Generating Negative Samples by Manipulating Golden Responses for Unsupervised Learning of a Response Evaluation Model

ChaeHun Park **Eugene Jang** **Wonsuk Yang** **Jong C. Park***
School of Computing
Korea Advanced Institute of Science and Technology
{ddehun,eugenej,derrick0511,park}@nlp.kaist.ac.kr

Abstract

Evaluating the quality of responses generated by open-domain conversation systems is a challenging task. This is partly because there can be multiple appropriate responses to a given dialogue history. Reference-based metrics that rely on comparisons to a set of known correct responses often fail to account for this variety, and consequently correlate poorly with human judgment. To address this problem, researchers have investigated the possibility of assessing response quality without using a set of known correct responses. Tao et al. (2018) demonstrated that an automatic response evaluation model could be made using unsupervised learning for the next-utterance prediction (NUP) task. For unsupervised learning of such a model, we propose a method of manipulating a golden response to create a new negative response that is designed to be inappropriate within the context while maintaining high similarity with the original golden response. We find, from our experiments on English datasets, that using the negative samples generated by our method alongside random negative samples can increase the model's correlation with human evaluations. The process of generating such negative samples is automated and does not rely on human annotation.[1]

1 Introduction

Automatic evaluation of responses can be difficult because multiple answers could be suitable for a single context. Well-known metrics often used in machine translation or text summarization, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), or ROUGE (Lin, 2004), are based on measuring n-gram overlap with a set of human-annotated golden answers. Compared to machine translation or text summarization systems, conversational systems have a wider range of acceptable responses to a given situation (dialogue history). This could explain the low correlation between n-gram-based evaluations and human-conducted evaluations for responses generated by conversation systems, as reported by Liu et al. (2016). They also suggested calculating the embedding similarities between responses and correct answers, and showed that these metrics had a higher correlation with human evaluations than n-gram-based metrics. As this method only rewards responses similar to ones in the fixed set of answer candidates, however, it still fails to account for other possible answers that are dissimilar to the known answers.

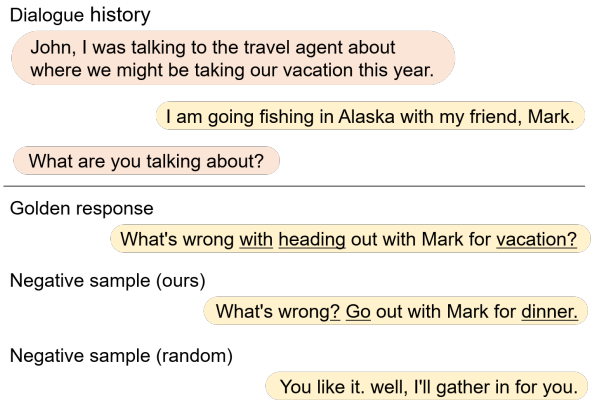

Figure 1: Example of the three different types of responses for a given dialogue history. Our method manipulates the original response *"What's wrong with heading out with Mark for vacation?"* to generate the negative sample *"What's wrong**? Go** out with Mark for **dinner.**"*

To solve this problem, Lowe et al. (2017) proposed a supervised regression model that makes predictions independent of correct answer candidates. Although they were able to achieve better correlation with human evaluations, their method depends on procuring a human-annotated dataset to learn from. Tao et al. (2018) used the Next-

* Corresponding author
[1]The code is available at https://github.com/nlpcl-lab/dialog-eval-hard-negative.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1525–1534
June 6–11, 2021.

Utterance Prediction (NUP) task to learn for automatic response evaluation. Their model, which is unsupervised, learned to distinguish an appropriate response from random negative samples (responses randomly taken from the training corpus). The model can evaluate the response quality by estimating the probability that the response occurs directly after the dialogue history. They also demonstrated that the probability-based evaluations highly correlated with human evaluations of response quality.

In this paper, we propose a method to create a negative sample by manipulating a golden response. The manipulation is carried out in three steps: (1) scoring each word, (2) selecting words to replace, and (3) replacing the selected words. In the first step, each word is assigned a score designed to determine how dependent the word is on the context. In the second step, we select all the words with a score above a threshold value, where higher scores indicate higher dependency to the dialogue history. In the third step, all previously selected words are masked and replaced with words predicted in their place by a pretrained language model (LM). Figure 1 shows an example of a negative sample generated by our method. When *"What's wrong with heading out with Mark for vacation?"* is the golden response, the tokens *"with"*, *"heading"*, *"vacation"*, and *"?"* were selected and replaced with *"?"*, *"Go"*, *"dinner"*, and *"."*, in that order.

We find that the model trained with our negative samples alongside random negative samples shows a higher correlation with human evaluations than the models trained only on random negative samples, in experiments using two datasets (Zhao et al., 2020). We also find evidence that automatic evaluation systems trained with the negative samples generated by our proposed method can make decisions closer to human judgment than those without.

The contributions of this paper are as follows:

(1) We introduce a method that automatically generates negative samples from the golden responses.

(2) We show that the negative samples can boost unsupervised learning of an automatic response evaluation model with experiment results.

(3) We conducted crowdsourcing and used its results to examine whether the negative samples generated by our method are actually negative.

2 Related Work

Liu et al. (2016) pointed out that the traditional n-gram overlap based metrics such as BLEU, METEOR, and ROUGE show low correlation with human evaluations when used to evaluate the results of an open-domain conversation system. They suggested measuring the similarity by comparing embeddings of a generated response to those of the golden response. Li et al. (2016) explored dialog system with textual feedback. Ghandeharioun et al. (2019) suggested the necessity of interactive human evaluation for dialogue systems, and proposed a self-play scenario to reduce the burden of human effort. Hashimoto et al. (2019) proposed a method to combine human assessments with the predictions of an evaluation model.

Lowe et al. (2017) proposed a supervised learning method to predict the quality of a response directly, rather than measuring the similarities with golden responses. Tao et al. (2018) showed that a model trained on the NUP task, in an unsupervised manner, can be used to predict the quality of a response that is generated by a system. Ghazarian et al. (2019) improved the previous work by using contextualized word embeddings. Mehri and Eskenazi (2020) proposed two unsupervised evaluation models: one based on masked language modeling (MLM) and another based on the response retrieval task using a pretrained LM. Pang et al. (2020) predicted the coherence and fluency of a response by estimating its likelihood using a LM.

Sai et al. (2020) emphasized the importance of adversarial negative samples for learning response evaluation, and released a dataset with human-curated adversarial negative responses. Their negative samples were manually curated, however, whose process can be both time-consuming and expensive. Wu et al. (2020) attempted to improve the performance of evaluation models for abstractive summarization by corrupting the golden summary and using it as a negative sample. In the machine translation task, Sellam et al. (2020) created paired data with synthetic examples, through methods such as back-translation and mask-filling with BERT (Devlin et al., 2019), and they used the paired data to pretrain the evaluation models. Our work introduces a method to create negative samples by manipulating the golden response to the dialogue history, and also suggests that the negative samples generated by the proposed method could be used to improve the unsupervised response evaluation model. The proposed method can be performed automatically without human effort.

3 Method

In this section, we describe our method to generate negative samples. The proposed method creates a negative sample by selecting and replacing specific word(s) in a golden response. The word selection is based on the difference between (a) the estimated probability that a word would appear in the response considering the dialogue history and (b) the estimated probability that the word would appear in the response when the dialogue history is not considered. An LM that can perform MLM can be used to estimate these probabilities. Words that have large differences in probability are selected and replaced with other words. When replacing a word with another word, an LM that can perform MLM can be used to predict the word that is most likely to appear in the position of the original word when the dialogue history is *not* given.

3.1 Scoring

The proposed method includes a scoring process to determine which words in the golden response are affected the most by the dialogue history. The score of a word is calculated by taking the difference between (a) the estimated probability of the word appearing in its position when the dialogue history is given and (b) the estimated probability of the word appearing in its position when the dialogue history is not given. This scoring process is performed independently for all words in the target response.

Specifically, to calculate the score of the *i*-th word (x_i) in the golden response, we first replace x_i with the `[mask]` token. Then the likelihood that the original word x_i appears in place of the masked token is calculated twice: once with the dialogue history and once without. The difference in the log-likelihood is used as the final score of each word, which is defined as

$$\begin{aligned}\text{score}(x_i \mid c, r_{/\text{i}}; \theta) = \log(\text{P}(x_i \mid [c; r_{/\text{i}}]; \theta)) \\ -\log(\text{P}(x_i \mid r_{/\text{i}}; \theta))\end{aligned} \quad (1)$$

, where x_i denotes the word to be scored, and $r_{/\text{i}}$ denotes the sequence of words in the golden response where x_i is masked. c denotes the dialogue history of the golden response, and $[;]$ the concatenation of two pieces of text. $P(x_i|[c; r_{/\text{i}}]; \theta)$ denotes the estimated probability that x_i would occur when the dialog history is considered. $P(x_i|r_{/\text{i}}; \theta)$ denotes the estimated probability that x_i would occur when the dialog history is not considered. θ denotes the parameters of the LM.

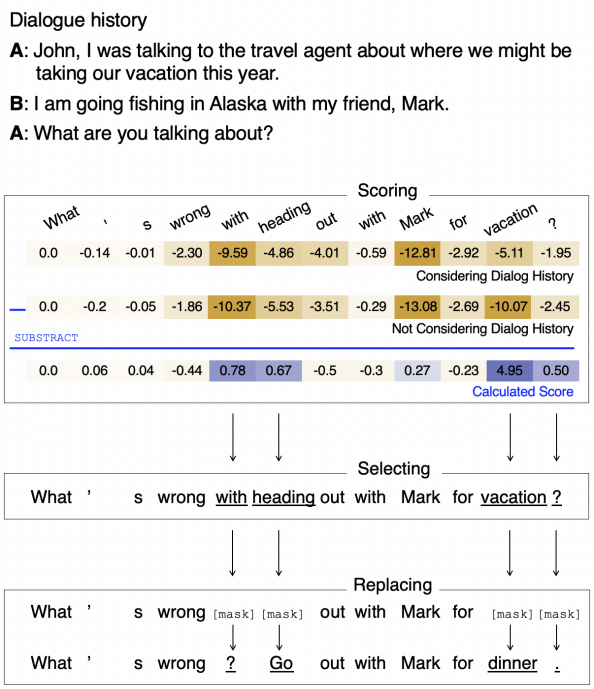

Figure 2: An illustration of our proposed method for generating a negative sample. The original response to the dialogue history, *"What's wrong with heading out with Mark for vacation?"*, is manipulated through the steps of scoring, selecting, and replacing.

Figure 2 shows an example of our proposed scoring process. The word *"vacation"* in the original response received the highest score among the words in the response. The words *"with"* and *"heading"* also scored higher than other words.

3.2 Selecting

For each sentence, we select words that scored higher than the threshold t. For example, in the case seen in Figure 2, if the threshold is 0.5, the words *"with"*, *"heading"*, *"vacation"*, and *"?"* will be selected. If none of the words receive a score higher than the threshold value, no words will be selected, and in this case, a negative sample cannot be generated.

We set the threshold t to 0.5 for our experiments. Using this threshold in our dataset, an average of 27.28% of tokens were selected for each response. Also, 94.89% of the responses contained at least one selected word, which means a negative sample could be generated for 94.89% of the cases.

3.3 Replacing

The selected words are then replaced using an LM. All selected words are replaced with `[mask]` tokens in the original response. Then the LM predicts, without considering the dialogue history, the words that are most likely to occur in the location of each masked word. If the LM predicts the original word, the second most likely word is used instead.

4 Experiments

4.1 Setting

4.1.1 Dataset

To measure the correlation between model predictions and human evaluations, we use the response-evaluation dataset proposed by Zhao et al. (2020). The dataset contains dialogue histories, machine-generated responses, golden responses, and appropriateness scores evaluated by human annotators. The scores were on a 5-point Likert scale, and each response was scored by four annotators. Six generative models, S2S (Sutskever et al., 2014), attentional S2S, HRED (Serban et al., 2016), VHRED (Serban et al., 2017), GPT2-sm and GPT2-md (Wolf et al., 2018), with three decoding algorithms, greedy decoding, ancestral decoding, and nucleus sampling (Holtzman et al., 2020), were used to generate the responses. They used DailyDialog (Li et al., 2017) and PersonaChat (Zhang et al., 2018). For each dataset, they trained a set of generative conversation models. Each of the 900 context-response pairs was randomly selected from the test set of the two datasets, and the annotators evaluated the appropriateness of each response to the context to construct two different evaluation datasets. The Krippendorff's alpha for this dataset was 0.815, suggesting reasonable inter-annotator agreement.

DailyDialog dataset consists of 13,118 multi-turn open-domain conversations written by human workers, and PersonaChat dataset consists of 12,875 multi-turn open-domain conversations written by human workers.

4.1.2 Models

The evaluation models used in the experiment are listed below. Among them, BLEU, ROUGE, METEOR, Embedding Average/Extrema/Greedy, and BERTScore are reference-based metrics that evaluate the quality of a response based on its similarity to the golden response. BERT-MLM, GPT2-coherence, BERT-retrieval (random-N), BERT-retrieval (ours) are unreferenced metrics that do not require golden responses. RUBER can be viewed as a hybrid metric that includes both reference-based and unreferenced approaches. Some of the reference-based metrics are simple comparison methods, rather than trainable models, but are presented along with other models because they can also be used to estimate the quality of responses. It should be noted that we do not compare the unsupervised approaches listed below with supervised approaches, such as the ones proposed by Lowe et al. (2017); Zhao et al. (2020), which require human-annotated response-evaluation pairs for training.

BLEU is a widely used metric for the machine translation task by measuring n-gram precision between multiple references and a hypothesis (Papineni et al., 2002).

ROUGE is a widely used metric for text summarization, which measures the n-gram recall (Lin, 2004). We use the F-score of ROUGE-L as an appropriateness score.

METEOR is a metric for the machine translation task, which considers both n-gram precision and n-gram recall of a hypothesis (Banerjee and Lavie, 2005).

Embeddding Average/Greedy/Extrema calculate the similarity between golden and generated responses using the embedding similarity to account for the diverse ways in which the golden response could be stated (Liu et al., 2016).

BERTScore is a recently proposed unsupervised metric based on the contextualized BERT embeddings (Zhang et al., 2020).

RUBER calculates the scores of reference-based and unreferenced metrics individually, then uses them to predict the final score (Tao et al., 2018). The reference-based metric measures the similarity between golden responses and generated responses based on their embedding similarity. The unreferenced metric is trained on the NUP task.

BERT-MLM sums the log-likelihood of each token in a response after masking it using an LM that is fine-tuned on a corpus, then uses the aggregated likelihood as the final score of the response (Mehri and Eskenazi, 2020).

GPT2-coherence measures the coherence between the dialogue history and a response by using a fine-tuned GPT2 model (Radford et al., 2019) to compute the averaged log-likelihood of the response (Pang et al., 2020).

BERT-retrieval (random-N) is a BERT-based model that is trained to distinguish a golden response from a negative sample (Mehri and Eskenazi, 2020), using the dialogue history. We refer to the original model by Mehri and Eskenazi (2020) as BERT-retrieval (random-1) since they used one random response as a negative sample, for a dialogue history. We refer to a variation of the model that uses two random negative samples for a dialogue history, as BERT-retrieval (random-2). This is to fairly compare with our model, which uses two negative samples for a dialogue history, as explained below.

BERT-retrieval (ours) is a model that has the same structure as the BERT-retrieval model. The difference is that our model utilizes the negative samples generated by the method that we propose. The model uses both the generated negative samples and the random negative samples. Specifically, during training, the model learns to distinguish a golden response from two negative samples: one generated from our method and one randomly sampled from the corpus.

4.1.3 Implementation Details

We trained the unreferenced models on the original DailyDialog dataset, and then evaluated them on the two response-evaluation datasets (Section 4.1.1). We split the conversations in the DailyDialog dataset in a sliding window manner to construct pairs of dialogue histories and corresponding responses. The maximum turn of the dialogue history was set to 5, following Zhao et al. (2020).

We use the pretrained BERT and GPT2 released by Wolf et al. (2018) for all of our relevant experiments.[2] A BERT model, fine-tuned on the DailyDialog train set with MLM for 1 epoch, was used for the scoring step of our proposed method (Section 3.1). The same model was used for the replacing step (Section 3.3). We used the threshold[3] of 0.5 for the selecting step (Section 3.2). We used Adam optimizer (Kingma and Ba, 2015) for training. We searched for hyperparameters for the BERT-retrieval (random-1) model, that maximize the (Pearson) correlation between human evaluations and model predictions on the response-evaluation dataset made from DailyDialog dataset (Section 4.1.1). The values found in this search (epoch=3, batch size=64, and learning rate=2e-5) were used for all the BERT-retrieval models (random-N, ours). The random seed was fixed for all experiments.

[2] `bert-base-uncased` and `gpt2-12layer` are used.

[3] We tested threshold values of 0, 0.5, 1, and 2, and found that using 0.5 as the threshold achieved the highest correlation with human evaluations; therefore we report only the experiment results with this value.

Model	DailyDialog r	DailyDialog ρ	Persona r	Persona ρ
BLEU	.08*	.05*	.19	.23
METEOR	.11*	.33*	.23	.18
ROUGE	.12	.09*	.25	.21
Embed. Average	.09*	.08*	.16	.17
Embed. Greedy	.18	.18*	.25	.24
Embed. Extrema	.16	.15*	.28	.27
BERTScore	.13	.12	.28	.26
RUBER	.28	.26	.07*	.04*
BERT-MLM	.32	.38	.35	.35
GPT2-coherence	.47	.47	.48	.48
BERT-rtv. (rand1)	.47	.47	.56	.60
BERT-rtv. (rand2)	.49	.48	.55	.58
BERT-rtv. (ours)	**.55**	**.56**	**.64**	**.66**

Table 1: The correlations between model predictions and human evaluations for each model, based on the two response-evaluation datasets. The highest score on each metric is highlighted in bold. All values with $p > 0.001$ are marked with *. **DailyDialog** and **Persona** denote the response-evaluation datasets made from DailyDialog and PersonaChat datasets, respectively. BERT-rtv. denotes the BERT-retrieval model. r and ρ mean the Pearson correlation and Spearman's rank correlation coefficient, respectively.

Model	DailyDialog r	DailyDialog ρ	Persona r	Persona ρ
drop-golden	.49	.48	.54	.57
shuffle-golden	.46	.45	.55	.59
score-w/o-history	.51	.52	.58	.58
select-random	.54	.55	.57	.59
replace-w-history	.52	.52	.57	.57

Table 2: The correlations between model predictions and human evaluations for each of the variations of our model, based on the two response-evaluation datasets.

4.2 Results

In Section 4.2.1, we check the correlations between the results of each evaluation model and human evaluations. In Section 4.2.2, an in-depth analysis of our proposed method is shown. In Section 4.2.3 we present examples that may suggest that automatic evaluation systems that have been trained with the proposed method can make deci-

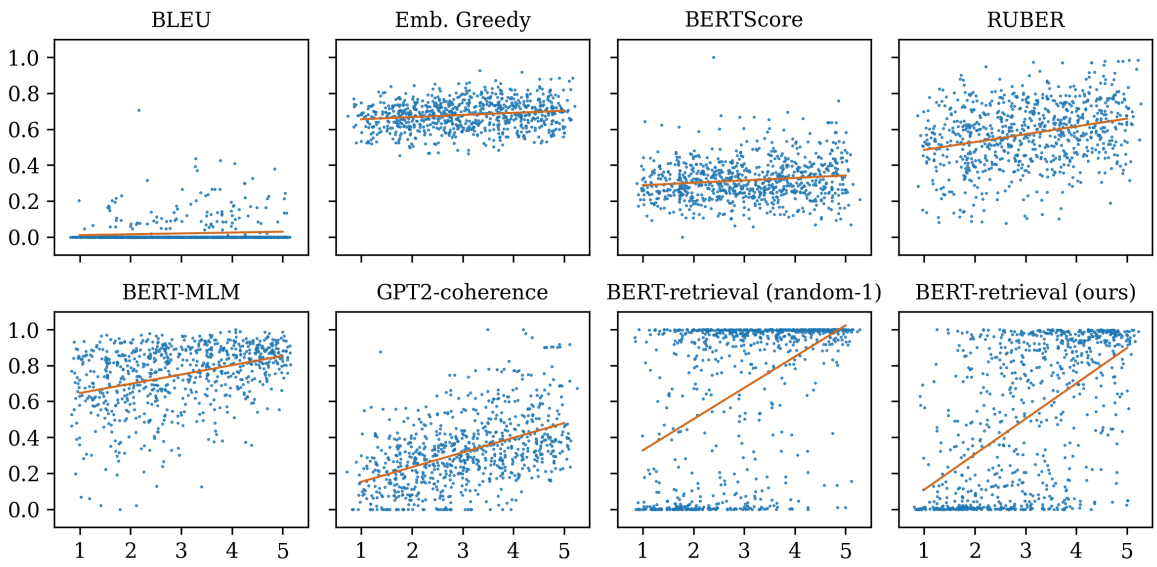

Figure 3: The scatter plots that show in detail the correlations between model predictions and human evaluations. Each of the plots contains 800 system-generated responses in the response-evaluation dataset made from DailyDialog dataset (Section 4.1.1). Each point indicates a response. Its x-value indicates the human evaluation score for the quality of the response, given on a 5-point Likert scale. Its y-value indicates the model prediction for the quality of the response, normalized into the range of [0, 1]. The orange line is a linear regression. We add a noise sampled from $\mathcal{N}(0, 0.09)$ into human score for better visualization, following previous studies (Lowe et al., 2017; Bak and Oh, 2020; Pang et al., 2020).

sions closer to human judgment than models that have not.

4.2.1 Correlation with Human Judgment

Table 1 shows the correlation between model predictions and human evaluations for each model, based on the two datasets. Pearson correlation (r) and Spearman's rank correlation coefficient (ρ) were used to measure the correlation between human score and model prediction. It should be noted that we excluded the scores of golden responses from the response-evaluation datasets and extracted 800 and 750 response-evaluation pairs from the DailyDialog and PersonaChat datasets, respectively. The model incorporating our negative sample method made predictions with higher correlation with human evaluations than the predictions made by BERT-retrieval (random-2), which uses the same number of negative samples for training. Among the baseline models, most of the reference-based metrics showed comparatively low performances. It is thought that these results support the observations made by previous studies suggesting that using the golden response as the "one and only" correct answer to evaluate responses can be ineffective. RUBER showed better performance than other reference-based models for the DailyDialog dataset, but showed low performance in evaluating PersonaChat responses. The GPT2-coherence model showed similar performance to the BERT-retrieval (random-1) model on the DailyDialog dataset, but relatively low performance in the PersonaChat dataset. It should also be noted that the hybrid and unreferenced models were trained on the DailyDialog dataset, and not on the PersonaChat dataset.

Figure 3 shows a scatter plot visualizing the human scores and model predictions for the response-evaluation dataset on DailyDialog. BLEU tended to predict low scores. This may suggest that there were only a few n-gram overlaps between the golden responses and the generated responses. The predictions of embedding-based metrics (Emb. Greedy and BERTScore) were concentrated on a specific range, and showed low correlation with human scores. The unreferenced or hybrid metrics (RUBER, BERT-MLM, GPT2-coherence, and BERT-retrieval (random-1)) show relatively higher correlations than the reference-based metrics. We can see that BERT-retrieval (ours) shows the greatest correlation among the models, with a correlation coefficient of 0.1974. The scatter plots suggest that false-positive predictions, which frequently occurred in the BERT-retrieval (random-1) predic-

tions, occurred less frequently in our model's predictions. However, the scatter plot for our model has a step-function-like appearance. Most of the responses received a score near 0 or near 1, and this is problematic because an ideal model should be able to match human scores even when the scores are moderate. This tendency is considered as a limitation of our model that must be addressed in the future work.

4.2.2 Model Analysis

We analyze our model, by performing experiments with some variations in making the negative samples to be used with the random negative sample: (1) *drop-golden*: Instead of following the steps of scoring, selecting, and replacing, we randomly drop some of the words in the golden response to create a negative sample, and use it with the random negative sample. (2) *shuffle-golden*: Instead of following the three steps, we randomly shuffle the words in the golden response to create a negative sample, and use it with the random negative sample. (3) *score-w/o-history*: We use the scoring function in Equation 1 without the first term, so that it only considers the probabilities within the sentence without the dialogue history. (4) *select-random*: Instead of using the scoring function proposed in Equation 1, we randomly select the words to be replaced. (5) *replace-w-history*: When replacing a word, we concatenate the dialogue history with the response so that the LM considers the dialogue history when replacing the masked words.

Table 2 shows the correlations between model predictions and human evaluations for the modified models above. Dropping or shuffling words in the golden response to make a negative sample shows similar or lower performance compared to using random responses (BERT-retrieval (random-1, random-2)). The correlation was lower when the dialogue history was not considered in the scoring process than when it was considered. We speculate that this is because it gives high scores not only to words important for the consistency of a conversation, but also to the words with low likelihoods in general. Randomly selecting the tokens shows lower correlation than using our proposed scoring function. Considering the dialogue history in the replacing process gives lower performance than when it is not considered. We speculate that providing the dialogue history makes predictions on the masked words that are more appropriate to the context, making the reconstructed response less appropriate as a negative sample.

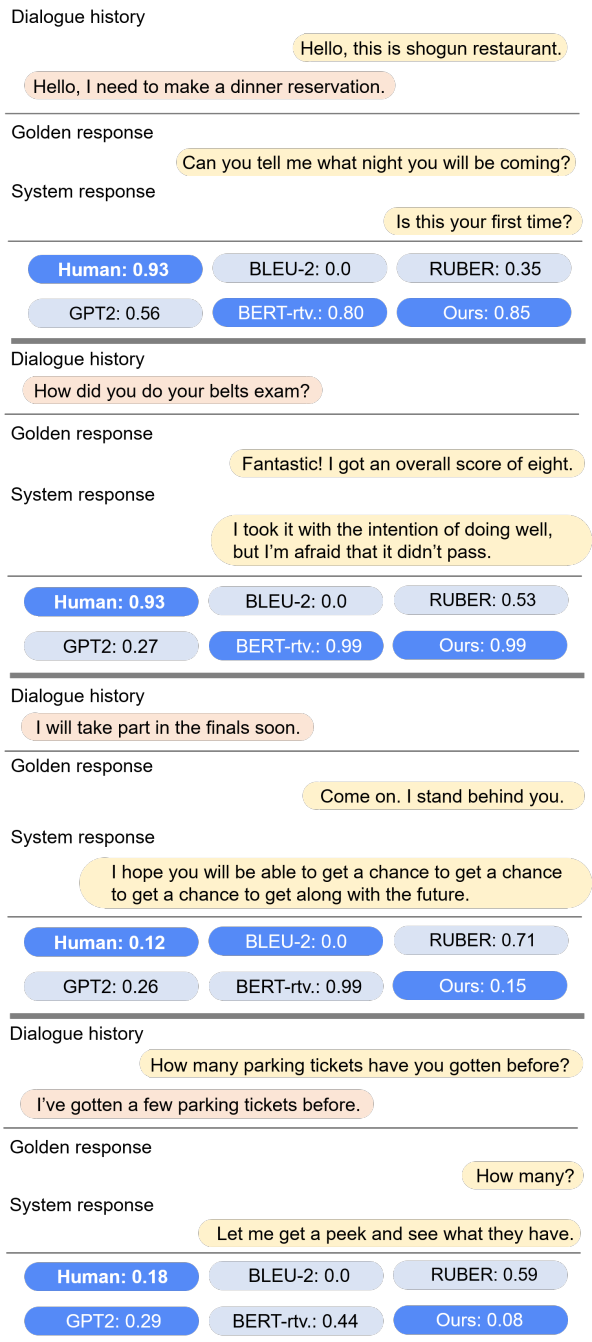

Figure 4: Some examples of cases in which our model predicted scores similar to human evaluations. GPT2 denotes the GPT2-coherence model. BERT-rtv. denotes the BERT-retrieval (random-1). All scores are normalized into the range of [0, 1].

4.2.3 Case Study

Figure 4 shows some of the evaluation results of each model on the DailyDialog dataset. The responses in the first and second examples are appropriate to the given dialogue history as suggested by the high human score. BLEU-2 gives a score of 0

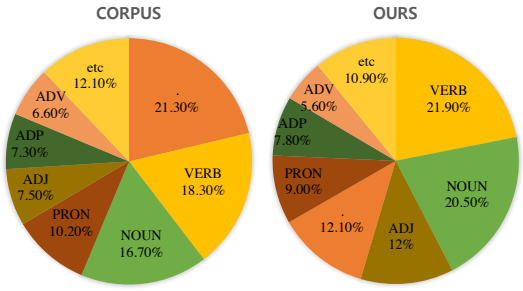

Figure 5: The POS tag distribution of the words in the original training corpus (left) and the selected words by our method (right).

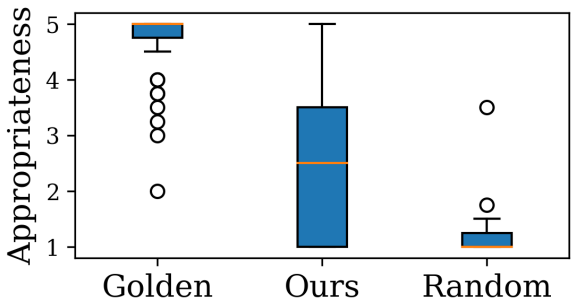

Figure 6: Box plot of scores for each type of responses. The average scores of each type are 4.65, 2.51 and 1.19. The standard deviations for the scores of each type are 0.67, 1.27, and 0.41 (from left to right).

because the response has no bi-grams shared with the golden response. RUBER and GPT2-coherence did not recognize the utterances as appropriate responses. BERT-retrieval (random-1) and BERT-retrieval (ours) gave relatively high scores to the responses, evaluating them as appropriate utterances. In the third example, the system response appears to be somewhat relevant to the given context because it includes some words (*"chance", "future"*) relevant to the phrase *"take part in the finals"*. A repetition of a phrase in this example (*"to get a chance"*) is believed to have contributed to the low human evaluation score (0.12). The RUBER and BERT-retrieval (random) models appear to lack this intuition, and instead evaluate the response as appropriate, possibly because some words appear relevant. Our proposed model scored the response with a relatively low score of 0.15, which was close to the human score. In the fourth example, the response is not coherent, but because it begins with a sentence *"Let me get a peek"*, it could have appeared as a coherent response to the previous dialogue about parking tickets. For this case, our proposed model and GPT2-coherence gave scores similar to human scores.

4.3 POS-tag distribution of selected words

We compute the Part-of-Speech (POS) tag distribution of selected words by our method and compare it with the original distribution of the DailyDialog corpus (Figure 5). [4] As we can see, the VERB and NOUN tags are the most frequently selected (21.9% and 20.5%, respectively), and their ratio is increased than in the original corpus (18.3% and 16.7%, respectively). Meanwhile, the ratio of punctuation tag (.) is highly decreased (from 21.3% to 12.1%). We suspect that the likelihood of the punctuation tag is more affected by local information from a response rather than dialog history.

4.3.1 Are the generated samples actually inappropriate?

To see whether the negative samples generated by our method are actually inappropriate, we conducted a survey through Amazon Mechanical Turk (AMT). We selected 40 dialogue history examples and prepared three types of responses for each dialogue: 1) the golden response, 2) a negative sample generated by our method, and 3) a randomly selected negative sample from the corpus. For each dialog, 4 annotators were asked to score the quality of the three responses. Following Lowe et al. (2017), we asked the question *"How appropriate is the response overall?"* for each context-response pair, and the evaluation was conducted on a 5-point Likert scale. The Fleiss' kappa and Krippendorff's alpha for the annotations were 0.63 and 0.63, respectively.

Figure 6 shows the survey results. The mean scores of golden and random responses were 4.65 and 1.19, respectively. The mean score of our negative samples was 2.51. The standard deviations for the scores of each response type were 0.67, 1.27, and 0.41 for the golden response, our negative sample, and the random response, respectively. We see that these results do not guarantee that all the generated negative samples are inappropriate. What we can assume, however, is that our method of manipulating a golden response generates a negative sample that is more inappropriate than the golden response. Table 3 shows two examples of the three different types of responses for a given dialog history with their survey results.

[4] We use the NLTK POS tagger (https://www.nltk.org/book/ch05.html) with universal tagset.

Dialog History
A: Sir, would you like some dessert now? **B**: Please show me the menu again. **A**: Here you are sir. The chocolate cake is very delicious.
Responses
Golden: No, thanks. I don't like chocolate. I'd like strawberry pie. (5) **Ours**: No, thanks. I don't have chocolate. I'll like some one. (1.5) **Random**: I basically believe in science over theology. I mean , I (...) (1)
Dialog History
A: Could you tell me something about your family ?
Responses
Golden: Ok. There are five people in my family, father, mother, elder brother, younger sister and I. (5) **Ours**: Ok. There are five children in my family, father, mother, and brother, and father my me. (3.25) **Random**: When do you want to move in? (1.25)

Table 3: Examples of three different types of responses for a given dialog history with their survey results. The highlighted words are newly generated by our method. The score of each response is underlined.

For a model learning to find the difference between appropriate and inappropriate responses, we speculate that the task of distinguishing the negative samples generated by our method from the golden responses would be more difficult than the task of distinguishing the randomly selected negative samples from the golden responses. We believe that this is because the generated negative samples can be inappropriate in more subtle ways than completely unrelated responses are. We suspect that learning with this more challenging setting have resulted in the performance gain that we discussed in Section 4.2.1. However, we believe that it will need a more in-depth semantic analysis on each of the cases, such as performing a more quantitative analysis (through an extensive human study, for instance) and further interpretation of the semantic relationships between the original golden responses and the modified negative samples according to the proposed method. We leave it as a future work.

5 Conclusion

In this paper, we proposed an automatic method for generating negative samples that can be used to train an unsupervised and unreferenced response evaluation model. We performed experiments to demonstrate that the proposed method can boost the unsupervised training of a response evaluation model. We analyzed the experiment results quantitatively, and examined some examples that show the distinct characteristics of our proposed method.

Acknowledgments

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government MSIT) (No. 2018-0-00582, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection).

References

JinYeong Bak and Alice Oh. 2020. Speaker sensitive response evaluation model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6376–6385.

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.

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*, pages 4171–4186.

Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating interactive human evaluation with self-play for open-domain dialog systems. In *Advances in Neural Information Processing Systems*, pages 13658–13669.

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.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1689–1701.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations*.

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

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2016. Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823*.

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*, pages 986–995.

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

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.

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*, pages 1116–1126.

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.

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

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

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

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. *arXiv preprint arXiv:2009.11321*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Iulian V 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 30th AAAI Conference on Artificial Intelligence*, pages 3776–3783.

Iulian Vlad 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. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, pages 3295–3301.

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

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2018. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 722–729.

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*.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621.

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*, pages 2204–2213.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations*.

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.