CORAL: Contextual Response Retrievability Loss Function for Training Dialog Generation Models

Anonymous ACL submission

Abstract

Natural Language Generation (NLG) represents a large collection of tasks in the field of NLP. While many of these tasks have been tackled well by the cross-entropy (CE) loss, the task of dialog generation poses a few unique challenges for this loss function. First, CE loss assumes that for any given input, the only possible output is the one available as the ground truth in the training dataset. In general, this is not true for any task, as there can be multiple semantically equivalent sentences, each with a different surface form. This problem gets exaggerated further for the dialog generation task, as there can be multiple valid responses (for a given context) that not only have different surface forms but are also not semantically equivalent. Second, CE loss does not take the context into consideration while processing the response and, hence, it grades the response irrespective of the context. To grade the generated response for qualities like relevance, coherence, etc., the loss function should depend on both the context and the generated response. To circumvent these shortcomings of the CE loss, in this paper, we propose a novel loss function, CORAL, that directly optimizes recently proposed estimates of human preference for generated responses. Using CORAL, we can train dialog generation models without assuming non-existence of response other than the ground-truth. Also, the CORAL loss is computed based on both the context and the response. Extensive comparisons on two benchmark datasets show that CORAL based models outperform strong state-of-the-art baseline models of different sizes.

1 Introduction

Choosing the right loss function is crucial for getting the expected behavior from deep learning based models trained for any task. While the token-level cross-entropy (CE) loss continues to excel in training natural language generation (NLG) models for various tasks, including dialog-response generation (Roller et al., 2021; Zhang et al., 2020), it is well accepted that CE is not the most appropriate loss function to use for training dialog generation models, and finding the right loss function for training dialog generation is still an open problem and an active area of research (Shen et al., 2017; Zhao et al., 2017; Saleh et al., 2020; Li et al., 2021).

The CE loss is computed by comparing the predicted token probabilities to the ground truth target sequence from the dataset. Thus, computation of CE loss is unconditional or context-free as it does not depend on the input prompt/context in the case of conditional NLG tasks like dialog generation. To be able to generate responses for qualities like relevance, coherence, etc., the loss function should ideally consider both the context and the generated response. While training any NLG model using the CE loss, the probability of the ground truth response is maximized. Here, we make an implicit assumption that the ground truth is the only response possible for the given context. This is a major concern as this property does not hold for most dialogs where each context may have a large number of possible responses (Dou et al., 2021)[1].

Previous attempts in training Seq2Seq dialog generation models using the CE loss have led to various complications. Mode collapse is one of the most common issues when training a Seq2Seq model with the CE loss, mainly at smaller scales (Li et al., 2016, 2021). Here, the model will just assign a high probability to one or more generic and bland responses e.g. “I don’t know", “I have a problem", “Yes", etc., irrespective of the context. Previous research works have also explored various augmentations to the model architecture (Serban et al., 2016; Zhao et al., 2019) and/or the loss function (Serban et al., 2017; Shen et al., 2017; Zhao et al., 2017; Li et al., 2021) to resolve the common

[1]These problems are specific to dialog systems only and may or may not apply to other NLG tasks.

problems with CE, as mentioned above.

To train dialog generation models that maximize some user-expected qualities, we propose to directly optimize an estimate of human perceived quality of a context-response pair. Sinha et al. (2020); Yeh et al. (2021) have shown that the output score of retrieval models correlates strongly with human perception of dialog response quality (given that it was trained on a dataset of domain similar to the target domain of application). Further, Santra et al. (2021) showed that representations learned by a response retrieval (using binary cross-entropy or a contrastive loss) capture important dialog understanding features even better than large-scale dialog generation models trained using CE. But that is not applicable for a generative setting.

These observations triggered us to propose a novel **Contextual Response Retrievability** loss function or **CORAL** that circumvents issues with CE loss and relies on a response retrieval model. Further, we leverage CORAL to design a learning framework for dialog generation models that treats score calculated by retrieval models as a reward function, referred to as **Response Retrievability Reward** (R_3). We then use reinforcement learning (RL) to train dialog models that maximize above reward between context and generations from the optimized network. To test the proposed loss function, we train transformer-based Seq2Seq models (Vaswani et al., 2017) using CORAL for open-domain dialog generation using DailyDialog dataset (Li et al., 2017) and for domain-specific dialog generation using DSTC7-Ubuntu dataset (Yoshino et al., 2019). We compare the performance against state-of-the-art CE-based Seq2Seq models and various other baselines using both automatic metrics and through human evaluation. To summarize, our contributions are as follows. (1) We propose CORAL loss function for training dialog generation models by directly optimizing for an estimate of human preference of a ⟨context, response⟩ pair. To the best of our knowledge, we are the first to propose a loss function for dialog generation model that also relies upon the context. (2) Further, we propose a recipe to train improved seq2seq dialog models which uses the CORAL loss in a reinforcement learning setup. (3) We experimentally prove the effectiveness of CORAL against strong baseline models using CE or its variants. We make the code publicly available[2]

2 Literature Review

The use of the Seq2Seq model for training chitchat (now known as open-domain) dialog generation model was first proposed by Ritter et al. (2011). It was realized very soon that, unlike other NLG problems such as NMT, it is almost impossible to train dialog generation models using the CE loss, at small scales (Serban et al., 2016, 2017; Zhao et al., 2017; Li et al., 2016). Serban et al. (2016) argued that standard RNNs fail to encode the context properly and proposed a hierarchical encoder-based architecture, HRED, that captures utterance-level and context-level information using separate RNNs. To capture the notion of multiple possible responses per context, Serban et al. (2017) further augmented their HRED model with a VAE. Several prior works have also tried to formulate reward functions for training dialog generation models using Reinforcement Learning Li et al. (2016); Sankar and Ravi (2019); Saleh et al. (2020); Zhao et al. (2019).

Although at small scales CE-based training leads to degenerate dialog generation models, at (extremely) large scales of training data and model size, models can generate diverse, coherent and interesting responses. DialoGPT (Zhang et al., 2020) is based on the GPT-2 (Radford et al., 2019) language model (LM) further finetuned on a large conversational corpus crawled from Reddit. Blenderbot (Roller et al., 2021) focuses on generating high-quality responses and is based on a Transformer-based Seq2Seq architecture. To improve the quality of responses, Blenderbot is finetuned on the Blended-Skill-Talk dataset (combination of multiple datasets, Smith et al., 2020) after pretraining on Reddit. DialogRPT (Gao et al., 2020) proposes a sample-and-rank method for candidates generated from DialoGPT. The ranker was trained using a dataset of upvotes/downvotes and the number of replies on Reddit comments. Another family of recently proposed dialog generation models (Wu et al., 2019; Komeili et al., 2021) does retrieve-and-refine for efficiently training large models.

In this paper, we focus on testing the proposed CORAL loss mainly at small scale because 1) the changes are more apparent for smaller scale models, and 2) it becomes more difficult to judge the efficiency and robustness of the approach at larger

[2] https://anonymous.4open.science/r/2022-CORAL-Anonymous/

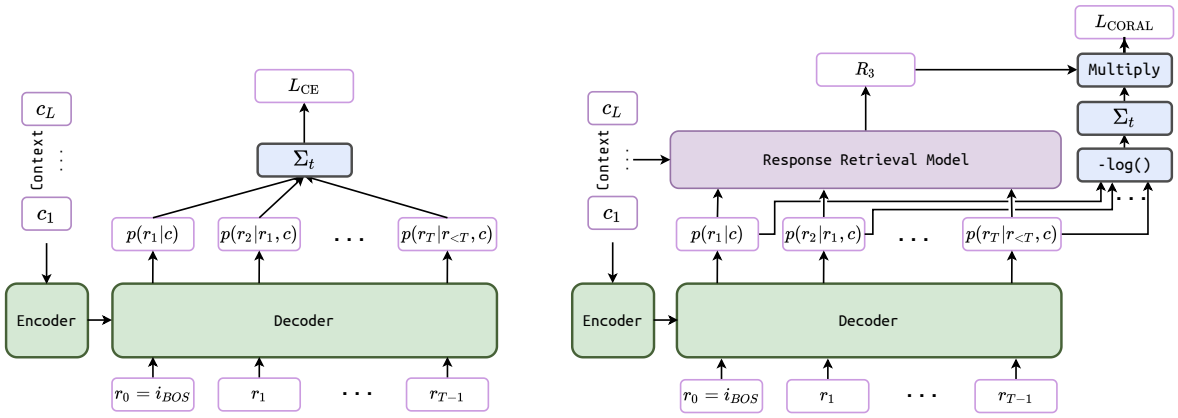


Figure 1: Schematics of Cross-Entropy (CE) and CORAL Losses at train time. The main idea of CORAL loss is to optimize a measure of the compatibility between the context and a candidate response (using a response retrieval model). Compared to CORAL, CE is more strict and trains the model only based on true response targets. CORAL also utilizes randomly sampled response targets and increases/decreases its probability of decoding based on R_3.

scales where all loss functions seem to work upto the same extent. Although we train small-scale models only, our small models outperform several large-sized models across multiple metrics.

3 Methods

First, using a response retrieval model, we design a reward function suitable for training dialog generation models (Section 3.1). Next, we apply reinforcement learning for training a Seq2Seq context-to-response generation model that maximizes the chosen reward function. We apply REINFORCE (Williams, 1992) to obtain the final differentiable objective function (Section 3.2), and use it to design our algorithm (Section 3.3) for training a Seq2Seq dialog generation model.

3.1 The R_3 Reward Function

Instead of optimizing perplexity or other heuristic based rewards with weak correlation to human feedback, we choose a reward function that directly optimizes an estimate of human preferences. Thus, we design reward functions based on response retrieval models. We avoid supervised model based estimates of human preferences (Lowe et al., 2017) for their known shortcomings (Sai et al., 2019).

Motivated by prior work of Sinha et al. (2020), we propose a reward function based on a (self-supervised) response retrieval model, ESIM (Chen and Wang, 2019), which has been shown to correlate strongly with human ratings of model generated responses (Yeh et al., 2021). Similarly, Santra et al. (2021) recently showed that response retrieval models actually optimize an estimate of the mutual information between context and response. Their experiments also prove that these representations are more feature-rich than large-scale pretrained generative dialog models (based on CE loss). Hence, we also experiment with their model, DMI, to design a reward function.

Contextual Response Retrievability Reward (R_3) We define the output score of an already trained response retrieval model for a context-response pair as the Contextual Response Retrievability Reward. This score is usually normalized between 0 and 1 and indicates the probability that the response is a valid (coherent and on-topic) continuation to the context. As the reward function would not be differentiable with respect to the target Seq2Seq-model parameters, we first formulate response generation as an RL task. Then, we apply REINFORCE (Williams, 1992) to obtain a differentiable objective function, as described next.

3.2 Final Objective Function (CORAL)

We pose the response generation problem as a reinforcement learning (RL) task. Each instance of the context-to-response generation task is considered as an episode in the RL formulation. The episode consists of several actions taken by the agent, in our case the decoder. Each action corresponds to generation of an output token. When the agent generates an EOS (end-of-sequence) token or has produced a max number of allowed tokens (T), the episode ends and the environment (the response retrieval model) generates the R_3 reward, for the $\langle$context c, generated response $r\rangle$ pair.

The updates to the Seq2Seq model weights are

then determined by the Episodic REINFORCE algorithmas follows.

$$\Delta W_{\text{CORAL}} = \eta R_3(c, r) \sum_{t=1}^{T} \frac{\partial \log P(r|c)}{\partial W_{\text{CORAL}}} \quad (1)$$

Loss function to be minimized (for an autoregressive decoder) can then be written as follows.

$$L_{\text{CORAL}} = -\eta R_3(c, r) \sum_{t=1}^{T} \log P(r_t|r_{<t}, c) \quad (2)$$

Algorithm 1 Training Algorithm for Seq2Seq Models using CORAL Loss (On-policy + Off-policy)

$D = \{(c, r^+)_i\}_{i=1}^{n}$ ▷ n Positive pairs from training dataset
$\theta^{(0)} \leftarrow$ Initialize Seq2Seq network weights
for $(c, r^+) \in D$ **do** ▷ Actual implementation uses batch gradient descent
  if $rand() > p_+$ **then**
    #Nucleus Sampling
    $r \leftarrow$ Sample $r^- \sim$ Nucleus$(c|\theta_{S2S}^{(k)})$
    # [or] RandomNegative Sampling
    $r \leftarrow$ Sample $r^- \sim$ Uniform(Training Utterance Pool)
  else
    $r \leftarrow r^+$ ▷ Use Positive Response
  $R_3(c, r) = f_R(c, r) - m$ ▷ f_R: Response Retrieval Model Score
  #Compute Decoder output token distribution
  $P_{S2S}(r_t|r_{<t}, c) \forall t \in [1, T]$
  $L_{\text{CORAL}} = -R_3(c, r) \sum_{t=1}^{T} P_{S2S}(r_t|r_{<t}, c)$
  #Update parameters of P_{S2S} using gradient descent on L_{CORAL}
  $\theta_{S2S}^{(k+1)} \leftarrow \theta_{S2S}^{(k)} - \alpha \nabla_{\theta_{S2S}^{(k)}} L_{\text{CORAL}}$

3.3 Training Algorithm

On-policy and Off-policy Training Using CORAL Loss RL training can be either on-policy or off-policy, depending on whether samples are generated from the parameterized policy network (the decoder in the Seq2Seq model in our case) or obtained from a dataset of human generated examples. For pure on-policy training, we will have to rely on response sequences randomly sampled from the decoder. But, because of the combinatorial complexity of the response space, it is highly unlikely that we would obtain any valid utterances/response candidates during on-policy training. Thus, to direct the model towards generating grammatically and semantically valid utterances, we mix on-policy and off-policy modes of training and refer to this method of sample generation as **mix-policy**. To control the amount of mixing, we introduce a hyperparameter called p^+ (detailed below). Since random sampling based decoding can sometimes generate very low probability tokens while generating a sequence, we use nucleus sampling instead of random sampling for generating on-policy samples. We also tried a method to

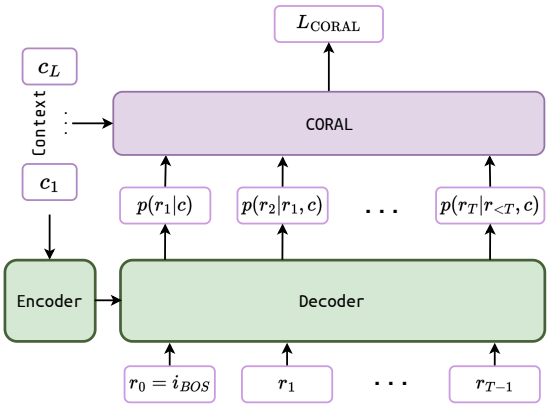


Figure 2: This figure shows the model architecture used for training the Seq2Seq dialog model using the CORAL loss. The context input to the Transformer-encoder is denoted by a token sequence $[c_1, c_2, ..., c_L]$. The response input sequence $r = [r_1, r_2, ..., r_T]$ denotes the positive or a randomly sampled negative response candidate.

exclusively utilize more off-policy samples, called RandomNegative sampling (or RNS), for generating more diverse off-policy samples during training. In RNS, we randomly sample utterances from the pool of all utterances from the training set and use it as a response candidate in the training process. In Algorithm 1, we show the exact steps used for training a Seq2Seq model using the CORAL loss and an existing dialog dataset D. Fig. 1 illustrates how the standard cross-entropy loss and the proposed CORAL loss differ from each other. Fig. 2 shows the architecture of our Seq2Seq model, which generates a response given the context using a Transformer encoder-decoder model trained under CORAL loss.

Hyperparameters of CORAL

(1) Probability of positive samples (p_+) denotes the probability with which we use the ground truth response for off-policy training.

(2) Margin (m) denotes the minimum reward that we expect from model generations. We use a fixed margin[3] value as the baseline reward for the RL training.

(3) Retrieval model: For implementing R_3, we experimented with these retrieval model architectures: ESIM (Chen and Wang, 2019), BERT (Devlin et al., 2019) and DMI (Santra et al., 2021).

Although CORAL is derived from quite a different viewpoint, under certain hyperparameter set-

[3]More flexible versions of margin are possible based on a learned value function (baseline function) or a critic function (in Actor-Critic formulations). We leave integration of such more advanced RL algorithms with the proposed CORAL-based learning framework as a future research direction.

tings ($m = 0, p^+ = 1, R_3 \in (0, 1)$) CORAL approximates a sample-weighted version of the CE loss. Also, training of a dialog generation model using CE may over-weigh generic responses more than the informative ones. A more detailed account of the similarities/differences between CORAL and CE are in the appendix. CORAL loss incorporates quality of the response as measured by a response retrieval model as a factor in computing the loss.

4 Experimental Setup

4.1 Model Setup

We use a standard transformer-based Seq2Seq (S2S) architecture for training a dialog generation model using the CORAL loss as shown in Fig. 2. We used the Word-piece tokenizer[4] from BERT (Devlin et al., 2019). We train our models using early-stopping, upto a maximum 50 epochs, based on validation-R_3 (average R_3 score of generated responses on the validation set). We use Adam optimizer with a peak learning rate of $1e-4$ that is warmed up (first 1000 steps) and decayed linearly. We use a single NVIDIA V100-16GB GPU-based system for training all our models.

For the retrieval models, ESIM has two LSTM encoding layers (Encoding and Composition Layers) for individually encoding the context and a candidate response, interleaved by a cross-attention layer. The sigmoid output from the ESIM model is used as the score for a context-response pair. Since both BERT and DMI are pretrained models, we add an MLP, with two hidden layers and sigmoid activation at the output, on top of the [CLS] token embedding. These models are finetuned on DailyDialog or DSTC7-Ubuntu for the response retrieval tasks. We obtain the reward (R_3) by subtracting the margin (m) (also between 0 and 1) from the output score (sigmoid output).

ESIM For faster learning, we also initialize the ESIM model's token embeddings with Blenderbot embeddings, which has a size of 2,560. The encoding and composition layers both used 1 layer LSTMs.

BERT and DMI We use the "bert-base-uncased" and the "DMI_Medium" checkpoints, respectively.

CORAL Our proposed CORAL model has 6 self-attention layers for both the encoder and the decoder. We use 8 self-attention heads and 1,024 as the size of hidden representations.

[4]We used the implementation by huggingface (Wolf et al., 2019) library (bert-base-uncased).

4.2 Data

We used the **DailyDialog** (DD) (Li et al., 2017) and **DSTC7-Ubuntu**[5] (Yoshino et al., 2019) datasets for all our experiments. DD is an open-domain dialog dataset in English. For training the retrieval model, we used randomly sampled utterances as negative samples. DSTC7-Ubuntu is a domain-specific dataset based on conversations from the Ubuntu IRC channel. DSTC7-Ubuntu contains majorly English conversations, with a small percentage instances from other languages. This is available directly as a dataset for training retrieval models with single positive and multiple negative responses per context. DD contains 76052, 7069 and 5740 context-response pairs for train, validation and test, resp. DTSC7-Ubuntu contains 470860, 23478 and 3247 resp. We make the code publicly available[6] and will release trained model checkpoints publicly upon publication of the paper.

4.3 Baselines

Small Scale Baselines:

(1) Mirror (Li et al., 2021): Seq2Seq model that extends CVAE (Shen et al., 2017), and is trained with a backward-reasoning loss function. It optimizes for generating both final and pre-final utterances in a bidirectional fashion. This is state-of-the-art loss function for training small-scale dialog generation models outperforming Shen et al. (2017); Zhao et al. (2017); Saleh et al. (2020).

(2) AdaLabel (Wang et al., 2021): Uses an adaptive label smoothing to prevent the model from being overconfident over a single choice. It also uses a soft-target distribution depending on the context, instead of usual one-hot distribution.

Large Scale Baselines:

(1) Blenderbot (Roller et al., 2021): Transformer-based S2S model pretrained on a large dialog corpus based on Reddit and finetuned on Blended-Skill-Talk dataset (Smith et al., 2020).

(2) DialoGPT (Zhang et al., 2020): GPT-2 (Radford et al., 2019) based language model further finetuned on dialogs from Reddit.

(3) DialogRPT (Gao et al., 2020): A response ranking model trained on a dataset of upvote/downvote and number of replies on Reddit comments. For generation purposes it reranks sampled responses from DialoGPT and returns the one with highest

[5]https://github.com/IBM/dstc-noesis

[6]https://anonymous.4open.science/r/2022-CORAL-Anonymous/

Table 1: Results for DailyDialog and DSTC7-Ubuntu datasets: From the results, we can see that by optimizing the contextual R_3 score directly, using REINFORCE, the CORAL model is able to produce coherent and diverse responses at the same time. The average length is reported to make sure that the model is not resorting to short utterances, such as "I don't think I know about *[topic_word]*", just to be coherent. **CORAL**$_x$ denotes a Seq2Seq model trained with CORAL loss. 'x' identifies the retrieval model used for the R_3 reward. Note that we used the DialoGPT-medium, DialogRPT-medium, Blenderbot checkpoints. *We have kept these large-scale baselines as a separate group, in the table.* Each value reported for CORAL models is computed as average of 5 runs.

		DSTC7-Ubuntu							
		Model	Avg. Len	BLEU	METEOR	Dist-1	Dist-2	MAUDE-ESIM	MAUDE-BERT
		Ground Truth	13.73	NA	NA	0.0922	0.5231	0.8583	0.7363
Large Models	Zero-shot	DialogRPT (ZS)	14.17	0.0714	0.0424	0.0351	0.1333	**0.7328**	**0.6813**
		DialoGPT (ZS)	8.59	0.0565	0.0369	**0.0504**	**0.1723**	0.7058	0.6557
		Blenderbot (ZS)	17.78	**0.0869**	**0.0571**	0.0266	0.0893	0.5683	0.5294
	Fine tuned	DialoGPT (FT)	6.22	0.0527	0.0459	**0.1101**	**0.4017**	0.8210	0.6610
		DialogRPT (FT)	13.63	0.0902	0.0628	0.0710	0.3219	**0.8731**	0.7485
		Blenderbot (FT)	16.98	**0.1210**	**0.0905**	0.0788	0.3821	0.8237	**0.7800**
Small Models		Mirror	6.19	0.0299	0.0387	0.0068	0.0145	0.5121	0.4926
		AdaLabel	15.70	**0.2022**	**0.1824**	0.0539	0.3099	0.7929	0.7046
		CORAL$_{\text{BERT}}$ (off-policy)	10.53	0.0949	0.0724	**0.0694**	**0.4093**	0.8534	0.8108
		CORAL$_{\text{BERT}}$ (mix-policy)	12.42	0.0970	0.0676	0.0630	0.3616	**0.8787**	**0.8477**
		DailyDialog							
		Model	Avg. Len	BLEU	METEOR	Dist-1	Dist-2	MAUDE-ESIM	MAUDE-BERT
		Ground Truth	11.97	NA	NA	0.0681	0.4061	0.8180	0.8603
Large Models	Zero-shot	DialoGPT (ZS)	8.89	0.0667	0.0445	**0.0415**	**0.1682**	0.6327	0.6016
		DialoRPT (ZS)	15.94	0.0809	0.0530	0.0273	0.1192	**0.6523**	0.6055
		Blenderbot (ZS)	17.27	**0.1049**	**0.0771**	0.0223	0.0977	0.6363	**0.6322**
	Fine tuned	DialoGPT (FT)	6.15	0.0752	0.0678	**0.0673**	**0.2957**	0.7878	0.7831
		DialogRPT (FT)	16.55	0.1073	0.0978	0.0353	0.1791	0.8086	**0.8317**
		Blenderbot (FT)	30.70	**0.1139**	**0.1081**	0.0355	0.1984	**0.8222**	0.8249
Small Models		Mirror	7.93	0.0618	0.0538	0.0371	0.1485	0.6223	0.4995
		AdaLabel	11.59	0.1208	0.0881	0.0381	0.2292	0.5814	0.4890
		CORAL$_{\text{BERT}}$ (off-policy)	9.68	0.1838	0.1656	**0.0462**	**0.2902**	0.7279	0.6526
		CORAL$_{\text{BERT}}$ (mix-policy)	10.50	**0.2241**	**0.2079**	0.0428	0.2760	**0.7418**	**0.6692**

rank predicted by DialogRPT.

4.4 Evaluation Metrics

We use a standard set of referenced evaluation metrics and a recently proposed reference-free metric for automatic evaluation.

BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) measure lexical match upto $n = 4$-grams between the predicted and ground truth response. Since a valid response generated by the model can be different from the ground truth, these two metrics cannot always capture the validity of a generation.

MAUDE (Sinha et al., 2020) is a recently proposed metric that can capture the suitability between a context and any response without a ground truth reference. MAUDE is also a metric based on a response retrieval model and thus our model effectively optimizes MAUDE through the CORAL loss function. (Yeh et al., 2021) observed that MAUDE performs particularly well when the model has been trained on a dataset from a similar domain to its target application domain. Hence, we report results for two MAUDE variants based on ESIM (Chen et al., 2017; Chen and Wang, 2019) and BERT (Devlin et al., 2019) finetuned on the target dataset (DSTC7-Ubuntu or DD).

The Distinct-n (Liu et al., 2016) metric measures the diversity of n-grams in the overall set of generated responses. This number should not be too low, and an ideal target for this is the Distinct-n of the ground truth responses.

5 Results and Discussions

5.1 Automatic Evaluation (Small Scale Models)

In Table 1, we present the results of automatic evaluation metrics for response generation. Results for small scale and large scale models are shown separately. For our proposed CORAL models, we used BERT-R_3 reward function with $m = 0.4$ and $p^+ = 0.8$ for DSTC7-Ubuntu and $m = 0$ and $p^+ = 0.8$ for DD dataset. Fig. 4 in Appendix

shows sensitivity analysis for hyperparameters p_+ and m.

The **diversity** of the generated responses, in terms of number of unique unigrams and bigrams, is indicated by the dist-1 and dist-2 metrics. All baselines, except AdaLabel, have very low diversity for the generated responses. The RNN-based Mirror model mostly generates the same bland response (e.g., *"I don't know how to do that"*, *"I don't know what to suggest"*) for most contexts leading to a low diversity score. Although DialogRPT produces relevant responses (as indicated by a high MAUDE score), the response produced by these models sometimes does not change after the initial few turns, and the model repeatedly generates the same utterance for all remaining turns which leads to a poor Dist-n score. The diversity of most baselines (except AdaLabel) is quite low compared to that of CORAL models. From Table 1, it is quite clear that only the proposed CORAL-based models are able to produce responses which are both diverse (high Dist-n) and coherent to the context (high MAUDE), while most baselines have failed to maintain this consistently.

BLEU and **METEOR** have been known to not correlate strongly with human judgments on the quality of generated responses, as shown by several prior works (Liu et al., 2016; Sinha et al., 2020; Yeh et al., 2021). Even in terms of these word overlap-based metrics, we outperform most of the baselines in DD. In case of DSTC7-Ubuntu, AdaLabel beats $\text{CORAL}_{\text{ESIM}}$ and $\text{CORAL}_{\text{DMI}}$.

In terms of **average length** of the generated responses, the pretrained models outperform other baselines and also the CORAL models. Though it should be noted that high generation length does not necessarily mean higher quality responses, as also indicated by the MAUDE scores.

Finally, in terms of the **MAUDE** metric, our proposed method, CORAL, outperforms all baselines by a significant margin. The diverse and high quality responses justify the choice and design of the CORAL loss function.

5.2 Human Evaluation Study

As automatic evaluation metrics cannot capture all the nuances of how humans assess a model generated response, we also run a crowdsourced human evaluation study for various models. Three different annotators rated a context-response pair in terms of engagement, fluency and relevance on a 0-2 scale: No (0), Somewhat (1), Yes (2). Detailed annotation questionnaire is *in the appendix*. This evaluation process was run on a 100 randomly selected contexts from DailyDialog test set.

Figure 3 shows that the CORAL-based models outperform the baselines in all the three aspects. CORAL-mixp (nucleus) outperforms CORAL-offp in engagement and relevance, whereas CORAL-offp is more fluent than CORAL-mixp.

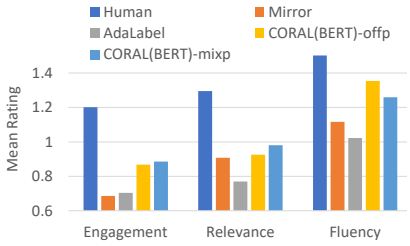


Figure 3: Human Evaluation Results - Mean ratings for Engagement, Relevance, and Fluency.

5.3 Ablation Studies

To better understand the effects of various hyperparameters (see Section 3.3) on the final trained model, we ran ablation studies on both the datasets. The complete set of results is displayed in Figure 4 and Table 3 in the appendix. All the comparisons are done based on the best average reward obtained by the model on validation set. In general, the mix-policy setup outperforms off-policy training routines. For the DailyDialog dataset, lower margin values tend to have higher R_3 scores. But, for DSTC7-Ubuntu, in case of mix-policy training, the R_3 score increases with the margin value. Off-policy training worked better with positive ground truth responses only, with RandomNegatives generally having a detrimental effect on the final reward achieved by the model.

5.4 Comparison with Large-Scale Models

Large-scale (in terms of data or model parameter size) pretrained language models have become a very prominent candidate for open-domain dialog generation models with the development of models like DialoGPT, Blender etc. Thus, we also compare our CORAL models with two variants of these large-scale models: zero-shot and finetuned versions. Because of the extensive pretraining done for training these large-scale models, we do not expect CORAL models to fully outperform these large-scale models. We observe that CORAL is able to beat all the zero-shot models for both the

Table 2: This table shows sample generations from different baseline models and our CORAL$_{\text{BERT}}$ model. Compared to Mirror, responses from our model feels like a more natural continuation to the context. The AdaLabel baseline also has a higher tendency to generate a generic/dull response to the context. A common issue that we notice in all small-scale models is their random behavior when they encounter a new concept unseen during training.

Context: I really need to start eating healthier. EOU I have to start eating better too. EOU What kind of food do you usually eat? EOU I try my best to eat only fruits, vegetables, and chicken. EOU			
Mirror: What kind of food do you want ?	**Blenderbot**: I try to eat a lot of veggies and lean meats.	**AdaLabel** i have no idea what i'm supposed to eat .	**CORAL$_{\text{BERT}}$**: I usually eat a bowl of dumplings.
Context: Room service, is there anything I can do for you? EOU Yes, I am afraid there are not enough outlets. EOU			
Mirror: Have you got any wine ?	**Blenderbot**: I'm not sure what you mean by that, but I'm sure there's something you can do.	**AdaLabel**: ok. thank you .	**CORAL$_{\text{BERT}}$**: OK, sir. Can I take your room key to the Room 20 minutes?
Context: Good evening, madam. Can I help you? EOU Yes. I bought this sheet here yesterday. EOU			
Mirror: Did it happen ?	**Blenderbot**: How much did you pay for it?? I'm a madam	**AdaLabel**: ok. when will you come back ?	**CORAL$_{\text{BERT}}$**: I'd like to buy a new dress.
Context: Could I have my bill, please? EOU Certainly, sir. EOU I'm afraid there's been a mistake. EOU I'm sorry, sir. What seems to be the trouble? EOU			
Mirror: Yes, I have been looking forward to it .	**Blenderbot**: I don't know what to do with my life. I feel like I have no purpose.	**AdaLabel** oh, i see. then, let me see. ... oh, this switch down to the shop tonight. i' ve got a lot	**CORAL$_{\text{BERT}}$** I'm sorry, sir. I've got you a wrong case. I've been waiting for the wrong medicine. I'm sorry.
Context: Are you busy tomorrow morning? EOU I'm free. What's up? EOU Someone has to pick up the boss at the airport. EOU			
Mirror: Oh, I see. Thank you .	**Blenderbot**: Yes, I have to get up early to go to work. How about you?	**AdaLabel**: i' m sorry to hear that. do you want to go to the hospital ?	**CORAL$_{\text{BERT}}$**: Yes. I'd like to. What time do you want to

datasets. Even though DailyDialog and Reddit-based pretraining datasets both comprises open-domain English conversations, CORAL$_{\text{BERT}}$ outperforms the zero-shot variants on DailyDialog and DSTC7-Ubuntu. CORAL$_{\text{BERT}}$ is even able to beat the finetuned version of DialoGPT and Blenderbot on the DSTC7-Ubuntu dataset. This proves the strength of the loss function and RL training paradigm proposed in this paper for dialog generation models.

5.5 Case Study: Generation Quality

We provide samples generated from CORAL$_{\text{BERT}}$ and some baseline models in Table 2. Since Blenderbot (Roller et al., 2021) model was trained with a much larger dialog corpus, it is able to generate good responses in many cases where small-scale models failed to understand the context leading to random utterances. Among the small scale models, CORAL$_{\text{BERT}}$ generally replies with more engaging and specific responses, whereas Mirror tends to produce generic responses most of the time. Blenderbot, however, fails to realize the switch in the turns sometimes and generates the same response repeatedly, only based on the context topic. This is the reason for low diversity scores of Blenderbot. For all small-scale models, because of limited training data, they lack knowledge of many entities that only appears in test set but not during training. In such cases, the small scale models just produce some unrelated utterance.

6 Conclusion

In this paper, we proposed CORAL, a novel loss function to circumvent these shortcomings of CE loss. Specifically, using CORAL, we can train dialog generation models without assuming a fixed ground-truth response and the value of the loss function is based on both the context and response. The CORAL loss is based on pretrained response retrieval models that, in prior literature, have been shown to correlate with human preferences. Experiments over two diverse benchmarks have shown that it comprehensively outperforms other small scale models and is even comparable to the large scale models. The proposed loss function will make it possible to train future models focused on maximizing human preference. We also hope that our work will motivate the NLP community to look for more suitable loss functions for training dialog generation models and to rely less on the cross-entropy loss.

We plan to extend this framework for training larger scale models that can capture more patterns from larger training data. We are also looking into the possibility of designing a learning curriculum for RL-based training using the mix-policy method. This will help make the training of the dialog generation model more efficient.

7 Ethical Considerations

Like many other pretrained language representation models, the proposed model may also have learned patterns associated with exposure bias. Interpretability associated with the output is rather limited, hence users should use the outputs carefully. The proposed model generates possible response candidates, and does not filter out any "problematic" candidates. Thus, for applications, where candidate responses could be problematic, (e.g., offensive, hateful, abusive, etc.), users should carefully filter them out before using the output from our model.

All the datasets used in this work are publicly available. We did not collect any new dataset as part of this work.

DailyDialog: The dataset was downloaded from http://yanran.li/dailydialog. DailyDialog dataset is licensed under CC BY-NC-SA 4.0.

DSTC7-Ubuntu: The dataset was downloaded from https://ibm.github.io/dstc-noesis/public/data_description.html#ubuntu. The dataset is available under MIT license.

References

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

Qian Chen and Wen Wang. 2019. Sequential attention-based network for noetic end-to-end response selection. *ArXiv preprint*, abs/1901.02609.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics.

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

Yao Dou, Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2021. Multitalk: A highly-branching dialog testbed for diverse conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12760–12767.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online. Association for Computational Linguistics.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2021. Internet-augmented dialogue generation. *ArXiv preprint*, abs/2107.07566.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2021. Improving response quality with backward reasoning in open-domain dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1940–1944.

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

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

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, Philadelphia,

Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ananya B Sai, Mithun Das Gupta, Mitesh M Khapra, and Mukundhan Srinivasan. 2019. Re-evaluating adem: A deeper look at scoring dialogue responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6220–6227.

Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, and Rosalind W. Picard. 2020. Hierarchical reinforcement learning for open-domain dialog. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8741–8748. AAAI Press.

Chinnadhurai Sankar and Sujith Ravi. 2019. Deep reinforcement learning for modeling chit-chat dialog with discrete attributes. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–10, Stockholm, Sweden. Association for Computational Linguistics.

Bishal Santra, Sumegh Roychowdhury, Aishik Mandal, Vasu Gurram, Atharva Naik, Manish Gupta, and Pawan Goyal. 2021. Representation learning for conversational data using discourse mutual information maximization. *arXiv preprint arXiv:2112.05787*.

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

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

Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 504–509, Vancouver, Canada. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Jasmine Wang, Ryan Lowe, William L. Hamilton, and Joelle Pineau. 2020. Learning an unreferenced metric for online dialogue evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2430–2441, Online. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yida Wang, Yinhe Zheng, Yong Jiang, and Minlie Huang. 2021. Diversifying dialog generation via adaptive label smoothing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3507–3520.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv preprint*, abs/1910.03771.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7281–7288. AAAI Press.

Yi-Ting Yeh, Maxine Eskenazi, and Shikib Mehri. 2021. A comprehensive assessment of dialog evaluation metrics. In *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*, pages 15–33.

Koichiro Yoshino, Chiori Hori, Julien Perez, Luis Fernando D'Haro, Lazaros Polymenakos, Chulaka Gunasekara, Walter S Lasecki, Jonathan K Kummerfeld, Michel Galley, Chris Brockett, et al. 2019. Dialog system technology challenge 7. *arXiv preprint arXiv:1901.03461*.

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

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. 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 1208–1218, Minneapolis, Minnesota. Association for Computational Linguistics.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

A Similarities and Differences between CORAL and CE Loss Functions

In this section, we explore the similarities and differences between the proposed CORAL loss and the CE loss function. Although CORAL is derived from quite a different viewpoint, under certain hyperparameter settings CORAL approximates a weighted version of the CE loss.

1. If we only consider positive samples as candidate responses and set the score range ($score \in [0, 1]$) and margin m ($m = 0$) such that R_3 is always greater than zero, CORAL is equivalent to a weighted version of CE.

2. Cross-entropy loss has always relied *strictly* on the positive responses in the dataset. CORAL utilizes both positive and negative response candidates.

3. Training of a dialog generation model using CE may over-weigh generic responses more than more informative ones as there is no mechanism for automatically assigning weight to different ⟨context,response⟩ pairs. CORAL has provision for assigning different weight for different ⟨context, candidate response⟩ pairs.

4. CORAL uses randomly sampled response candidates for training which allows us to utilize more samples of ⟨context,response⟩ pairs during training. This provides a richer training signal from the same dataset.

5. CE loss decomposes to a token level comparison between the predicted and the target token. Its main goal is to increase the probability of the tokens in ground truth response strictly in the given form and order. CORAL loss works quite differently as it treats responses as whole units. It will either increase or decrease probability of responses as a whole, based on their semantics and compatibility to the context.

B Limitations

We have trained a small version of the proposed CORAL model. It will be great to see if the gains due to CORAL loss lead to similar improvements for large scale models as well.

We experimented with English datasets only. While we hope that these results will generalize to models trained on multi-lingual datasets; empirical validation needs to be done.

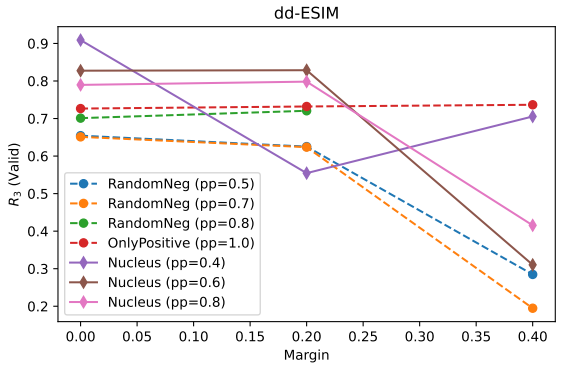
C Human Annotation Guidelines

For each of the eighteen dialog qualities, the detailed instructions and examples are shown below. These instructions were available for the worker to expand for each question.

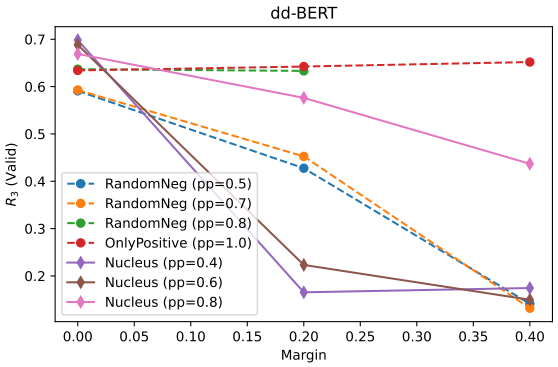
C.1 Engaging

A response is considered engaging if it can engage the user. This might be an inquisitive question or an interesting response that can be followed-up on.

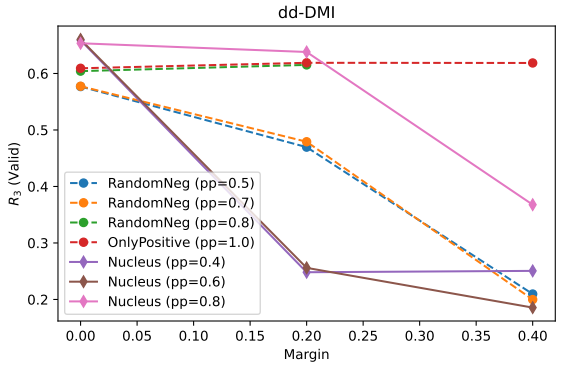
- No: the response is boring and does little to engage the user.
  - Hi there.
  - Oh wow! That's cool!



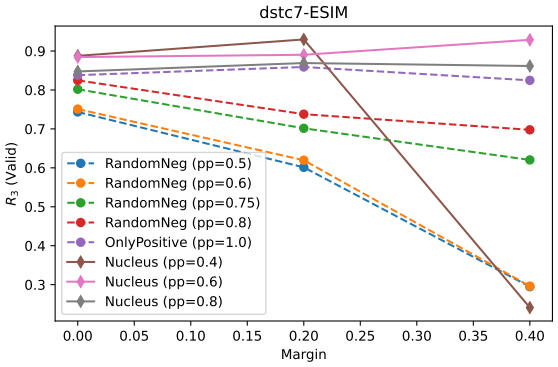
(a) CORAL$_{\text{ESIM}}$ (DailyDialog)



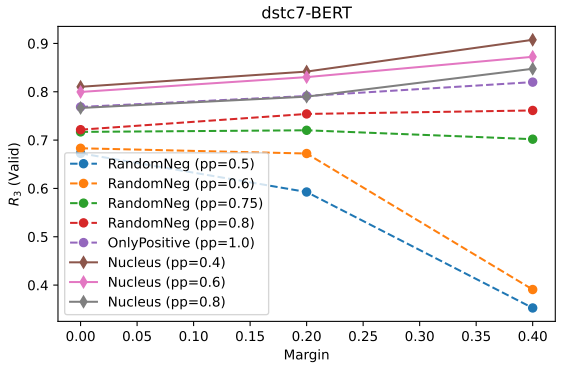
(b) CORAL$_{\text{BERT}}$ (DailyDialog)



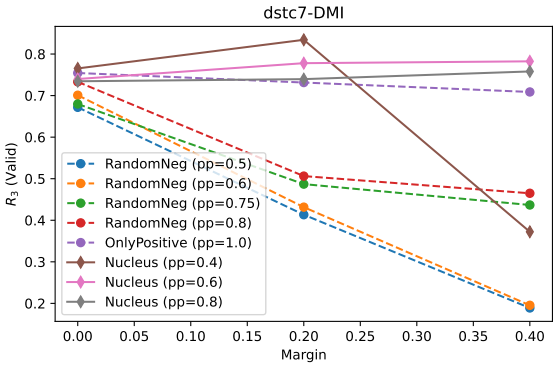
(c) CORAL$_{\text{DMI}}$ (DailyDialog)



(d) CORAL$_{\text{ESIM}}$ (DSTC7-Ubuntu)



(e) CORAL$_{\text{BERT}}$ (DSTC7-Ubuntu)



(f) CORAL$_{\text{DMI}}$ (DSTC7-Ubuntu)

Figure 4: Hyperparameter Sensitivity Analysis/Ablation Studies: These plots showcase the effect of p^+ and margin on the final validation-R_3 score obtained by the corresponding CORAL model. Each lineplot corresponds to a single p^+ value as indicated by the legend. *Note: The R_3 values are not comparable across any two plots.*

- Somewhat: the response is not particularly engaging but still leaves room for follow-up.
  - My favourite colour is blue.
  - Nope. I'm not very good with cooking.
- Yes: the response is actively engaging the user and trying to move forward the conversation.
  - I have a feeling that if you can dodge a wrench you can dodge a ball.
  - What kind of shows do you like?

C.2 Relevant

A response is considered relevant if it is on topic and a valid/reasonable continuation of the conversation. An irrelevant response is one that is completely off topic, or abruptly tries to change the topic of conversation.

- No: the response is irrelevant. It is completely off topic or abruptly tries to change the topic of conversation.
  - User: I think I'm starting to feel sick actually System: My favorite movie is Parasite
  - User: Can you play dodge ball? System: What kind of shows do you like?
- Somewhat: the response is a valid continuation of the conversation, but is somewhat generic or tangential to the main topic. For example, it begins with a generic acknowledgement of the topic then changes topics.
  - User: Have you tried seafood pancake? System: It's okay. I like beef bulgogi, but I'm not a huge fan of kimchi.
  - User: It's going great, I just finished reading '1001 inventions that changed the world' System: That sounds interesting. My favorite movie is Parasite.
- Yes: the response is relevant, on topic and a valid continuation of the conversation.
  - User: lol... wanna join us for dodge ball? System: Sounds like fun! I'm in!
  - User: I think I like most foods. I really like steak System: Steak is good. I especially love it when it's well done.

C.3 Fluent

A response is considered fluently written if it follows all the rules of English grammar and is coherent.

- No: the response is not fluently written.
  - System: Therefore you are. Still how long? I don't know if I can.
  - System: OK let's do it. Pardon me? Why it shouldn't be?
- Somewhat: the response is somewhat fluent, but makes a minor mistake or is confusing.
  - System: I am 18 of your Earth years.
  - System: Glad to hear it. What's new? Oh. Maybe you should get some rest?
- Yes: the response is written fluently.
  - System: I have a feeling that if you can dodge a wrench you can dodge a ball.
  - System: The Last Airbender, and Arrested Development are my favorites.

Table 3: Automatic evaluation results on reward-type (ESIM, BERT, DMI) based ablations for CORAL-based models. For each reward-type, we report the best results here among all the possible configuration of p^+ and m. This selection was done based on validation set.

	DSTC7-Ubuntu							
	Model	Avg. Len	BLEU	METEOR	Dist-1	Dist-2	MAUDE-ESIM	MAUDE-BERT
	Ground Truth	13.73	NA	NA	0.0922	0.5231	0.8583	0.7363
Small Models	$\text{CORAL}_{\text{ESIM}}$ (off-policy)	10.28	0.1729	0.1581	**0.0744**	**0.4242**	0.8590	0.7600
	$\text{CORAL}_{\text{BERT}}$ (off-policy)	10.53	0.0949	0.0724	0.0694	0.4093	0.8534	0.8108
	$\text{CORAL}_{\text{DMI}}$ (off-policy)	10.39	**0.2194**	**0.2111**	0.0707	0.4151	0.8274	0.7309
	$\text{CORAL}_{\text{ESIM}}$ (mix-policy)	8.54	0.1342	0.1219	0.0716	0.4038	0.8691	0.7534
	$\text{CORAL}_{\text{BERT}}$ (mix-policy)	12.42	0.0970	0.0676	0.0630	0.3616	**0.8787**	**0.8477**
	$\text{CORAL}_{\text{DMI}}$ (mix-policy)	6.93	0.0904	0.0779	0.0729	0.3848	0.8628	0.7913
	DailyDialog							
	Model	Avg. Len	BLEU	METEOR	Dist-1	Dist-2	MAUDE-ESIM	MAUDE-BERT
	Ground Truth	11.97	NA	NA	0.0681	0.4061	0.8180	0.8603
Small Models	$\text{CORAL}_{\text{ESIM}}$ (off-policy)	10.00	0.1910	0.1727	0.0459	0.2840	0.7402	0.6474
	$\text{CORAL}_{\text{BERT}}$ (off-policy)	9.68	0.1838	0.1656	**0.0462**	**0.2902**	0.7279	0.6526
	$\text{CORAL}_{\text{DMI}}$ (off-policy)	9.44	0.1727	0.1567	0.0461	0.2844	0.7229	0.6234
	$\text{CORAL}_{\text{ESIM}}$ (mix-policy)	11.01	0.2118	0.1970	0.0397	0.2541	**0.7577**	0.6588
	$\text{CORAL}_{\text{BERT}}$ (mix-policy)	10.50	0.2241	0.2079	0.0428	0.2760	0.7418	**0.6692**
	$\text{CORAL}_{\text{DMI}}$ (mix-policy)	10.91	**0.2327**	**0.2190**	0.0433	0.2635	0.7410	0.6680