PNEG: Prompt-based Negative Response Generation for Dialogue Response Selection Task

Nyoungwoo Lee[1], **ChaeHun Park**[2], **Ho-Jin Choi**[3], and **Jaegul Choo**[2]

[1]Scatter Lab [2]KAIST AI [3]KAIST

nyoungwoo@scatterlab.co.kr, {ddehun, hojinc, jchoo}@kaist.ac.kr

Abstract

In retrieval-based dialogue systems, a response selection model acts as a ranker to select the most appropriate response among several candidates. However, such selection models tend to rely on context-response content similarity, which makes models vulnerable to adversarial responses that are semantically similar but not relevant to the dialogue context. Recent studies have shown that leveraging these adversarial responses as negative training samples is useful for improving the discriminating power of the selection model. Nevertheless, collecting human-written adversarial responses is expensive, and existing synthesizing methods often have limited scalability. To overcome these limitations, this paper proposes a simple but efficient method for generating adversarial negative responses leveraging a large-scale language model. Experimental results on dialogue selection tasks show that our method outperforms other methods of synthesizing adversarial negative responses. These results suggest that our method can be an effective alternative to human annotators in generating adversarial responses. Our dataset and generation code is available at https://github.com/leenw23/generating-negatives-by-gpt3.

1 Introduction

In retrieval-based dialogue systems, the response selection model aims to predict the most appropriate response among multiple candidates retrieved for a given conversation context (Zhou et al., 2018; Wu et al., 2019; Chen et al., 2022). The selection model is generally trained to distinguish a related (positive) response from randomly sampled negative responses on training datasets, but such a model generally poses the following problems.

First, randomly selected negatives are often too easy to distinguish because they are totally irrelevant to the dialogue context (Li et al., 2019; Lin et al., 2020). In this case, the model is more likely to predict the response only by relying on the superficial content similarity of the context-response pairs (Yuan et al., 2019; Sai et al., 2020; Whang et al., 2021). These models are vulnerable to adversarial responses with high content similarity to the dialogue context, and fail to distinguish subtle differences in various contexts in real-world scenarios. Second, random sampling causes sampling bias, such as containing false negatives for a given dialogue context (Zhou et al., 2022). These biases inherent in the training datasets hinder the accurate prediction of the selection model, resulting in performance degradation.

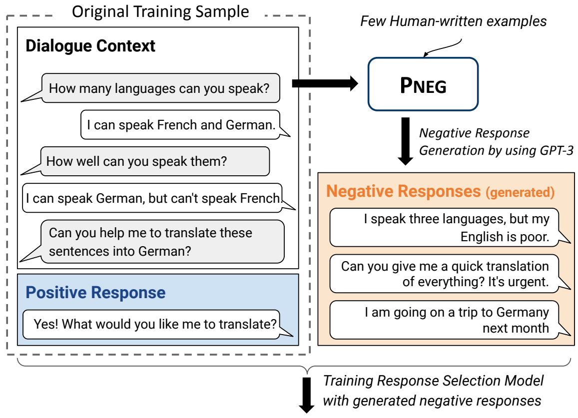

Figure 1: A conceptual pipeline of prompt-based negative response generation.

To mitigate this problem, recent studies have proposed various methods to synthesize and leverage adversarial negative training samples so that the selection model can learn features beyond content similarity (Srivastava et al., 2020; Kaushik et al., 2021). However, existing methods for synthesizing adversarial negative responses (Ebrahimi et al., 2018; Alzantot et al., 2018; Zhang et al., 2019; Qiu et al., 2021; Gupta et al., 2021) still have limitations in creating human-like responses. The most reliable method is to collect human-written adversarial negatives (Sai et al., 2020), but it is not scalable because it is expensive and time consuming.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 10692–10703
December 7-11, 2022 ©2022 Association for Computational Linguistics

To overcome these limitations, we note that large-scale language models such as GPT3 can be utilized as a low-cost data labeler (Wang et al., 2021). In this paper, we present PNEG, a **P**rompt-based **NE**gative response **G**eneration method leveraging a large-scale language model (Figure 1). In PNEG, we can cheaply collect human-like adversarial negatives by using an in-context learning-based data augmentation method with human-written samples as demonstration examples.

Experimental results on the dialogue response selection task show that the selection model that trained the negative samples from PNEG has better performance than other baselines. We then conduct quality evaluation and ablation studies to analyze the validity of PNEG. Consequently, we argue that PNEG can be an efficient alternative to human annotators in generating adversarial responses.

Our contributions are as follows:

- We propose PNEG, a **P**rompt-based **NE**gative response **G**eneration method.
- Our method can generate high-quality adversarial negative responses only with a few human-written examples.
- We show that our method outperforms other baselines across multiple model architectures on the response selection task.

2 PNEG: Prompt-based NEgative response Generation

Large-scale language models such as GPT-3 (Brown et al., 2020) can augment fluent text training samples using natural language prompts and in-context examples (Yoo et al., 2021; Schick and Schütze, 2021; Bae et al., 2022; Liu et al., 2022). By extending such direction to the response selection task, we propose PNEG, a **P**rompt-based **NE**gative response **G**eneration method for robust response selection models. Our method consists of three steps: (1) selecting demonstration examples, (2) constructing a prompt containing examples and target dialogue context, and (3) generating adversarial negative responses by inputting the prompt into GPT-3. The generated negative responses are used as training samples for response selection task.

2.1 Step 1: Example Selection

We first sample a total k demonstration examples from an example set E to construct a prompt for in-context learning of GPT-3 (A in Figure 2). We consider *DailyDialog++* (Sai et al., 2020) as an example set to generate training samples that have similar patterns as human-written high-quality linguistic patterns. The example set consists of a dialogue context, a positive response, and multiple human-written negative responses. We uniformly sampled examples from the dialogue dataset. The context and human-written negative responses are used in the following prompt construction step.

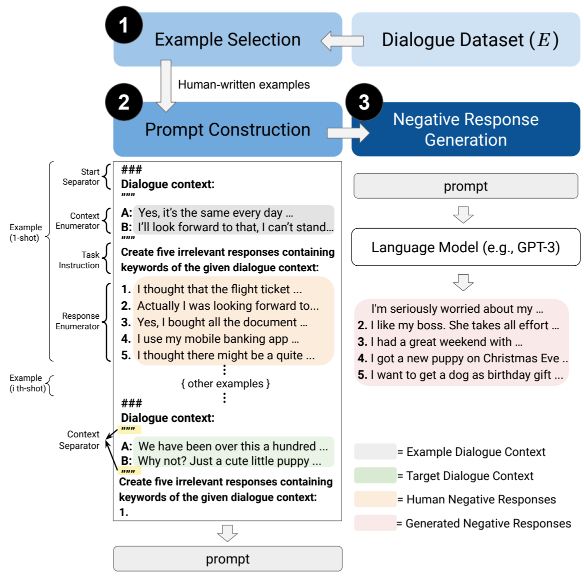

Figure 2: Overall pipeline of PNEG.

2.2 Step 2: Prompt Construction

Prompt Design We first design a prompt to use as an input to GPT-3. Large-scale language models are generally pre-trained to generate responses with high coherence for a given context, but our method aims to generate adversarial responses with high content similarity but low coherence. In particular, natural language prompt-based reasoning is sensitive to prompt changes (Jin et al., 2022), so we have to carefully design prompts to generate accurate negative responses. We devise several prompts to evaluate generative quality, inspired by related works (Yoo et al., 2021; Schick and Schütze, 2021; Min et al., 2022). We evaluate the devised prompts by the response selection task, and then select the finest prompt template based on the results of Appendix B.

Prompt Construction Our prompt is constructed with the chosen prompt template, k number of examples, and the target dialogue context that we aim to generate multiple negative responses. The prompt template consists of three components to clarify the role of each example and target context: (1) a task instruction I written in natural language, (2) an enumerator to receive each utterance from examples and the target context, and (3) a separator to separate each example or dialogue context in the prompt. The details of each component in the prompt template are shown in Figure 2.

2.3 Step 3: Negative Response Generation

Now, GPT-3 can generates augmented negative responses following our input prompt (C in Figure 2). The examples within the prompt encourage the language model to generate negative responses of similar patterns to the human-written negative responses. The task instruction directly guides the model to understand the target task and the relationship between a dialogue context and corresponding negative responses in the examples.

3 Experimental Setup

3.1 Dialogue Response Selection Task

We evaluate our method on the dialogue response selection task. We train the selection model with 11 response candidates consisting of 1 positive response, 5 random and 5 adversarial negative responses per context. We report the R@1 and mean reciprocal rank (MRR) score. For evaluation, we use random and adversarial test datasets which consist of the 6 candidates for each context.

3.2 Datasets

DailyDialog++ We use the *DailyDialog++* (Sai et al., 2020) dataset for our overall experiments. This dataset consists of 16,900, 1028, and 1142 dialogue contexts in training, validation, and test datasets, respectively. Since only the subset of 9259 contexts in the training dataset contains adversarial responses, we use them as our training dataset. Each context has five adversarially curated negative responses written by human annotators.

PersonaChat We also use the *PersonaChat* dataset (Zhang et al., 2018) on the response selection task. This dataset consists of 8938, 1000, and 968 dialogue conversations in training, validation, and test datasets, respectively. We use 8938 contexts for training, and concatenate the persona sentences in front of the context. Since there are no human-written adversarial negative responses in this dataset, we create an adversarial test dataset by sampling one response from the context and including it in the candidate responses following Gupta et al. (2021) and Whang et al. (2021).

3.3 Baselines

We compare our method with following baselines.

Random Randomly sampled responses.

Human (Sai et al., 2020) Human-written adversarial responses in *DailyDialog++* dataset.

BM25 (Karpukhin et al., 2020) Retrieved responses from BM25 (Robertson and Zaragoza, 2009).

Semi-hard (Li et al., 2019) Retrieved responses from training dataset based on their similarity between positive response with a margin of α. We perform a static sampling using sentence-BERT (Reimers and Gurevych, 2019) with α as 0.07 following Gupta et al. (2021).

Mask-and-fill (Gupta et al., 2021) This method first randomly masks the words in a answer response, and infill them using masked language modeling conditioned on a random context.

Key-sem (Gupta et al., 2021) This method generates new responses conditioned on keywords in the context using GPT-2 (Radford et al., 2019).

PNEG (Ours) GPT-3 generated adversarial negative responses by using our method, PNEG.

3.4 Models

We train dialogue selection models with different negative responses described in §3.3. The models are based on cross-encoder architecture, and three different pre-trained language models are used in experiments: 1) BERT (Devlin et al., 2019), 2) RoBERTa (Liu et al., 2019), and 3) ELECTRA (Clark et al., 2019). The implementation details are provided in Appendix A.

4 Experiments

4.1 Performance on Response Selection Task

We compare our method with the baselines for the response selection task[1] (Table 1). On the adversarial test dataset, we notice that PNEG consistently outperforms other baselines across different model architectures. The PNEG especially shows the most similar performance to the human baseline, which

[1]The negative responses samples generated by each method are provided in Table 10.

suggests that our method can be an effective alternative to human annotators for collecting adversarial negative samples. In the random test dataset, Semi-hard performs better than other baselines including PNEG and even the human baseline. This tendency indicates that the robustness of the models to the adversarial test dataset does not always lead to the random test dataset. We speculate that these results are due to data distribution shifts according to different negative response sampling strategies (Penha and Hauff, 2021). Retrieval-based methods like BM25 and Semi-hard would suffer from finding appropriate negative responses when the dialogue corpus for response retrieval is limited.

We also compare our method with the baselines for the response selection task on the *PersonaChat* (Table 2). Although PNEG generates negative responses using human-written examples from *DailyDialog++*, it shows better performance than other baselines in the adversarial test dataset. Such results prove the scalability of PNEG across multiple dialogue datasets, because our method can automatically generate adversarial negative samples with only a few human-written samples. It means that our method can easily transfer to other domain conversations only with minimal human effect. Since we used fewer samples for *PersonaChat* experiments, the response selection models overall show lower performance than the *DailyDialog++* experiment.

4.2 Synthetic Dataset Quality

We evaluate the quality of adversarial responses with predictive scores of the response selection model and context-response similarity model[2]. We assume that the higher the prediction score of the selection model for the adversarial negative response, the more effective it is for the robust training of the model. The evaluation results are shown in Figure 3. In both models, the prediction score for negative responses generated by PNEG is higher on average than other responses. The difference is more significant in the score of the selection model, suggesting that PNEG can produce more effective adversarial responses that are confused with the positive response. Although Semi-hard samples negative responses using similarity scores from Sentence-BERT, the responses have lower scores than other methods because the sampling pool is limited. We additionally provide human evaluation results for synthetic quality in Appendix D.

[2]The experiment details are provided in Appendix C.

Model	Method	Test Dataset			
		Random		Adversarial	
		R@1	MRR	R@1	MRR
BERT	Random	0.865	0.923	0.674	0.806
	BM25	0.845	0.911	<u>0.857</u>	0.915
	Semi-hard	**0.881**	**0.934**	0.672	0.804
	Key-sem	0.864	0.923	0.842	0.909
	Mask&fill	<u>0.869</u>	<u>0.926</u>	0.856	<u>0.916</u>
	PNEG	0.867	0.924	**0.937**	**0.964**
	Human	0.870	0.926	0.954	0.974
RoBERTa	Random	0.879	0.932	0.658	0.797
	BM25	0.879	0.932	0.865	0.920
	Semi-hard	**0.892**	**0.937**	0.660	0.797
	Key-sem	<u>0.889</u>	**0.937**	<u>0.868</u>	<u>0.924</u>
	Mask&fill	0.873	0.927	<u>0.868</u>	0.922
	PNEG	0.882	<u>0.933</u>	**0.942**	**0.967**
	Human	0.891	0.938	0.955	0.975
ELECTRA	Random	0.893	<u>0.941</u>	0.705	0.823
	BM25	0.853	0.916	<u>0.900</u>	<u>0.940</u>
	Semi-hard	**0.908**	**0.949**	0.730	0.840
	Key-sem	<u>0.895</u>	0.940	0.869	0.929
	Mask&fill	<u>0.895</u>	<u>0.941</u>	0.877	0.923
	PNEG	0.873	0.928	**0.951**	**0.972**
	Human	0.896	0.941	0.967	0.982

Table 1: Performance in the dialogue response selection task on Random and Adversarial test datasets based on the *DailyDialog++*. Among the methods except for human baseline, the best result is shown in **bold**, and the second-highest result is <u>underlined</u>.

Approach	Test Set		Mean
	Random	Adversarial	Rand + Adv.
Random	**0.815**	0.316	0.566
Semi-hard	<u>0.814</u>	0.338	0.576
BM25	0.718	<u>0.637</u>	<u>0.678</u>
PNEG (Ours)	0.774	**0.684**	**0.729**

Table 2: Performance in the dialogue response selection task on Random and Adversarial test datasets based on the *PersonaChat*. Among the methods except for human baseline, the best result is shown in **bold**, and the second-highest result is <u>underlined</u>.

4.3 Varying the Size of Example Set

We study the effect of the size of the example set E containing examples for prompt construction on the performance of the response selection task. As shown in Table 3, even if the size of E becomes extremely small (e.g., 0.1%), the performance on the adversarial test dataset hardly decreases. These results show that our method can efficiently generate high-quality adversarial responses by collecting only a few demo examples. To increase the diversity of examples, we further try +REUSE, which continuously adds the negative responses generated by PNEG to E. However, the 0.1%+REUSE has a

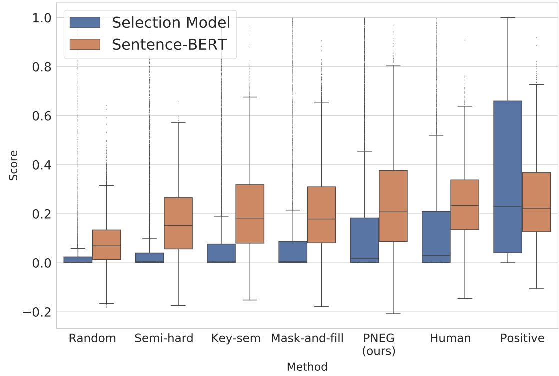

Figure 3: Box plot of prediction scores (blue) and similarity score (orange) for each type of response. The prediction scores are linearly normalized into the [0,1].

Sub. (%)	$\|E\|$	**Test Set (R@1)**		**Mean**
		Random	Adv	Rand + Adv
0.1 +REUSE	$9{+}\alpha$	0.852	0.899	0.876
0.1	9	0.843	0.938	0.891
1	93	0.845	0.936	0.891
10	926	0.852	0.936	0.894
100 (PNEG)	9259	0.877	0.941	0.909

Table 3: Performance on the size changes of E containing examples used to construct prompts of our method.

significant performance drop in the adversarial test dataset. These results shows that example quality is more important than the diversity of examples to optimize the quality of the generated negative responses.

5 Conclusion

In this paper, we present PNEG, a prompt-based adversarial negative response generation method for response selection task. PNEG can collect high-quality adversarial negative responses at a lower cost than human annotators. Our experiments on dialogue response selection tasks show that negative responses generated by PNEG can improve the discriminating power of the selection models. In future work, we are interested in designing more confused adversarial responses and how to effectively utilize them for training the selection models.

Limitations

In this section, we present two major limitations of our work. First, PNEG has not yet been tested on a larger dataset. We expect our method can produce effective adversarial training samples in large-scale datasets at a reasonable cost. However, since the performance of sampling-based methods can improve on sufficiently large training datasets, it is necessary to compare such methods with PNEG. Second, there is a limitation to leveraging the adversarial negative response for training the selection model in the dialogue response selection task. The selection model that learns the adversarial negative response improves performance in the adversarial test dataset with the same distribution of candidates, but decreases performance in the random test dataset. This is not the result we were hoping for, and we need a close analysis to solve this phenomenon.

Ethics Statement

We clarified compliance with the ACL Ethics Policy in this study. In particular, in the human evaluation experiment, we recruited participants fairly, designed experiments that require minimal effort, and provided adequate rewards to participants.

Acknowledgements

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques, No. 2013-2-00131, Development of Knowledge Evolutionary WiseQA Platform Technology for Human Knowledge Augmented Services, and No. 2019-0-00075, Artificial Intelligence Graduate School Program(KAIST))

References

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.

Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee, and Woomyoung Park. 2022. Building a role specified open-domain dialogue system leveraging large-scale language models. *arXiv preprint arXiv:2205.00176*.

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.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Wei Chen, Yeyun Gong, Can Xu, Huang Hu, Bolun Yao, Zhongyu Wei, Zhihao Fan, Xiaowu Hu, Bartuer Zhou, Biao Cheng, Daxin Jiang, and Nan Duan. 2022. Contextual fine-to-coarse distillation for coarse-grained response selection in open-domain conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4865–4877, Dublin, Ireland. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

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.

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.

Prakhar Gupta, Yulia Tsvetkov, and Jeffrey Bigham. 2021. Synthesizing adversarial negative responses for robust response ranking and evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3867–3883, Online. Association for Computational Linguistics.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. 2022. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2763–2775, Dublin, Ireland. Association for Computational Linguistics.

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, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. 2021. Explaining the efficacy of counterfactually augmented data. *International Conference on Learning Representations (ICLR)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*.

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

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.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. 2019. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Gustavo Penha and Claudia Hauff. 2021. On the calibration and uncertainty of neural learning to rank models for conversational search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 160–170.

Yao Qiu, Jinchao Zhang, Huiying Ren, and Jie Zhou. 2021. Challenging instances are worth learning: Generating valuable negative samples for response selection training. *arXiv preprint arXiv:2109.06538*.

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 Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

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.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. 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.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, 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*, volume 35, pages 14041–14049.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2018. Transfertransfo: A transfer learning approach for neural network based conversational agents. In *NeurIPS Workshop on Conversational AI*.

Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. 2018. Learning matching models with weak supervision for response selection in retrieval-based chatbots. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 420–425, Melbourne, Australia. Association for Computational Linguistics.

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.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. 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.

Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. 2022. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130, Dublin, Ireland. 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.

k	Test Set		Mean
	Random	Adv	Rand + Adv
0	0.799	0.841	0.820
1	0.856	0.893	0.875
2 (PNEG)	0.859	0.928	0.894

Table 4: Ablation study on the number of examples k in the prompts of our method. (k = 0, 1, and 2)

A Implementation Details

The inference on GPT-3 was carried out via the Open AI API Beta Access program. We used the largest GPT-3 model, *davinci*. It takes an average of $0.03 for generating negative responses for one sample of the dialogue dataset, which is approximately one-tenth of the cost of collecting through humans. Each inference consumes an average of 600 tokens and takes an average of 4.28 seconds. For the balance between diversity and quality of synthetic samples from our method, PNEG, we set the temperature to 0.8 and both frequency penalty and presence penalty to 0.4.

For training of selection models, we predict the score of each context-response pair for every responses in a candidate responses and use cross entropy loss to maximize the score of the context-positive response pair. We use the pre-trained language models[3] released by Wolf et al. (2018) for experiments. We use the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate as 2e-5, and set the maximum epochs to 3. We use the validation loss to select the best model. The random seed is fixed, and the batch size is set to 16 per GPU on machines with 2 Nvidia TITAN RTX GPUs. We repeated the experiments three times with different random seeds and report the average performance.

B Performance on Prompt Changes

In this section, we study the performance according to the prompt changes of PNEG on the response selection task.

B.1 Number of Examples (k)

We analyze the effect of the number of examples k in the prompts of our method on the response selection model. The results are in Table 4. Our method has the highest performance when using two examples (k=2), but using one example (k=1) also can be a reasonable alternative. The performance of prompts without examples (k=0) is rapidly degraded due to frequent occurrence or false-negative generation. These results show that it is important to provide an adequate number of examples to minimize the occurrence of false-negative responses.

We also measured the Jaccard similarity and length correlation between generated responses and each example in the prompt to qualitatively analyze the effect of the example on the generated responses. As shown in Table 5, the Jaccard similarity and length correlation coefficient are measured higher when k=1 than when k=2, and the generated responses are more affected by the closer example. Such contamination effect can increase the effectiveness of the in-context example as guidance of the task, but it can also limit the diversity.

[3]`bert-base-uncased`, `roberta-base` and `google/electra-base-discriminator` are used.

k	position (pos/k)	Jaccard Similarity	Length Correlation	
			Pearson	Spearman
1	1/1	**0.046**	**0.376**	**0.351**
2	1/2	0.031	0.154	0.174
2	2/2	0.035	0.339	0.293
2	all	0.041	0.342	0.324

Table 5: Correlation of generated negative responses in our method with the few-shot examples (k>0). We measure the Jacquard similarity and length correlation between the example and the generated response.

Type	k	Test Set		Mean
		Random	Adv	Rand + Adv
I_dir	2	0.877	0.941	0.909
I_pos	2	0.857	0.940	0.898
I_imp	0	0.788	0.800	0.796

Table 6: Ablation studies on task instruction changes in the prompt of PNEG. The I_{pos} and the I_{pneg} are follows 2-shot setting, and the I_{imp} follows zero-shot setting.

B.2 Task Instruction Type (I)

We compare the performance of PNEG according to changes in the task instruction. We design the following three types: (1) direct instruction (I_{dir}), (2) direct instruction with a positive response (I_{pos}), and (3) implicit instruction (I_{imp}). We expect that I_{pos} can generate more challenging negatives by referring to the positive response, and I_{imp} can generate diverse responses due to the reduced constraints in the prompt. As shown in Table 6, I_{pos} show lower performance than I_{dir} in the random test dataset. Since I_{imp} is vulnerable to false-negative generation, it has the lowest performance in both random and adversarial test datasets.

Approach	Pred. Score	Similarity
Random	$-2.749_{2.48}$	$0.078_{0.09}$
Semi-hard	$-2.051_{2.83}$	$0.161_{0.13}$
Mask-and-fill	$-1.925_{3.26}$	$0.207_{0.17}$
Key-sem	$-1.956_{3.34}$	$0.212_{0.17}$
PNEG (Ours)	$-0.598_{3.53}$	$0.241_{0.20}$
Human	$-0.279_{3.13}$	$0.242_{0.15}$
Positive	$2.779_{2.25}$	$0.256_{0.17}$

Table 7: Automatic evaluation results for response quality. **Pred. Score** and **Similarity** indicate the predicted score of each response by selection model and the similarity score between each response and the context measured by Sentence-BERT, respectively. The mean and standard deviation of each score are reported in the $mean_{std.}$ format.

Approach	Rand neg	Hard neg	False neg
Mask-and-fill	56.6%	41.0%	2.4%
PNEG (Ours)	43.6%	52.2%	4.2%
Human	47.4%	51.2%	1.4%

Table 8: Human evaluation results to verify the quality of synthetic adversarial negative responses.

C Automatic Evaluation Results

Table 7 shows statistics on the scores of each model for automatic evaluation in §4.2. For the evaluation, we first divide the training dataset of *DailyDialog*++ by 8:2 and use it as a training and test dataset, respectively. Then we train the selection model using BERT with randomly sampled negatives. For the context-response similarity model, We use a pre-trained Sentence-BERT. Among the negative responses, human-written responses and our responses usually get the high predictive score than other negative responses. In terms of similarity score, our negative responses show high similarity with dialogue contexts. We speculate that the higher similarity of our responses with the dialogue contexts can improve the robustness of response selection models models by encouraging them to learn the features beyond superficial context-response similarity.

D Human Evaluation

For human evaluation, we randomly sampled 100 data consisting of a dialogue context and 5 negative responses from three different method (Mask-and-fill, PNEG, and Human). Each response is evaluated by three human annotators. We recruited a total of 9 human annotators (6 males and 3 females) for the human evaluation. Human annotators classify the type of each negative response as random, hard, and false negative according to the review criteria described in the *DailyDialog*++. The type of each data is basically determined by a majority, and if the evaluation result is a tie, such data is determined to be a random negative type. Table 8 shows the human evaluation results. Our PNEG has a slightly higher false negative ratio than Mask-and-fill, but shows the highest hard negative ratio. In future work, we may consider soft labeling (Wu et al., 2018; Chen et al., 2020) or label smoothing (Müller et al., 2019) techniques to alleviate this problem.

Aug.	Dataset num	Test Set Random	Test Set Adv	Mean Rand + Adv
PNEG	9259	0.877	0.941	0.909
+ 5000	14259	**0.889**	0.946	0.917
+ 10000	19259	0.886	**0.950**	**0.918**
+ 15000	24259	0.871	0.937	0.904
+ 20000	29259	0.877	0.947	0.912

Table 9: Performance on our method with data augmentation techniques on additional 5,000, 10,000, 15,000, and 20,000 augmented dataset in the dialogue response selection task.

E Data Augmentation

We conduct data augmentation experiments by synthesizing adversarial negative responses to the additional datasets. For the experiment, we use the dialogue contexts in the original *DailyDialog* dataset that has no duplication with the contexts in *DailyDialog*++. The results are shown in Table 9. Data augmentation using our method generally leads to improved performance. However, if the training dataset is already large enough, the model can properly generalize it (Wei and Zou, 2019). In our experiments, the performance of the selection model is saturated, if the dataset is augmented by more than 10,000 (about 100%).

F Frequent Error Types in GPT-3 Generation

During our experiments, we often observed the weird generation results of GPT-3. The frequent error types in generated results of GPT-3 can be roughly categorized as follows: (1) n-gram or word repetition, (2) containing too many "__" or "_ _", (3) out of numbering rules. We generate negative responses with GPT-3 for the given context until there is no error response that is aforementioned. Note that false negative is a semantic error type that needs to be evaluated by humans.

G Prompt Construction used in PNEG

The PNEG prompt is as follows:

```
###
Dialogue context:
"""
A: How about taking the damaged portion at a lower price?
B: What kind of price did you want?
A: I was thinking of 30% off.
"""
Create five irrelevant responses containing keywords of the given dialogue context:
1. I have not completed the portions of the children, ...
2. Shall I inquire about the price of the plane tickets ...
3. I have been thinking up new ways of supplying money ...
4. My car roof was not damaged in the accident.
5. I purchased a different kind of dress in the shopping mall ...
###
Dialogue context:
"""
A: No, but that was a random change of subject.
B: It may have been random, but have you?
A: I haven't lately.
"""
Create five irrelevant responses containing keywords of the given dialogue context:
1. Yeah, Our society is annoying. They keep on changing ...
2. I am not sure which subject I am going to take. Lately, ...
3. I don't know that day Prof. Lesley was randomly picking up ...
4. Today In college some random guy came and started talking ...
5. Have you seen Tina lately? I am feeling weird as ...
###
Dialogue context:
"""
A: Paul, a company called me for an interview.
B: That's great! You need to prepare for it.
A: How?
"""
Create five irrelevant responses containing keywords of the given dialogue context:
1.
```

Then, GPT-3 generate following negative responses:

```
I don't know how to handle an interview.
2. How could I prepare for the interview?
3. To be honest man, I don't know what to say.
4. You are probably thinking that I could not do well in the interview, right?
5. Did you remember to register for the interview?
```

This prompt is particularly sensitive to the task instruction. For example, GPT-3 generates emotionally negative responses when the word *"irrelevant"* in the task description is changed to *"negative"*. In addition, if we remove *"containing keywords of -"* from the task instruction, the generation quality of GPT-3 is degraded.

Context	**A**: I am sorry to tell you that you failed in the job interview in our company. **B**: It is tough to accept it, I mean. But can you tell me why? **A**: It is hard to tell. Maybe it is because you're not resourceful enough to be a manager.
Positive	**1.** Well what can I say! I am speechless, anyways thanks for calling. **2.** The interview was quite good, I answered everything I was asked for, and so what seems to be ... **3.** Well I was confident! It is really shocking that I couldn't clear it. **4.** But why! What is the matter? My performance was good. **5.** Anyways never mind, I will keep on trying, thank you for your information.
Random	**1.** Yogurt can be incredibly healthy. Unfortunately, most yogurts found in the grocery store are ... **2.** Well! I was really scared; maybe I need few more lessons. **3.** I too don't have any idea what they mean. **4.** Okay, you can have a day off. **5.** Well! Again it depends on family to family as I said to you earlier.
BM25	**1.** It is tough to accept it, i mean. But can you tell me why? **2.** I am indeed. What can you tell me? **3.** I am sorry, sir. Can you tell me your room number? **4.** Can you tell me why you replied to our advertisement? **5.** What's the matter? Please tell me. Maybe i can help you.
Key-sem	**1.** It's hard to say, but it sure was. You did a good job as a probationary employee, didn't you? **2.** No problem. The company will give me the job interview. **3.** No, that's better. I think we'd better do that at the least convenient time. **4.** I don't know exactly. I've never had any difficulty in starting a business. Why did you ask me to come in to talk about your business? **5.** Sorry, i'm not very resourceful enough to go there.", "I'm sorry. I'm too disappointed.
Mask-and-fill	**1.** No, that is not quite what i was thinking. What can you tell me? **2.** Can you tell me why do you want to get a taste for that? **3.** It is hard to tell. Get used to this kind of weather. **4.** First, i am looking for a shop to fix my bathroom. It's only the second time my water bill has been paying the bills recently. **5.** Anyways never mind giving up. I'm on my way out now, thank you for your information.
PNEG (Ours)	**1.** If you spend most of your time on the phone, then you are certainly not unemployed. **2.** It is hard to tell the difference between our company's products and those of my friend's company. **3.** I have tried to keep track of cross-cultural communication skills through the internet. **4.** You are so resourceful that you can work with people from all over the world. **5.** My partner has no resources to go over the financial reports. I am confused why she wants to be a part of this company again.
Human	**1.** Are you sorry for yourself for not being resourceful? **2.** It is tough to accept that I should attend the meeting without having lunch in the afternoon. **3.** Tomorrow there is a job interview in Titan company. **4.** She marched into the office and demanded to speak to the manager but she failed to talk. **5.** They were resourceful and talented people in our school.

Table 10: Negative response examples from various methods for a dialogue context with its positive responses. The dialogue context and its multiple positive responses are from *DailyDialog++* dataset.

H Adversarial Negative Response Samples from PNEG and other baselines

Table 10 shows sample negative responses from various methods and 5 human-written relevant responses for a single dialogue context.

I Task Instruction Type (I_{dir}, I_{pos}, and I_{imp})

The following task instructions are the detailed specification of the task instruction types used in Section 4.3.

I_{dir}=`Create five irrelevant responses containing keywords of the given dialogue context:`
`1.`

I_{pos}= `Relevant response: You need to know something about the company.`
`Create five irrelevant responses containing keywords of the given dialogue context:`
`1.`

I_{imp}=`Suddenly, B(or A) makes an awkward response. The response appears to be okay at first glance, but it's irrelevant to the conversation.`
`B (or A):`

Context	**A**: I am sorry to tell you that you failed in the job interview in our company. **B**: It is tough to accept it, I mean. But can you tell me why? **A**: It is hard to tell. Maybe it is because you're not resourceful enough to be a manager.
PNEG (OURS) ***with optimal configuration***	
E=100% k=2 I=I_{pneg}	**1.** If you spend most of your time on the phone, then you are certainly not unemployed. **2.** It is hard to tell the difference between our company's products and those of my friend's company. **3.** I have tried to keep track of cross-cultural communication skills through the internet. **4.** You are so resourceful that you can work with people from all over the world. **5.** My partner has no resources to go over the financial reports. I am confused why she wants to be a part of this company again.
Varying E with {10%, 1%, 0.1%, and 0.1%+REUSE}, k=2, I=I_{pneg}	
E=10%	**1.** We must consider the ways to reduce our expenditure. **2.** We simply have to make more of an effort if we are going to succeed in the future. **3.** I already know that I want that job - what is your final offer? **4.** I am scared that our luck has run out and we will never win again. **5.** This is my personal opinion ."
E=1%	**1.** I don't know how to thank you for the kind of lesson I received from you. **2.** My sister would like to study in your school next year, but she doesn't want to leave the place. **3.** Our new teacher is rather young and pretty. What a pity that she's married. **4.** We will go on a trip this Saturday as my parent promised me. **5.** I am going on a date with my boyfriend tonight, but it doesn't seem too good as he told me that he would ask for a break up tomorrow night.
E=0.1%	**1.** I am the best and I know it. You can't beat me. **2.** I got a very good score in my English class. What's the big deal? **3.** I will be back to take revenge. **4.** The interview was held in the manager's room, who is currently out of town for a business trip. **5.** Tough to accept it, but I'm here to accept it with an open heart and welcome the next round
E=0.1% +REUSE	**1.** Please give me a raise in salary, please. **2.** I think you're an interesting manager. **3.** Here is the money I owe you. **4.** I really appreciate your kindness. **5.** Your answer is totally off my expectation! That's so amazing.
Varying k with {0,1}, E=100%, I=I_{pneg}	
k=1	**1.** Would you like to drink a cup of coffee together? **2.** I learned the recipe for a new dish yesterday and would like to try making it. **3.** Would you please give me some money now? **4.** You look tired; I'll take your job and do it myself then. **5.** Do you think you could go shopping with me later? I want to know what kind of houses are ...
k=0	**1.** I don't know why I failed to win the race. Maybe it is because my legs were too short. """, **2.** I am sorry to tell you that you failed in the job interview in our company. **3.** Well! I am stunned. That's it? Thanks for telling me at least a reason. **4.** Sir, I am sorry to tell you that you failed in the job interview in our company. **5.** It is hard to tell . Maybe it is because you're not resourceful enough to be a manager.
Varying (I,k) with {(I_{pos}, 2), (I_{imp}, 0)}, E=100%	
k=2 I_{pos}	**1.** Thank God, otherwise it would have been worse. **2.** What? I am not able to listen because the person next to me is shouting. **3.** Oh no! can I reduce my cell phone bill by getting to hear this news? **4.** I need the electric blanket in my room to survive this winter season. **5.** MHe's probably referring to the lack of managerial skills that I have established till now.
k=0 I_{imp}	**1.** I want you to think again. **2.** I already have the new pair of glasses. **3.** I used to be a soccer player when I was in high school. **4.** What a coincidence! How many first-class stamps do you have on hand? **5.** I have never been encouraged enough to have a good start.

Table 11: Example of negative responses generated by PNEG with varying the components. E, k, and I indicates the size of example dataset, number of examples, and task instruction type, respectively. The optimal configurations that are used in PNEG are E=100%, k=2, and $I = I_{pneg}$.

J Adversarial Negative Response Samples from PNEG with Changing Prompts

Table 11 shows sample negative responses from PNEG with varying size of example dataset (E), number of examples in a context (k), and the task instruction type (I), following our ablation studies. Note that dialogue context in Table 11 is same with Table 10.