goal is to create reliable automatic methods for *synthesizing* adversarial negative responses.

The most common approach to generating natural language adversarial examples is to paraphrase or insert typos, synonyms, or words relevant to the context in the inputs (Iyyer et al., 2018; Ebrahimi et al., 2018; Alzantot et al., 2018; Zhang et al., 2019). In open domain conversations, however, a context can have a wide range of possible responses with varied forms and semantics. Small lexical

[1]Code and data are publicly available https://github.com/prakharguptaz/Adv_gen_dialogue

| | Error category | Description | Sample responses |
|---|---|---|---|
| **C-ent** | Incorrect entities or actors (R,G) | Incorrect subject or object of verbs or presence of one or more incorrect entities or coreference. | *Context:* I am so happy that you are doing okay.<br>*Response:* My friend is always happy. |
| **C-time** | Incorrect Time expressions (R) | Use of incorrect time expressions or tense of verbs. | *Context:* What are you going to do on Monday?<br>*Response:* Yesterday, I celebrated my daughter's wedding anniversary. |
| **C-cont** | Contradictory or extraneous details (R,G) | Presence of details which make the response inconsistent within itself or contradict the context | *Context:* A: I don't know why I bothered to come here.<br>B: Did you enjoy your stay?<br>*Response:* I enjoyed the concert a lot. |
| **C-speaker** | Incorrect speaker turn (R) | The response is relevant to the conversation but from the wrong speaker. | *Context:* What starting salary would you expect here?<br>*Response:* If you work overtime, I will pay you extra salary. |
| **C-follow** | Does not directly address the context (R,G) | The response does not follow immediately from the context. | *Context:* What would you like for main course sir?<br>*Response:* I know very well how to make noodles, and I taught one of my friends. |
| **C-strat** | Incorrect strategies (R,G) | Use of incorrect dialogue act, emotion, persona or style | *Context:* I can't find the paper clips.<br>*Response:* Ok, great work. |
| **C-lang** | Poor language (G) | Presence of poor grammar, incorrect sentence structures or repetitions | *Context:* Do you have mixed drinks available here?<br>*Response:* Yes. This order is divided by 16 divided for main main ones of order. |

Table 1: Error categories prevalent in inappropriate responses with high context-response semantic relatedness. We present 7 categories with their descriptions and sample context and response pairs. For each category we also indicate whether it is frequently observed in Retrieval (R) or Generation (G) models. Models which simply learn to associate response coherence with content similarity often ignore these errors. Our approaches create adversarial negative data for training dialogue models by introducing such errors in context relevant utterances.

variations via substitutions and paraphrasing do not provide adequate coverage over the possible space of adversarial responses, and they can also lead to generation of false negatives due to the open-ended nature of dialogues. Creating adversarial dialogue responses is thus different, and can be more challenging than in other natural language domains.

We propose two approaches for adversarial response creation: 1) a mask-and-fill approach that corrupts gold responses related to the context but retains content similarity, and 2) a keyword-guided generative approach that uses concepts from the context to generate topically relevant but incoherent responses. These approaches do not require additional annotations, are black-box (do not need access to model parameters), and are easily adapted to new datasets and domains.

The main contributions of this paper are: 1) We identify and discuss error patterns present in retrieval and generation model outputs, which are difficult to detect due to high content similarity; 2) To the best of our knowledge, we are the first to propose automatic approaches for creating adversarial responses for dialogue model training in a black-box setting; and, 3) We demonstrate that our proposed approaches achieve better performance compared to strong baselines on two datasets on dialogue classification, ranking and evaluation tasks.

## 2 Properties of Adversarial Responses

Models trained using randomly sampled negative examples tend to assign high scores to responses with high content similarity with the context, and often ignore other important factors necessary for response appropriateness and coherence. Therefore, we aim to generate adversarial negative responses which have high content similarity with the context, but which still possess factors rendering the responses inappropriate to the context. We present the categorization of such factors or error types which can make a response inappropriate in Table 1. For each category, we provide its description and sample context-response pairs. To create this categorization, we manually analyzed responses present in outputs of generative models, candidates of retrieval sets, and human written adversarial dialogue responses (Sai et al., 2020). Categories C-ent, C-time and C-cont are errors related to various inconsistencies and logical flaws in the responses and indicate poor response *appropriateness*. Categories C-speaker, C-follow and C-strat are error types specific to the dialogue setting and indicate poor response *coherence*. Category C-lang indicates poor response *fluency*. Our categorization of errors is inspired by the categorization suggested by Pagnoni et al. (2021) for factuality of summarization, and Higashinaka et al. (2019); Ko et al.

(2019) and Sato et al. (2020) for dialogue. These categories inform our approaches as well as error analysis.

## 3 Methodology

For a given dialogue context $C$ and its gold response $R_g$, our goal is to generate an adversarial response $R_a$ such that while achieving high scores from dialogue ranking or evaluation models, it should not be a valid response to the context $C$. Dialogue ranking and evaluation models trained with such hard synthetic negative responses should learn to associate response relevance with features beyond content similarity, and hence become robust against spurious features.

The adversarial responses should satisfy the following criteria: 1) have high content similarity with input contexts; 2) have one or more errors (Table 1) which make the response inappropriate to the context; 3) be hard training examples, that is, they should likely be misclassified by current models as correct; and 4) sufficiently cover errors which occur naturally in model generated responses and retrieval candidates, and therefore they should be plausible and diverse. We propose two approaches for synthesizing adversarial negative examples - a mask-and-fill approach and a keyword-guided generation approach which we discuss next.

### 3.1 Mask-and-fill Approach

This approach modifies and corrupts original utterances related to a context as shown in Figure 1. It consists of two steps: 1) masking, where one or more tokens of an original utterance are masked out; and 2) infilling, where the masked out tokens are substituted with new tokens. For a context $C$, the set of original utterances consists of:

- Set of ground truth responses of the context - $R_g$.
- Set of utterances from the context - $U_c$.
- Set of retrieved responses based on context - $R_e$.

**Masking**: We use the hierarchical masking function from Donahue et al. (2020) which selectively masks spans at the granularities of words, n-grams, and sentences. We apply the masking function to each utterance multiple times to get up to 3 masked versions per utterance. Each utterance is constrained to have at least two masked spans. The spans are selected randomly for masking following Donahue et al. (2020).

**Infilling:** We extend the Infilling Language Model (ILM) from Donahue et al. (2020) for dialogue response infilling (Figure 1). The ILM model is a GPT-2 (Radford et al., 2019) based language model. For any piece of text $t$ with some spans masked with [blank] tokens, it is trained to predict the blanked spans in $t$ as a sequence generation problem. Each blank is infilled with an n-gram which can consist of one or more tokens. For generating adversarial responses, infilling is done by conditioning on random contexts $C_{rand}$ instead of the original context $C$ to introduce various categories of errors (Table 1). For example in Figure 1, conditioning on a random context leads to the infilling of "the marriage" in the response, introducing error of type C-ent. For the context "Did you stay your stay at our hotel?" it generates a response "I enjoyed at lot at the marriage". By corrupting the three types of utterances $R_g, U_c$ and $R_e$, this approach is able to introduce errors covering the 7 categories in Table 1.

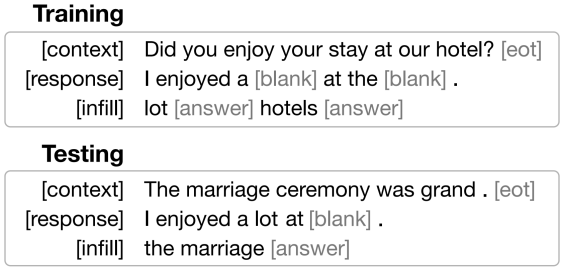


Figure 1: *Mask-and-fill* approach using ILM model. ILM is trained to infill n-grams in place of blanks in a response. Tokens after [infill] replace the [blank] tokens. During training, *Mask-and-fill* learns to infill responses conditioned on the correct context. During testing, it infills the response conditioned on a random context which introduces errors in the response.

**Preventing false negatives:** Accidentally incorporating false negatives during training can lead to the model learning to misclassify appropriate responses. However due to the open-ended nature of dialogue responses, preventing generation of false negatives is not trivial. In addition to conditioning on random contexts, we incorporate the following mechanisms during infilling to further reduce false negative generation:

- *Semantics of substitution*: We only select token substitutions which were not present in the tokens which were blanked. We also lower the generation probability of the blanked tokens' top 10 related words based on GloVe embedding (Pennington et al., 2014) similarity by a factor of 100. This ensures that the blanks are not infilled by the originally blanked tokens or any related words.
- *Degree of substitution* - To ensure that the gen-

**Training**

| | |
|---|---|
| [context] | How long did it take you to get your license? |
| [keywords] | month [sep] license |
| [response] | It took me 1 month to get the license |

**Testing**

| | |
|---|---|
| [context] | We should visit the park today. |
| [keywords] | license |
| [response] | We will bring our license and documents. |

Figure 2: Keyword-guided approach for adversarial response generation. During training, the model learns to generate a response conditioned on its keywords and the correct context. During testing, it generates the response conditioned on a random context and keywords extracted from the correct context. The generated response thus shares content with the test context but does not directly address the context.

erated negative response is sufficiently different from the original utterance, we filter out the original utterance if the number of words in the utterance after stop-word removal is less than 2. We also filter a generated response if the difference in count of non stop-words between the original and generated response is less than 2.

**Improving fluency:** The ILM model often generates responses with poor grammar or structure. To improve the fluency of the adversarial response sets, we first generate up to 4 different infilled variations of the masked original utterances, then score them using a GPT-2 based scorer named lm-scorer[2]. We then select the desired number of responses from this larger set.

### 3.2 Keyword-guided Approach

This approach generates adversarial responses using keywords from the context as guidance, as shown in Figure 2. The base generative architecture is a GPT-2 based dialogue model and it is trained to generate responses conditioned on the context and the response keywords. For adversarial response generation, the generation is conditioned on a random context $C_{rand}$ and keywords from the test context $C$. In Figure 2, for the context "How long did it take you to get your license?" it generates a response "We will bring our license and documents." To create the keyword set $K$ for a response, the model selects n number of keywords randomly from the set of all keywords extracted from the context $C$, where n is chosen randomly between 1 to 3 for every context. Keyword extraction is performed using Rake (Rose et al., 2010).

[2] https://github.com/simonepri/lm-scorer

We call this model *Key-context*. Since the generation is conditioned on keywords from context $C$, the generated response shares some content and semantics with the test context. However, since it is also conditioned on a random context $C_{rand}$, the generated response also incorporates entities, time expressions, speaker role, dialogue act, and other details based on $C_{rand}$. Since the generation model is not perfect, it also introduces errors related to fluency. Hence, the model is able to introduce errors covering the 7 categories in Table 1.

*Key-context* only uses keywords from the context to induce content similarity with the context. However, responses can have high content similarity due to the presence of similar concepts rather than just keywords. To introduce content similarity at concept level, we expand the keyword set $K$ with their top 10 most related words based on their GloVe embeddings. We use the gensim library[3] to find the most related words. For example, the related words for the keyword "christmas" are "holidays" and "easter". We replace a keyword in keyword set $K$ with one of its related words with a probability of $0.5$. We call this variant *Key-sem*.

### 3.3 Classification Model

Our classification model architecture is based on the Speaker-Aware Bert (SA-Bert) model (Gu et al., 2020). Given a dialogue context $C = \{C_1, C_2, \ldots, C_h\}$ with $C_k$ denoting $k_{th}$ utterance in the context, a response $r$ and a label $y \in \{0, 1\}$, the goal of the dialogue model $M$ is to learn a score $s(C, r)$ by minimizing cross-entropy loss function for the binary classification task. To calculate $s(C, r)$, $C$ and $r$ are concatenated, with a prepended [CLS] token. The output vector $\mathbf{E}_{[CLS]} \in \mathbb{R}^H$ for the [CLS] token is used as the aggregated representation for the context-response pair classification. The final prediction is made as $\hat{y} = softmax(\mathbf{W}\mathbf{E}_{[CLS]})$, where $\mathbf{W} \in \mathbb{R}^{2\times H}$. SA-Bert model incorporates speaker information in two ways. First, an additional speaker embedding is added to the token representations which indicates the speaker's identity for each utterance. Second, a [EOT] token is added at the end of each speaker turn. Before fine-tuning Bert model on the classification task, we first adapt Bert to the dataset by using the standard masked language model objective (Devlin et al., 2019).

[3] https://radimrehurek.com/gensim/

## 4 Experiments

We test our approaches and baselines on dialogue classification, ranking and evaluation tasks.

### 4.1 Training Details

We use the base-uncased checkpoints for BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020) from the Hugging Face transformers library (Wolf et al., 2020). We trained the models with maximum sequence length of 128, maximum number of training epochs set to 3, Adam optimizer with initial learning rate of 5e-5 with linear decay, batch size of 60 per GPU on machines with 4 Nvidia 2080Ti GPUs. For generation, we use temperature of 0.9, nucleus sampling with $p$ equal to 0.9 and minimum length of 5. We repeat each experiment three times (five times for BERT-based models) with different random seeds, use the validation split to select the best model, and report the mean metric values. Validation was done every 200 batches.

### 4.2 Experimental Setup

#### 4.2.1 Datasets

We use two open-domain dialogue datasets: *DailyDialog++* (Sai et al., 2020) and PersonaChat (Zhang et al., 2018). DailyDialog++ consists of 16900 dialogue contexts in train set, 1028 in validation set and 1142 in the test set. Each context contains 5 positive responses and 5 random negative responses. It also contains 5 adversarial responses per context collected through crowdsourcing where annotators were instructed to create negative responses with high content similarity with the context. A subset of 9259 out of the 16900 training contexts have 5 human-written adversarial negative responses. It has two test sets, adversarial test set and random test set, based on the type of the negative response. *PersonaChat* dataset (Zhang et al., 2018) is a corpus of human-human persona-conditioned conversations consisting of 8938 dialogues in the train set. We sample 2 random context-response pairs from each dialogue with a total of 17876 contexts for training. We prepend the persona utterances to the dialogue contexts in our experiments. Since there is no human-created adversarial test set available for PersonaChat dataset, we construct an artificial adversarial dataset by randomly selecting an utterance from the dialog context and inserting it in the set of candidate responses following Jia and Liang (2017) and Whang et al. (2021). The adversarial test set for each context consists of the ground truth response, one utterance selected from the dialog context, and 8 random negative responses. The random test set consists of 9 random negative responses.

#### 4.2.2 Metrics

For classification task, we report the accuracy following (Sai et al., 2020). For ranking task, we report standard ranking metrics - Recall $R_n@k$ and mean reciprocal rank (MRR). For DailyDialog++, $n$ is 6 in Recall as candidates consist of one positive response with 5 negative responses. For PersonaChat, $n$ is 10. For both classification and ranking tasks, we report results separately for the adversarial and the random test sets.

The dialogue evaluation task comprises of scoring or rating a response for its quality. For this task, we report the correlation of model scores with human provided ratings. We leverage the human ratings released by the following sources: 1) 600 ratings for response "sensibility" from (Zhao and Kawahara, 2020) with inter-rater agreement $> 0.6$ (Krippendorff's $\alpha$ (Krippendorff, 2018)). The responses consist of outputs from hierarchical recurrent encoder decoder (HRED) model with Attention (Serban et al., 2016) and Variational HRED model with attention (Serban et al., 2017); 2) 700 ratings for response quality from (Zhao et al., 2020). The responses are from 6 different generative models - Seq-2-Seq (Sutskever et al., 2014), attentional Seq-2-Seq, HRED, VHRED, GPT2-small, and GPT2-medium (Wolf et al., 2019) with greedy decoding, ancestral sampling, and nucleus sampling based decoding (Holtzman et al., 2020). The inter-rater agreement is 0.815 (Krippendorff's $\alpha$), and 3) Since the first two sources do not cover retrieval model outputs, we additionally collect quality ratings for 100 responses from a retrieval model's (Poly-Encoder (Humeau et al., 2020)) selected responses and 100 human written responses with moderate inter-annotator agreement (Cohen's Kappa 0.45 (Cohen, 1968)). All data points belong to the Dailydialog dataset and ratings are scaled between 0–1. By combining these sources we have a total of 1500 ratings for different context-response pairs.

#### 4.2.3 Baselines

We compare the following approaches of creating adversarial negative response sets.

| Model | Approach | Adversarial test set | | | Random test set | | |
|---|---|---|---|---|---|---|---|
| | | Accuracy | R@1 | MRR | Accuracy | R@1 | MRR |
| Poly-encoder | Random | - | 0.684 | 0.806 | - | **0.849** | **0.914** |
| | Mask-and-fill (Ours) | - | 0.758 | 0.856 | - | 0.821 | 0.897 |
| | Key-sem (Ours) | - | **0.788** | **0.877** | - | 0.828 | 0.902 |
| | Human | - | 0.847 | 0.913 | - | 0.831 | 0.902 |
| Electra | Random | 77.74 | 0.915 | 0.748 | 89.58 | 0.957 | **0.927** |
| | Mask-and-fill (Ours) | **87.24** | 0.945 | **0.893** | **89.61** | **0.959** | **0.927** |
| | Key-sem (Ours) | 86.24 | **0.951** | 0.881 | 89.47 | 0.957 | 0.924 |
| | Human | 91.94 | 0.984 | 0.967 | 87.95 | 0.944 | 0.911 |
| Bert | Random | 77.82 | 0.906 | 0.742 | **89.34** | **0.959** | **0.923** |
| | Semi-hard (Li et al., 2019) | 79.05 | 0.913 | 0.756 | 89.32 | 0.956 | **0.923** |
| | Token-subs (Kryscinski et al., 2020) | 77.23 | 0.901 | 0.783 | 88.60 | 0.950 | 0.906 |
| | BM25 (Karpukhin et al., 2020) | 84.42 | 0.936 | 0.872 | 87.68 | 0.948 | 0.902 |
| | Mask-and-fill (Ours) | **87.45** | **0.946** | **0.904** | 88.32 | 0.951 | 0.918 |
| | Key-context (Ours) | 86.23 | 0.939 | 0.891 | 88.16 | 0.953 | 0.922 |
| | Key-sem (Ours) | 87.02 | 0.944 | 0.897 | 89.31 | 0.954 | 0.916 |
| | Human (Sai et al., 2020) | 91.22 | 0.987 | 0.973 | 88.04 | 0.943 | 0.901 |

Table 2: Performance on classification and ranking tasks on DailyDialog++ test sets. Mask-and-fill and Key-sem approaches consistently perform the best across all model architectures compared to baselines on the Adversarial test set, just short of models trained with human created adversarial data. Poly-encoder's accuracy is not available as it ranks candidates relative to each other.

**Human** (Sai et al., 2020) Human written adversarial responses.

**Random** Responses sampled from random contexts.

**Semi-hard** (Li et al., 2019) Sampling scheme which selects samples from a batch based on their similarity scores with a margin of $\alpha$ from the positive response score. We perform static sampling and use Sentence-Bert (Reimers and Gurevych, 2019) for semantic similarity calculation with $\alpha$ set to the recommended value of 0.07.

**Token-subs** (Kryscinski et al., 2020) Training data is generated by applying a series of rule-based transformations on the positive responses. Transformations include pronoun, entity and number swapping, sentence negation and noise injection.

**BM25** Top responses returned by BM25 (Robertson and Zaragoza, 2009) based on similarity with the context. Any ground truth response is removed from this response set if present by chance. This baseline is inspired from Karpukhin et al. (2020) and Lin et al. (2020) and has shown strong performance in passage and response retrieval.

**Mask-and-fill** Our approach that infills utterances conditioned on random contexts.

**Key-context** Our approach that generates responses conditioned on test context keywords and random context history.

**Key-sem** Our approach similar to Key-context which additionally conditions on words semantically related to the keywords in the context.

For each context, adversarial train sets are created by adding 5 random negative responses to the set of 5 negative responses created from the above approaches. If an approach create more than 5 responses, we randomly select 5 from them.

For dialogue evaluation, we compare the above approaches with BLEU, METEOR (Banerjee and Lavie, 2005), embedding based metrics SkipThought (Kiros et al., 2015), Vec Extrema (Forgues et al., 2014), and RUBER (Tao et al., 2018) and BERTScore (Zhang et al., 2020a).

#### 4.2.4 Models

We experiment with following architectures for ranking and evaluation models in our experiments: 1) Bert (Devlin et al., 2019). We use the SA-Bert model (Gu et al., 2020), 2) Electra (Clark et al., 2020), pre-trained with a replaced token detection objective and employs a generator-discriminator framework, and 3) Poly-encoders (Humeau et al., 2020), allows for fast real-time inference by precomputing each candidate response representation once, and then ranking candidate responses for retrieval by attending to the context.

### 4.3 Results and Discussion

In this section, we compare the performance of our approaches with the baselines on dialogue classification, ranking and evaluation tasks.

**Performance on classification** Our proposed approaches Mask-and-fill and Key-sem achieve the highest classification accuracy on the adversarial test set (Table 2), a few percentage short of the Human baseline. The closest baseline is BM25 which has a gap of 3% in accuracy compared to our

| Approach | Adversarial test set | | Random test set | |
|---|---|---|---|---|
| | R@1 | MRR | R@1 | MRR |
| Random | 0.905 | 0.820 | 0.963 | **0.914** |
| Semi-hard | 0.906 | 0.820 | **0.964** | 0.913 |
| Token-subs | 0.895 | 0.825 | 0.958 | 0.901 |
| BM25 | 0.925 | 0.859 | 0.940 | 0.874 |
| Mask-and-fill (Ours) | **0.933** | **0.871** | 0.952 | 0.890 |
| Key-sem (Ours) | 0.920 | 0.856 | 0.947 | 0.884 |

Table 3: Performance on ranking task on PersonaChat dataset with Bert architecture. Our approaches perform better than all baselines on the adversarial test set.

| Approach | Pearson | Spearman |
|---|---|---|
| BLEU-2 | 0.046 | 0.004 |
| METEOR (Banerjee and Lavie, 2005) | 0.081 | 0.007 |
| SkipThought (Kiros et al., 2015) | 0.059 | 0.069 |
| Vec Extrema (Forgues et al., 2014) | 0.157 | 0.150 |
| BERTScore (Zhang et al., 2020a) | 0.208 | 0.198 |
| RUBER (Tao et al., 2018) | 0.253 | 0.282 |
| Random | 0.296 | 0.313 |
| Semi-hard (Li et al., 2019) | 0.299 | 0.315 |
| BM25 (Karpukhin et al., 2020) | 0.310 | 0.350 |
| Token-subs (Kryscinski et al., 2020) | 0.324 | 0.388 |
| Mask-and-fill (Ours) | 0.338 | 0.361 |
| Key-sem (Ours) | **0.382** | **0.401** |
| Human (Sai et al., 2020) | 0.348 | 0.371 |

Table 4: Comparison of approaches on dialogue evaluation. Trainable metrics are based on Bert architecture. For all entries except for the ones underlined, t-test *p-value* $< 0.05$. Mask-and-fill and Key-sem perform better than all baselines including the Human baseline.

approaches. Token-subs, which applies transformations on positive responses to corrupt them, does not fair well on this task. This indicates that simple transformations do not provide good coverage of semantic variations present in the adversarial test responses. Our approaches achieve similar performance across different model architectures, demonstrating their generalizability. Unsurprisingly, the Human baseline performs strongly as the training and test data were created in the same manner and have similar distributions. On the random test set, the performance of all approaches is either very close or lower than the Random baseline. Since the similarity between correct responses and the context is generally a lot higher than between random responses and the context in the random test set, Random baseline performs better since it associates coherence mostly with semantic similarity. Finally, our analysis shows that all baselines tend to assign low scores to valid responses which do not address a context directly. For example, for the context "Will you join us for the concert?", if the response is "It is supposed to rain this week.", models assign it a low score. Such scenarios require understanding of social and commonsense related factors. We leave addressing this limitation to future work.

**Performance on ranking** On the DailyDialog adversarial test set, Mask-and-fill and Key-sem approaches achieve the best Recall and MRR, closely followed by BM25 baselines (Table 2). The trends of the ranking metrics are similar to those observed for accuracy metrics. Our approaches perform better than the Human baseline on the random test set. On PersonaChat dataset, Mask-and-fill and Key-sem perform better than the baselines (Table 3), especially on the adversarial test set. This demonstrates the extensibility of our approach across datasets. Mask-and-fill performs better than Key-sem as the keyword sets contain a lot of keywords from the persona because of which responses have high content similarity with the persona rather than with the context. The poor performance of the Random baseline provides evidence that training models using random negative candidates does not make the models robust against hard test cases during testing. BM25 is a strong baseline for both datasets since retrieved responses also provide coverage over errors of various categories. However, retrieved response quality and diversity depends on the size of the retrieval pool. Furthermore, a stronger retrieval mechanism can lead to higher false negatives. While the variation in BM25 response sets is constraint by the size of the dataset, and they provide lesser coverage over categories C-cont, C-strat and C-lang (Table 1), our approaches have no such constraints.

**Performance on dialogue evaluation** To study the performance of various approaches on real systems, we compare them on the task of Dialogue evaluation or scoring. We measure the correlation between the scores predicted by the approaches in Table 4 with human provided ratings. Reference based metrics like BLEU-2, METEOR, SkipThought and Vec Extrema achieve very low correlations, similar to findings reported in prior art (Liu et al., 2016; Gupta et al., 2019). BERTScore and RUBER achieve moderate correlation. Our approach Key-sem achieves the best correlations, followed by Mask-and-fill. BM25's performance is lower than that of our approaches, but it is higher than the Random and Semi-hard approaches. Although Token-subs did not achieve high performance on the classification and ranking tasks, it performs well on this task. This is likely because real model outputs contains more of

| | |
|---|---|
| Context | A: Julia, will you be my wife?<br>B: I'm sorry, Steven.<br>C: Please, Julia, I have made proposal to you five times . I really want to share the rest of my life with you. |
| Random | (1) Yes of course it's a promise.<br>(2) It's better to go somewhere else.<br>(3) Let me first look at your work, how you have done it.<br>(4) Being in love is a deep experience while having a crush is shallow.<br>(5) Sometimes I don't understand, what is your problem? |
| Mask-and-fill | (1) You can't force me for to do that. They are designed for people of all ages and religions.<br>(2) There you are. I'll have to make my own lunch!<br>(3) I majored in economics. I really want i hope i can get some practical experience in life with you.<br>(4) We will go to, and to meet some of the children who are visiting at school.<br>(5) It takes time to learn. Bless you, baby! |
| Key-sem | (1) And what about the potatoes? Steven, i don't know.<br>(2) Sorry, there is no problem.<br>(3) Your wife didn't like it. Please don't tell me she is really interested in gardening.<br>(4) I really want to go inside. It's really cold outside.<br>(5) Really? I really want to pay a visit. I really want to spend the rest of my time enjoying this meal. |
| Human | (1) I want to finish my home work by five and then I am going to take rest.<br>(2) Follow these five tips, and you'll write a winning project proposal every time.<br>(3) I met my wife a three to four times before the marriage.<br>(4) Its difficult to live a life in a Dorze tribal area.<br>(5) I shared a large number of ideas with the wedding planner. |

Table 5: Sample adversarial responses from various approaches. Random responses are sampled from random dialogues. Human written responses are from the DailyDialog++ dataset. Mask-and-fill and Key-sem approaches create responses which are semantically related and yet inappropriate responses to the context.

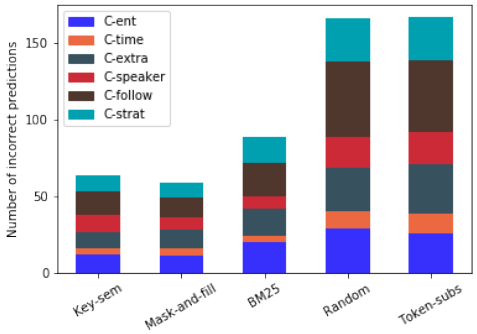


Figure 3: Analysis of error types for different approaches on DailyDialog++ predictions. C-lang error type is not present in DailyDialog++. Mask-and-fill and Key-sem achieve a more uniform distribution over error categories compared to other approaches.

the factual inconsistencies and contradictions that this approach captures, than what the adversarial test sets contain. Key-sem performs better than Mask-and-fill on evaluation since while Mask-and-fill only modifies utterances related to the context, Key-sem can freely generate more diverse adversarial responses for training. Also, Key-sem achieves higher correlation than Human baseline. This may be because it is difficult for humans to create erroneous responses with distributions similar to the ones in model generated or selected responses, especially error types like C-speaker, C-strat and C-lang. In contrast, our approaches provide good coverage over all error types.

**Analysis of errors types** We analyze the classification outputs of various approaches on the DailyDialog++ adversarial test set and report the types of misclassifications by each approach in Figure 3. We first select a subset of test data where at least one of the approaches misclassifies the adversarial response as positive. We then manually categorize the types of errors presented in Table 1 for 200 randomly selected contexts from this subset. Each response can have multiple error types. C-follow and C-extra are the dominant error types which are misclassified by baselines Random, BM25 and Token-subs. Key-sem and Mask-and-fill approaches achieve improvement in all error types compared to baselines and have a more uniform error distribution. While Key-sem performs better on C-extra, Mask-and-fill is better on C-follow and C-speaker.

**Adversarial response examples** We present sample responses from our approaches along with Random and Human baseline responses in Table 5. Random approach generates responses which are easily distinguishable from ground truth responses. Mask-and-fill approach modifies either the ground truth response, utterances from the context or BM25 retrieved responses. It modifies these utterances to introduce corruptions such as non-contextual tokens, extraneous entities, incorrect time expressions, affective words or contradictions which makes the response either inappropriate or incoherent to the context, but it remains topically

similar to the context. In Key-sem the dialogue acts, some entities and other tokens of the generated response depend on a random context the response is conditioned on, which also makes the response inappropriate or incoherent to the context.

## 5 Related Work

Dialogue response ranking and evaluation are important tasks in dialogue domain because even the recent large pretrained-language model based architectures (Zhang et al., 2020b; Humeau et al., 2020; Adiwardana et al., 2020; Roller et al., 2021; Gupta et al., 2021) have been shown to be susceptible to creating inconsistent, ungrammatical and incoherent responses (Roller et al., 2021). Traditional word-overlap based metrics like BLEU have been shown to be ineffective for dialogue response scoring (Liu et al., 2016; Gupta et al., 2019). Recently trainable metrics such as ADEM (Lowe et al., 2017), RUBER (Ghazarian et al., 2019) and USR (Mehri and Eskenazi, 2020) have been proposed for these tasks. However, since they are trained using negative samples obtained from random contexts, they are also prone to the spurious pattern of content similarity.

Adversarial or counterfactual data creation techniques have been proposed for applications such as evaluation (Gardner et al., 2020; Madaan et al., 2020), attacks (Ebrahimi et al., 2018; Wallace et al., 2019; Jin et al., 2020), explanations (Goodwin et al., 2020; Ross et al., 2020) or training models to be robust against spurious patterns and biases (Garg et al., 2019; Huang et al., 2020). Adversarial examples are crafted through operations such as adding noisy characters (Ebrahimi et al., 2018; Pruthi et al., 2019), paraphrasing (Iyyer et al., 2018), replacing with synonyms (Alzantot et al., 2018; Jin et al., 2020), rule based token-level transformations (Kryscinski et al., 2020), or inserting words relevant to the context (Zhang et al., 2019). While these approaches are optimized to change the predictions of a target model by perturbing the inputs, our approaches are more general and are not optimized towards any target model. Polyjuice (Wu et al., 2021) and FactCC (Kryscinski et al., 2020) proposed approaches for model-agnostic general-purpose counterfactual generation. These approaches change the model's prediction by creating small edits through substitutions and insertions to the inputs. They are not applicable to our setting where we aim to flip the gold label, that is, convert a valid response to an adversarial response, while the model prediction should ideally remain the same to create hard training examples. Furthermore small perturbations do not provide good coverage over the adversarial response space and can create false negative responses. Adversarial semantic collisions (Song et al., 2020) aims to generate texts that are semantically unrelated but judged as similar by NLP models to expose model vulnerabilities. However, the outputs which are unrelated to the context are not useful for adversarial training as they are easy to classify.

Finally, negative sampling strategies have also been studied for creating hard negative samples in context of visual embeddings (Faghri et al., 2018; Guo et al., 2018), knowledge graphs (Kotnis and Nastase, 2017), document retrieval (Saeidi et al., 2017; Karpukhin et al., 2020) and response retrieval (Li et al., 2019; Lin et al., 2020). In this work we compare and build upon past work and are the first to propose generative approaches for adversarial negative response creation in dialogue.

## 6 Conclusion

This paper introduces approaches for synthesizing adversarial negative responses for training more robust dialogue response ranking and evaluation models. To synthesize a rich and comprehensive set of responses, we present and analyze categories of errors which affect the models. Our proposed approaches do not require any manual annotation and achieve high performance in dialogue classification, ranking and evaluation tasks across two datasets. These results demonstrate the promise of synthetic negative examples for improving open domain dialogue. Future work, we will explore synthesizing adversarial test sets and methods for finer grained, controlled adversarial response generation.

## Acknowledgements

We thank Amy Pavel, Alissa Ostapenko, Rishabh Joshi, Artidoro Pagnoni and the anonymous reviewers for providing valuable feedback. This work was funded by the Defense Advanced Research Planning Agency (DARPA) under DARPA Grant N6600198-18908, and the National Science Foundation under Awards No. IIS1816012 and IIS2007960. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

## References


D. Adiwardana, Minh-Thang Luong, D. So, J. Hall, Noah Fiedel, R. Thoppilan, Z. Yang, Apoorv Kulshreshtha, G. Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *ArXiv*, abs/2001.09977.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019. Retrieval-guided dialogue response generation via a matching-to-generation framework. 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 1866–1875, Hong Kong, China. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

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.

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *NeurIPS, modern machine learning and natural language processing workshop*, volume 2.

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.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA. Association for Computing Machinery.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Goodwin, Koustuv Sinha, and Timothy J. O'Donnell. 2020. Probing linguistic systematicity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1958–1969, Online. Association for Computational Linguistics.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. *Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots*, page 2041–2044. Association for Computing Machinery, New York, NY, USA.

Guibing Guo, Songlin Zhai, Fajie Yuan, Yuan Liu, and Xingwei Wang. 2018. Vse-ens: Visual-semantic embeddings with efficient negative sampling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Prakhar Gupta, Jeffrey Bigham, Yulia Tsvetkov, and Amy Pavel. 2021. Controlling dialogue generation with semantic exemplars. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3018–3029, Online. Association for Computational Linguistics.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2019. Improving taxonomy of errors in chat-oriented dialogue systems. In *9th International Workshop on Spoken Dialogue System Technology*, pages 331–343. Springer.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *International Conference on Learning Representations*.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR*.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. *International Conference on Learning Representations (ICLR)*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3294–3302, Cambridge, MA, USA. MIT Press.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019. Linguistically-informed specificity and semantic plausibility for dialogue generation. 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 3456–3466, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhushan Kotnis and Vivi Nastase. 2017. Analysis of the impact of negative sampling on link prediction in knowledge graphs. *arXiv preprint arXiv:1708.06816*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Jia Li, Chongyang Tao, Wei Wu, Yansong Feng, Dongyan Zhao, and Rui Yan. 2019. Sampling mat-

ters! an empirical study of negative sampling strategies for learning of matching models in retrieval-based dialogue systems. 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 1291–1296, Hong Kong, China. Association for Computational Linguistics.

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.

Zibo Lin, Deng Cai, Yan Wang, Xiaojiang Liu, Haitao Zheng, and Shuming Shi. 2020. The world is not binary: Learning to rank with grayscale data for dialogue response selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9220–9229, Online. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

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

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2020. Generate your counterfactuals: Towards controlled counterfactual generation for text. *arXiv preprint arXiv:2012.04698*.

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

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

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.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20.

Alexis Ross, Ana Marasović, and Matthew E Peters. 2020. Explaining nlp models via minimal contrastive editing (mice). *arXiv preprint arXiv:2012.13985*.

Marzieh Saeidi, Ritwik Kulkarni, Theodosia Togia, and Michele Sama. 2017. The effect of negative sampling strategy on capturing semantic similarity in document embeddings. In *Proceedings of the 2nd Workshop on Semantic Deep Learning (SemDeep-2)*, pages 1–8, Montpellier, France. Association for Computational Linguistics.

Ananya B Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

Shiki Sato, Reina Akama, Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. 2020. Evaluating dialogue generation systems via response selection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 593–599, Online. Association for Computational Linguistics.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

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.

Congzheng Song, Alexander Rush, and Vitaly Shmatikov. 2020. Adversarial semantic collisions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4198–4210, Online. Association for Computational Linguistics.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112. Curran Associates, Inc.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Taesun Whang, Dongyub Lee, Dongsuk Oh, Chanhee Lee, Kijong Han, Dong-hun Lee, and Saebyeok Lee. 2021. Do response selection models really know what's next? utterance manipulation strategies for multi-turn response selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Automated, general-purpose counterfactual generation. *arXiv preprint arXiv:2101.00288*.

Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A sequential matching framework for multi-turn response selection in retrieval-based chatbots. *Computational Linguistics*, 45(1):163–197.

Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 65–75, Tokyo, Japan. Association for Computational Linguistics.

Chunyuan Yuan, Wei Zhou, Mingming Li, Shangwen Lv, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2019. Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. 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 111–120, Hong Kong, China. Association for Computational Linguistics.

Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating fluent adversarial examples for natural languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5569, Florence, Italy. Association for Computational Linguistics.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

Tianyu Zhao and Tatsuya Kawahara. 2020. Multi-referenced training for dialogue response generation. *arXiv preprint arXiv:2009.07117*.

Tianyu Zhao, Divesh Lala, and Tatsuya Kawahara. 2020. Designing precise and robust dialogue response evaluators. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Online. Association for Computational Linguistics.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

## A Quality of negative candidates

We perform a human evaluation experiment to test the number of false negative responses created by the different approaches. Three in-house annotators were asked to go through the set of 5 adversarial negative responses from 5 different approaches for 100 randomly selected contexts. They were instructed to report the number of responses which are appropriate responses for the context, which in this case is the number of false negatives. After annotating separately, annotators finally discussed the responses marked as appropriate and aggregated the results. We observe that Human baselines responses had 2, Random baseline had 5, Mask-and-fill had 3, Key-sem had 4 and BM25 had 10 false negative responses in the set of 500 responses (100 contexts, with 5 adversarial responses each). This shows that our approaches do not generate high number of false negatives. BM25 on the other hand leads to a relatively higher number of false negatives which can impede the learning process of the models.

## B Experiments with Masking

We experiment with two procedures for masking in the Mask-and-fill approach: 1) Random masking, which masks contiguous chunks of tokens some probability p. We leverage the masking function from (Donahue et al., 2020) which can selectively mask spans at the granularities of words, n-grams, and sentences. 2) Importance masking, which keeps the most important tokens in a response relevant to the context and masks the rest. For Importance masking, we leverage the matching model from (Cai et al., 2019) which is trained to estimate the sequence-level quality $s(q, r)$ of a response $r$ for a given query $q$. They decompose the sequence level matching score between a context and a response into a set of token-level scores as follows:

$$
\begin{aligned}
s(q,r) &= \mathbf{x}_q^T W^s \mathbf{x}_r \\
&= \mathbf{x}_q^T W^s \sum_{k=1}^{m} \omega_k \left(\mathbf{r}_k + \mathbf{e}_{r_k}\right) \\
&= \sum_{k=1}^{m} \omega_k \mathbf{x}_q^T W^s \left(\mathbf{r}_k + \mathbf{e}_{r_k}\right) \quad = \sum_{k=1}^{m} \omega_k s_k
\end{aligned}
$$

where $s_k = \mathbf{x}_q^T W^s \left(\mathbf{r}_k + \mathbf{e}_{r_k}\right)$, and $x_r$ is the weighted sum of a Bert Transformer encoder outputs $r_k$ as well as their initial vector representations $e_k$. The importance of each response token $k$ to the context is estimated by $s_k$. We mask out any token with importance weight $\omega_k$ less than the average $\omega$ and only retain tokens highly relevant to the context following Cai et al. (2019). In our initial experiments we found that the Importance masking procedure lead to worse performance than Random masking. The adversarial test set accuracy on DailyDialog adversarial test set was 85.43% compared to the 87.45% accuracy using Random masking. Our analyses showed that Importance masking masked out about 50% of the response tokens, and the infills generated by the ILM model were mostly poor in fluency as the number of masked tokens was high. We therefore finally used Random masking for Mask-and-fill.

## C Sample Model Generated Responses

In continuation of sample responses presented in Table 5 of the main paper, we present some more sample responses from different approaches along with Random and Human baseline responses in Table 6.

## D Additional Implementation Details

For BM25 approach, we use the open source implementation from transformer rankers[4]. The DailyDialog++ dataset contains 16900 dialogue contexts but only 9259 of those have adversarial negative responses for the Human baseline. For the results reported in Table 4, all approaches from Random and below use the Bert architecture and trained using DailyDialog domain data. Additionally, RUBER is also trained on the DailyDialog++ dataset. The approaches above Random in the table do not require training. Each approach predicts a score for the set of 1500 responses created using a set of generative and retrieval models as detailed in section 4.2.2. Sentence-Bert used in Semi-hard sampling scheme is fine-tuned on the datasets used in this paper.

For the Mask-and-fill approach, the model takes in the following sequence of inputs: {[context] $C_1$ [eot],.., [eot] $C_h$ [response] $r$-with-blanks [infill] $B_1$ [answer],.., $B_l$ [answer]}, where $C_{c=1}^{h}$ represents a context with h utterances, $r$ the response and $B_{b=1}^{l}$ are the tokens blanked in the response. [eot] is used to indicate end of turn. To generate a set of 5 adversarial responses in the Mask-and-fill approach, we first create 4 masked versions

[4]https://github.com/Guzpenha/transformer_rankers

| Context or Model | Utterances |
|---|---|
| Context | A: OK . Now I'll put the dishes in the sink.<br>B: Thank you . I'll wash them.<br>C: OK . Then I will clean the table. |
| Random | (1) May I have your name and number, Sir?<br>(2) Then I hope to get the other documents by end of the day.<br>(3) She is very experienced in that area, including all the years in college<br>(4) I am in Computer Science department.<br>(5) Sure, you can talk to him. |
| Mask-and-fill | (1) Ok. Thank you. But, what are you going to do with him?<br>(2) Uh, it's time to wake up. I will clean it up and then go to bed.<br>(3) Oh, thank you. You have helped a lot.<br>(4) Don't, thank you very much for saying it like that. Only in your opinion do you have to wear proper clothes.<br>(5) Yes, thank you! But, i am not satisfied with the work you've done. |
| Key-sem | (1) Sorry, he didn't put the dishes on the table.<br>(2) Ok. We'll clean up the room now. I can walk you through the process.<br>(3) Don't forget to wash the dishes and put away the clothes.<br>(4) In my field, i put on quite a few weight -bearing exercise in order to improve my lung capacity.<br>(5) Thank you for your understanding. What are your recipes for tableware? |
| Human | (1) I just now saw the news that the boat was sinking due to heavy goods.<br>(2) I want to thank my friend because he helped me to wash my dress at school camp.<br>(3) Nowadays, table fans are getting very cheap online.<br>(4) I know that using a facial scrub can make your skin look beautiful, clean and soft.<br>(5) I gifted a sink to my friend for his house warming ceremony. |
| Context | A: Can you tell me what's my responsibility in this position?<br>B: Yes, of course . You would be responsible for the development of software products.<br>C: I see . This is my advantage. |
| Random | (1) Okay! That sounds great to me.<br>(2) Well! How much will it cost per kg?<br>(3) Well! You can pay it on monthly or yearly basis, it is upto you.<br>(4) I usually spend those days with my family and it is quite fun you see.<br>(5) What type of games do you like to play? |
| Mask-and-fill | (1) Yes. Maybe he is just looking for some publicity. You are responsible, too.<br>(2) I see. Then we will all get on our own.<br>(3) That's nice. And i would be willing to take them for that.<br>(4) You also have to work on the meetings to be more focused. I need to add some training.<br>(5) What kind of software do they use now? |
| Key-sem | (1) Let me see, in your brochure, what kind of promotion you're promising?<br>(2) Tell me about it. What do you think? Will you marry her?<br>(3) Of course. Of course there are many things online. Tell me about it.<br>(4) Yes, i appreciate your cooperation. The development of the l / c is our utmost priority.<br>(5) Thank you. I do want to get him a diamond ring. He's responsible for development of the etv. |
| Human | (1) Of course, the museum is in the closing stage because of financial issues.<br>(2) I was searching on some websites for the junior engineer position to develop my knowledge in the hardware field.<br>(3) I see, is there any terms and condition that I have to sign for this position in your company?<br>(4) Of course, you must provide me the full details about our company's financial position by today evening.<br>(5) Of course, My friend is very much interested to work in a software company. Can you give him a chance in your company? |

Table 6: Outputs from different approaches for negative response set creation. Random responses are unrelated to the contexts. Mask-and-fill and Key-sem approaches create responses which are highly similar to the content of the contexts, and hence the model needs to learn factors important for response coherence and appropriateness such as presence of correct entities, time expressions, strategies and others.

of every utterance related to the context ($R_g, U_c$ and $R_e$). ILM model then generates 4 infills per masked utterance. Thus each utterance gets 16 different modified versions. All these modified utterances are then ranked using the lm-scorer library and we select the top 5. BM25 similarity is used to create the retrieved response set.

For the Keyword-guided approaches, the model is given as input the context $C$, keywords from the ground truth response $K$, and the ground truth response $r$ as shown in Figure 2. Specifically, the model takes in the following sequence of inputs - {[context] $C_1$ [eot],.., [eot] $C_h$ [keywords] $K_1$ [sep],..,[sep] $K_n$ [response] $r$}. For both ap-

proaches during training, positive responses and negative responses are interleaved, i.e. each positive response is followed by one random and one adversarial response.