# Diversifying Neural Dialogue Generation via Negative Distillation

Anonymous ACL submission

## Abstract

Generative dialogue models suffer from serious generic response problems, limiting their applications to a few toy scenarios. Recently, an interesting approach, namely negative training, has been proposed to alleviate this problem by reminding the model not to generate high-frequency responses during training. However, its performance is hindered by two issues, ignoring low-frequency but generic responses and bringing low-frequency but meaningless responses. In this paper, we propose a novel negative training paradigm, called negative distillation, to keep the model away from the undesirable generic responses while avoiding the above problems. First, we introduce a negative teacher model that can produce query-wise generic responses, and then the student model is required to maximize the distance with multi-level negative knowledge. Empirical results show that our method outperforms previous negative training methods significantly.

## 1 Introduction

In the past few years, data-driven response generation (Vougiouklis et al., 2016; Vinyals and Le, 2015; Shang et al., 2015) has achieved impressive performance, drawing continuously increasing attention from academic and industry. Conventionally, with the guidance of maximum likelihood estimation (MLE), neural dialogue models are expected to maximize the probability of generating the corresponding reference given any query. Unfortunately, due to the many-to-one phenomenon (see Table 1), a characteristic of the dialogue task (Csáky et al., 2019), these models are prone to produce safe but generic responses (e.g., *I don't know* (Li et al., 2016)), which sets an obstacle for the generative dialogue system to be deployed widely. Some researchers tried to redesign the objective of models to meet the requirement of diverse responses instead of MLE, such as MMI (Li et al., 2016), AdaLabel (Wang et al., 2021), and IAT (Zhou et al., 2021). Besides, several studies (Holtzman et al., 2020; Kulikov et al., 2019) proposed more advanced decoding strategies to alleviate the problem of generic responses. Indeed, the above methods boost the diversity of responses by reminding the model what should be said.

However, inspired by negative training (Kim et al., 2019; Ma et al., 2021), we argue that it is also necessary to tell the dialogue model what not to say. To alleviate the problem of generic responses, He and Glass (2020) negatively updates the parameters when identifying the high-frequency responses. Li et al. (2020a) punishes the behaviors of generating repetitive or high-frequency tokens by using the unlikelihood objective (Welleck et al., 2020).

Although the negative-training based methods enhance the diversity of responses, there still exists two drawbacks: First, they regard high-frequency tokens or utterances as negative candidates. However, the high-frequency response problem is only a sub-problem of the generic response problem (He and Glass, 2020). It means that the responses that are low-frequency but generic will escape from punishment. Even worse, we have observed that some generic responses are followed by a low-frequency but meaningless subsequence to avoid being identified as high-frequency, which inevitably sacrifices the fluency of responses (see Analysis). Second, these methods ignore the implicit negative knowledge in neural networks that characterizes negative candidates at multiple levels. We contend that it is more effective to conduct negative training using richer information (e.g., hierarchical representation).

To tackle the above problems and further improve the diversity of responses, we propose a novel negative training paradigm called *Negative Distillation* (ND). Conventional knowledge distillation (KD) (Hinton et al., 2015; Jiao et al., 2020) takes the teacher as a positive role model and in-

| | Query | Response | Entropy Score |
|---|---|---|---|
| 1: | Oh, honey, you made a mistake. | **I don't know** how to do it. | 8.61 |
| 2: | Would you like regular car wash package ? | **I don't know** what you mean. | 8.75 |
| 3: | I'm looking for the airport.<br>Can you tell me how to get there? | No, sorry. **I don't know**. | 10.69 |
| 4: | That's cathy. She is pretty, isn't she?<br>She likes dancing. Ask her to dance. | Well, **I don't know**. | 12.14 |
| 5: | It doesn't matter. You gotta find what she's interested in and go with that. | **I don't know** . . . | 6.82 |

Table 1: The *many-to-one* phenomenon in DailyDialog. All the above five queries have the same **I don't know**-like responses. The corresponding *source entropy* (Csáky et al., 2019) scores are much higher than the median score (0.92) of the whole training set. This phenomenon will lead to the generic response problem.

duces the student to imitate. Differing from that, we train the teacher as a negative role model and remind the student to get rid of those bad behaviors.

Specifically, we first collect a negative training set by using a filtering method called *Source Entropy* (Csáky et al., 2019). This filtering method can retrieve all *many-to-one* cases of the raw dataset. Note that the "one" is usually a generic response. Then, we train a dialogue model on the above subset as the negative teacher. Given queries, the negative teacher can provide a set of negative candidates (i.e., generic and dull responses) that the student is prone to generate, which avoids the first drawback mentioned before. Therefore, the student obtains query-wise bad behaviors for *Negative Distillation*. To conduct the negative update holistically, we design two negative objectives, including soft unlikelihood loss on the prediction layer and reverse square error on the intermediate layer. In this way, the negative distillation fully exploits multi-level negative knowledge to force the student to generate non-generic responses.

Our contributions are summarized as follows:

- We propose a novel and effective negative training paradigm called *Negative Distillation*. It constructs query-wise generic responses as the negative candidates.
- We design two negative objectives to utilize multi-level information to further boost the performance of negative distillation.
- We perform extensive experiments and detailed analysis to verify the effectiveness of the negative distillation framework and the superiority compared with previous negative training methods.

## 2 Method

In this section, we first introduce the negative teacher, then describe the negative distillation on the prediction layer and the intermediate layer, respectively, and finally present the progressive optimization objective. Algorithm 1 shows the whole training details.

### 2.1 Background

**Dialogue Generation with MLE** Take $Q = \{q_1, q_2, ..., q_{T_q}\}$ and $R = \{r_1, r_2, ..., r_{T_r}\}$ as the (*query, response*) pair, where $T_q$ and $T_r$ represent the length of query and response, respectively. The generative dialogue model aims to learn a conditional probability distribution $p_\theta(R|Q)$. The maximum likelihood estimation (MLE) is usually used to train the model, which can also be expressed as minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{MLE}} = -\sum_{i=1}^{T_r} \log p_\theta \left(r_i \mid r_{<i}, Q\right). \quad (1)$$

Considering one characteristic of the dialogue task, i.e., allowing the response to be varied, the *many-to-one* phenomenon occurs in the dialogue corpora frequently. However, with the MLE-based training, this phenomenon will cause the model to produce generic responses.

**Unlikelihood Training** Unlikelihood (UL) loss (Welleck et al., 2020) is proposed for the model to address the problem of undesirable behaviors (e.g., repetitive or high-frequency tokens). It forces the

model to minimize the probability of generating negative candidates, which is formulated as:

$$\mathcal{L}_{\mathrm{UL}} = -\sum_{i=1}^{T_r} \sum_{r_c \in \mathcal{C}_t} \log\left(1 - p_\theta\left(r_c \mid r_{<i}, Q\right)\right), \tag{2}$$

where $\mathcal{C}_t$ consists of negative candidates (e.g., overuse frequent words) that are also a sub-set of the vocabulary.

**Knowledge Distillation** The traditional knowledge distillation (KD) usually transfers useful knowledge from a large and strong teacher network $T$ to a small student network $S$. The distillation loss is used to align the soften predictions of the teacher and the student, denoted as $f^T(x)$ and $f^S(x)$:

$$\mathcal{L}_{\mathrm{KD}} = \sum_{x \in \mathcal{D}} L\left(f^T(x), f^S(x)\right), \tag{3}$$

where $L(\cdot)$ is a measurement function that calculates the distance of different probability distributions, $x$ is the input text, and $\mathcal{D}$ denotes the training set.

In this work, we replace the positive teacher in vanilla KD with a negative teacher, aiming to provide negative knowledge for the student to conduct negative training and avoid undesirable behaviors.

## 2.2 Negative Teacher

To improve the diversity of responses, the dialogue model should be told which responses are generic. For negative distillation, a negative teacher is required to produce possible generic responses given any query. In this work, we adopt the widely used Transformer (Vaswani et al., 2017) as the underlying model for both teacher and student. We introduce the *Source Entropy* filtering method (Csáky et al., 2019) to identify and collect the many-to-one cases for the negative training set. The source entropy is defined as:

$$H_{src}(r, \mathcal{D}) = -\sum_{(q_i, r) \in \mathcal{D}} p(q_i|r) \log p(q_i|r), \tag{4}$$

where $p(q_i|r)$ is the conditional probability calculated based on the relative frequency of (*query, response*) pairs, $r$ is a response, $q_i$ is the query corresponding to the response $r$, and $\mathcal{D}$ represents the raw training set. A higher source entropy indicates that the response $r$ corresponds to more queries, i.e., the *many-to-one* problem is serious. We select the top 50% dialogue pairs $(q, r)$ with a high source entropy as the negative training set $\mathcal{D}_N$, which contains a much higher proportion of generic responses than the raw training set.

After that, we train the teacher $N$ on the negative training set $\mathcal{D}_N$ by Equation 1. The teacher will naturally produce generic responses for any input query. More importantly, it will provide richer negative knowledge for the student, including soft logits in the prediction layer and implicit features in the intermediate layers.

## 2.3 Negative Distillation

In this section, we conduct the negative distillation for the student based on the multi-level negative knowledge.

**ND for Prediction Layer** The soften logits in the prediction layer contain more information than the ground-truth labels, such as the similarity between labels (Wang et al., 2021). Therefore, conventional KD transfers knowledge by narrowing the gap between the probability distributions of the teacher $T$ and the student $S$:

$$\mathcal{L}_{KD} = -\sum_{i=1}^{T_r} \sum_{k=1}^{|\mathcal{V}|} p_T\left(r_i = k \mid r_{<i}, Q\right) \cdot \log p_S\left(r_i = k \mid r_{<i}, Q\right). \tag{5}$$

As for negative distillation, the extra knowledge in soften logits of the negative teacher reflects how to generate dull responses based on the input query. Therefore, we propose a soft unlikelihood loss to maximizing the distance between the predictions of the negative teacher $N$ and the student $S$:

$$\mathcal{L}_{pred} = -\sum_{i=1}^{T_r} \sum_{k=1}^{|\mathcal{V}|} p_N\left(r_i = k \mid r_{<i}, Q\right) \cdot \log\left(1 - p_S\left(r_i = k \mid r_{<i}, Q\right)\right), \tag{6}$$

where $p_N$ and $p_S$ are calculated by:

$$p^i = \frac{\exp\left(z_i/t\right)}{\sum_j \exp\left(z_j/t\right)}, \tag{7}$$

where $t$ is a temperature coefficient that is used to soften the probability distribution over words.

It should be emphasized that previous negative training methods only use the high-frequency words or phrases with one-hot representation as the targets, which ignores the rich information existing in the soften logits (e.g., the generic words

have similar probabilities). In the Analysis section, we demonstrates the superiority of soften logits compared with hard targets (i.e., one-hot representation).

**ND for Intermediate Layer** In addition to the output knowledge from the prediction layer, there is also some implicit knowledge embedded in the intermediate layers, such as hidden states and attention matrices. To keep the student away from undesirable behaviors (i.e., producing generic responses) more effectively, we further consider the above knowledge into negative distillation. Specifically, the distance between features of the negative teacher and the student should also be increased. In this work, we propose a new measurement function, called mean reverse square error (MRSE), to calculate this distance:

$$\mathcal{L}_{MRSE}(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{n}\sum_{i=1}^{n} \exp^{-SE(\boldsymbol{A}_i, \boldsymbol{B}_i)}, \quad (8)$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are the feature matrices of the negative teacher and the student, respectively, and $n$ is the number of elements of each matrix.

Due to the responses generating in the decoding phrase, we only conduct negative distillation on the intermediate layers of the decoder. For each decoder layer, the negative distillation objective of hidden states is defined as:

$$\mathcal{L}^l_{hid} = \mathcal{L}_{MRSE}(\boldsymbol{H}^l_N, \boldsymbol{H}^l_S), \quad (9)$$

where $\boldsymbol{H}^l_N$ and $\boldsymbol{H}^l_S$ are the output hidden states of the $l^{th}$ decode layer of $N$ and $S$, respectively.

As the attention weights can learn substantial linguistic knowledge (Clark et al., 2019), it is beneficial for the student to further conduct negative distillation on the attention matrices, which is computed as follows:

$$\boldsymbol{A} = \frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}, \quad (10)$$

$$\text{Attention}\,(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\boldsymbol{A})\boldsymbol{V}, \quad (11)$$

where $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the matrices of queries, keys, and values, respectively, and $d_k$ is a scaling factor. Following Jiao et al. (2020), the attention matrix $\boldsymbol{A}$ is chosen to calculate the distance rather than its softmax version $\text{softmax}(\boldsymbol{A})$. Similar to Equation 9, the negative distillation objective of attention matrices is formulated as:

$$\mathcal{L}^l_{att} = \mathcal{L}_{MRSE}(\boldsymbol{A}^l_N, \boldsymbol{A}^l_S)), \quad (12)$$

where $\boldsymbol{A}^l_N$ and $\boldsymbol{A}^l_S$ are the attention matrices of the $l^{th}$ decoder layer of $N$ and $S$, respectively.

**Algorithm 1** Negative Distillation

**Input:** $\mathcal{D}$: The raw training set; $H_{src}$ : The *Source Entropy* filtering method; $N$ and $S$: The negative teacher and the student.

```
% Collection of negative training set.
[Data_entropy] ← Calculate_data_entropy(D, H_src) using Eq.4
Index_list ← Sort([Data_entropy])
D_N ← Extract_top_data(D, Index_list, 50%)
% Training of negative teacher.
repeat
    Optimize N by minimizing L_mle(N) on D_N using Eq. 1
until Convergence
% Negative distillation.
repeat
    Optimize S by minimizing L(S) on D using Eq. 13
until Convergence
```

**Output:** $S$ : The trained student.

## 2.4 Progressive Optimization

The overall loss, combining the above negative distillation objectives and the MLE objective, is denoted as:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{mle} + \alpha(\mathcal{L}_{pred} + \sum^{l} \mathcal{L}^l_{hid} + \sum^{l} \mathcal{L}^l_{att}), \quad (13)$$

where $\alpha$ is a hyper-parameter that balances the importance of supervised learning and negative distillation. For negative distillation, it would be better that the student has the ability to say something before it is reminded of what not to say. Thus, we perform a progressive distillation that first warms up the negative distillation ratio and then colds it down gradually. Inspired by the derivative of sigmoid function:

$$\sigma'(z) = \sigma(z)(1-\sigma(z)) = \frac{e^{-z}}{(e^{-z}+1)^2}, \quad (14)$$

which shows a trend of gradual rise-fall, we define the balance coefficient $\alpha$ as:

$$\alpha = \lambda * \frac{e^{-z}}{(e^{-z}+1)^2}, \quad (15)$$

where $\lambda$ controls the peak value and $z$ is calculated by:

$$z(s) = \beta * (s - \gamma), \quad (16)$$

| Datasets | Train | Valid | Test | Vocab |
|---|---|---|---|---|
| DailyDialog | 68k | 6.8k | 6.8k | 17,930 |
| OpenSubtitles | 200k | 20k | 10k | 21,177 |

Table 2: Statistics of two dialogue datasets in the experiments.

where $s$ is the training step, and $\beta$ and $\gamma$ control the telescopic and translation transformation, respectively.

# 3 Experiments

## 3.1 Datasets

In our experiments, two widely used dialogue datasets are employed to evaluate the proposed method: **DailyDialog**, which collects conversations that are similar to human daily communication (Li et al., 2017b), and **OpenSubtitles**, which consists of large-scale dialogues extracted from movie subtitles (Tiedemann, 2009). In this work, we focus on the single-turn dialogue generation, thus we pre-process these two datasets into the (query, response) pairs. Table 2 provides the statistics of both datasets.

## 3.2 Experimental Settings

We take the Transformer-based sequence-to-sequence model (Vaswani et al., 2017) as the underlying model for all approaches. Following the settings of Transformer in Csáky et al. (2019), both encoder and decoder contain 6 layers, in which the self-attention module has 8 attention heads and the number of feed-forward units is 2048. The size of hidden states is set to 512 and the dimension is 64 for query, key, and value. Dropout (Srivastava et al., 2014) is used for the self-attention module, the feed-forward layer, and the activation layer, and the rate of all three is set to 0.1. We also use label smoothing (Szegedy et al., 2016) and the smoothing value is 0.1. The batch size is set to 256. We use the Adam optimizer (Kingma and Ba, 2015) and employ the *warm-up* (He et al., 2016) trick to adjust the learning rate during training. The warm-up steps $s_{\text{wp}}$ are 128000 and 256000 for DailyDialog and OpenSubtitles, respectively. The learning rate is computed as follows:

$$lr = \frac{2 \cdot \min(\frac{1}{\sqrt{s}}, \frac{s}{\sqrt{s_{\text{wp}}^3}})}{\sqrt{d_{\text{model}}}}, \quad (17)$$

where $lr$ is the learning rate at the $s^{th}$ step of training and $d_{\text{model}}$ is the size of hidden states. We implement all approaches with Pytorch 1.7, and conduct all experiments on RTX 3090.

For the proposed approach, both the negative teacher network and the student network have the same settings in terms of the network architecture and hyper-parameters. $\lambda$ in Equation 15 is set to 4, making the peak value equal to 1. $\gamma$ is 25600 and $\beta$ is $6/\gamma$. For the temperature coefficient $t$, we simply set it to 1.

## 3.3 Baselines

We compare the proposed negative distillation (**ND**) approach with the standard Transformer and two existing negative training approaches:

- **Standard** The vanilla Transformer-based sequence-to-sequence model with the MLE-based training (i.e., the cross-entropy based loss).
- **NT** (Negative Training) (He and Glass, 2020) During training, it first counts the frequency of all generated utterances and then conducts the negative update based on the high-frequency utterances.
- **UL** (Unlikelihood Training) (Li et al., 2020a) Different from **NT**, it calculates the frequency of all generated words instead of utterances and penalizes the high-frequency words by introducing an unlikelihood loss term.

All the baselines are performed with the same architecture and hyper-parameters as ours. For **NT**, the threshold $r_{\text{thres}}$ is set to 1% and the weight coefficient $\lambda_{\text{POS}}$ is set to 1 as the authors' suggestion. For **UL**, we search the mixing hyper-parameter $\alpha$ in $[1, 10, 100, 1000]$ and 1000 is selected for its best performance. Both **NT** and **UL** are refined on the well-trained **Standard** model. Following He and Glass (2020); Li et al. (2020a), we use greedy search as the decoding strategy for all baselines and our method. We also evaluate the performance with beam search (size 5) and obtain similar results (see 3.6 for details).

## 3.4 Automatic Evaluation

**Metrics** To evaluate whether negative distillation can effectively reduce the generic responses, we adopt **Dist-{1,2,3}** (distinct) (Li et al., 2016) to reflect the lexical diversity of the generated responses. It is a widely used metric that counts the proportion of unique unigrams/bigrams/trigrams. **LF** (low-frequency token ratio) (Li et al., 2020b) further

| Models | Dist-1 ↑ | Dist-2 ↑ | Dist-3 ↑ | LF ↑ | KL-1 ↓ | KL-2 ↓ | BLEU-3 ↑ | BLEU-4 ↑ |
|---|---|---|---|---|---|---|---|---|
| Standard | 0.0089 | 0.0313 | 0.0576 | 0.102 | 0.96 | 0.53 | **0.384** | **0.395** |
| NT | 0.0059 | 0.0293 | 0.0760 | 0.070 | 1.05 | 1.84 | 0.226 | 0.183 |
| UL | 0.0062 | 0.0319 | 0.0882 | 0.075 | 1.07 | 1.88 | 0.228 | 0.187 |
| ND | **0.0145** | **0.0678** | **0.1447** | **0.158** | **0.65** | **0.26** | 0.381 | 0.388 |
| Standard | 0.0020 | 0.0071 | 0.0147 | 0.022 | 2.19 | 1.40 | **0.355** | 0.353 |
| NT | 0.0011 | 0.0045 | 0.0108 | 0.014 | 2.26 | 2.39 | 0.255 | 0.216 |
| UL | 0.0015 | 0.0060 | 0.0151 | 0.018 | **1.85** | 1.97 | 0.303 | 0.269 |
| ND | **0.0027** | **0.0102** | **0.0218** | **0.029** | 2.10 | **1.23** | **0.355** | **0.355** |

Table 3: Automatic evaluation results using greedy search on DailyDialog (Up) and OpenSubtitles (Down). The best/second-best results are **bold**/underlined. "↑" means higher is better. "↓" means lower is better.

| vs. Models | Informativeness | | | Kappa | Relevance | | | Kappa | Fluency | | | Kappa |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Win(%) | Tie(%) | Lose(%) | | Win(%) | Tie(%) | Lose(%) | | Win(%) | Tie(%) | Lose(%) | |
| Standard | 77.3 | 18.7 | 4.0 | 0.456 | 48.0 | 34.7 | 17.3 | 0.453 | 14.7 | 76.0 | 9.3 | 0.491 |
| NT | 31.3 | 38.7 | 30.0 | 0.669 | 54.7 | 32.0 | 13.3 | 0.421 | 91.3 | 8.0 | 0.7 | 0.497 |
| UL | 44.7 | 28.7 | 26.7 | 0.411 | 66.0 | 23.3 | 10.7 | 0.425 | 92.7 | 7.3 | 0.0 | 0.614 |

Table 4: Results of human evaluations on DailyDialog. Our framework has a higher win rate than baselines.

measures the diversity of responses by calculating the ratio of low-frequency words in the generated responses. The threshold of low frequency is set to 100. Besides, it is necessary to verify whether the models can ensure consistency while improving diversity. So we use **KL-{1,2}** (KL divergence) (Csáky et al., 2019), which measures the distribution distance between the generated and the ground-truth responses, to reflect how well a model can approximate the ground-truth unigrams/bigrams distributions. **BLEU** (Chen and Cherry, 2014) is also reported and it measures n-gram overlap between the generated and the ground-truth references.

**Results** Table 3 shows the results obtained at the lowest point of the validation loss. We can see that our approach outperforms all baselines in diversity (**Dist** and **LF**) by a significant margin on both datasets, demonstrating that **ND** can effectively alleviate the generic response problem by using multi-level negative information. The **KL** and **BLEU** scores of **ND** are close to or better than **Standard**, which verifies that our method can maintain the consistency of responses while improving its diversity. To some extent, both **NT** and **UL** improve the diversity of words, especially for trigrams, but the low **LF** scores indicate that they reduce the high-frequency words but fail to increase the number of low-frequency's. What's worse, their **BLEU** and **KL-2** scores sharply decline. It suggests that previous negative training approaches may harm the consistency and fluency of responses dramatically, which is not in line with the goals of the dialogue system. Our method obtains similar results with beam search. Please refer to 3.6 for details.

### 3.5 Human Evaluation

Apart from automatic evaluations, we conduct human evaluations to further verify the effectiveness of our method. We randomly select 50 samples from the test set of DailyDialog, and three well-educated annotators are invited to judge which of the responses generated by **ND** and baselines is better (i.e., win, tie or loss) in terms of informativeness, relevance, and fluency. Informativeness reflects how much the information related to the query is contained in the generated response. Relevance reflects how likely the generated response is coherent to its query. Fluency reflects how likely the generated response is produced by human.

Table 4 summarizes the human evaluation results. We can see that the proposed approach is overall better than all baselines. Specifically, **ND** achieves better performance than **Standard** in terms of informativeness and relevance, and remains competitive in fluency. Compared with both **NT** and **UL**, our approach shows significant advantages, especially in fluency. It indicates that their punishment for high-frequency tokens or utterances will lead to a serious non-fluency and inconsistency problem. We use Fleiss's kappa (Fleiss, 1971) to measure the inter-annotator agreement.

### 3.6 Experimental Analysis

We conduct extensive analysis on DailyDialog to investigate the effectiveness of the negative distillation in more details.

| Models | Dist-2 | Dist-3 | LF | KL-2 | BLEU-4 |
|---|---|---|---|---|---|
| ND | .0678 | .1447 | .158 | .26 | .388 |
| w/o $\mathcal{L}_{pred}$ | .0529 | .1084 | .145 | .39 | .397 |
| w/o $\mathcal{L}_{att}$ | .0517 | .1032 | .138 | .26 | .392 |
| w/o $\mathcal{L}_{hid}$ | .0365 | .0677 | .109 | .62 | .380 |
| w/o $\mathcal{L}_{neg}$ | .0313 | .0576 | .102 | .53 | .395 |

Table 5: Ablation studies of different negative distillation objectives in **ND**.

**Ablation Study** We study the effects of different negative distillation objectives by ablating the prediction layer distillation (w/o $\mathcal{L}_{pred}$), the attention distillation (w/o $\mathcal{L}_{att}$), the hidden state distillation (w/o $\mathcal{L}_{hid}$), and the whole negative distillation (w/o $\mathcal{L}_{neg}$, i.e. Standard). The results in Table 5 show that all three proposed negative distillation objectives are useful for improving the diversity. The significant decline in w/o $\mathcal{L}_{hid}$ indicates that the negative information in intermediate layers is very important for **ND**. w/o $\mathcal{L}_{att}$ is better than w/o $\mathcal{L}_{hid}$, attributing to the more abundant information in hidden states.

**Does *source entropy* work?** To verify whether the *source entropy* filtering method can collect the generic responses, we select the top 50% and the bottom 50% of the sorted training set as $\mathcal{D}_t$ and $\mathcal{D}_b$, respectively. Then we train $N_t$ and $N_b$ on the corresponding sub-sets. From Table 6, we can see that $N_b$ outperforms $N_t$ in all the diversity-related metrics, indicating the effectiveness of *source entropy*.

| Models | Dist-1 | Dist-2 | Dist-3 | LF |
|---|---|---|---|---|
| $N_t$ | 0.0024 | 0.0078 | 0.0134 | 0.0331 |
| $N_b$ | **0.0040** | **0.0121** | **0.0215** | **0.0444** |

Table 6: Effect of the *source entropy* filtering method.

**Can the negative knowledge be transferred?** We take $N_t$ and $N_b$ as the negative teachers for the students $S_t$ and $S_b$, respectively. Then we conduct negative distillation on both $S_t$ and $S_b$. The results in Table 7 demonstrate that $S_t$ obtains more gains in diversity than $S_b$, indicating $S_t$ gets rid of more negative knowledge. It can be further verified by the results of previous analysis that $N_t$ has more negative knowledge than $N_b$.

| Models | Dist-2 | Dist-3 | LF | KL-2 | BLEU-4 |
|---|---|---|---|---|---|
| $S_t$ | **0.0678** | **0.1447** | **0.158** | **0.26** | **0.388** |
| $S_b$ | 0.0409 | 0.0844 | 0.097 | 0.40 | 0.386 |

Table 7: Effect of negative knowledge.

| Models | Dist-2 | Dist-3 | LF | KL-2 | BLEU-4 |
|---|---|---|---|---|---|
| Standard | .0313 | .0576 | .102 | .53 | .395 |
| ND (fixed $\alpha$) | .0392 | .0793 | .123 | .42 | .386 |
| ND | .0678 | .1447 | .158 | .26 | .388 |

Table 8: Effect of progressive distillation.

| Models | Dist-1 | Dist-2 | Dist-3 | LF |
|---|---|---|---|---|
| ND (random target) | 0.0040 | 0.0109 | 0.0170 | 0.053 |
| ND (hard target) | 0.0136 | 0.0620 | 0.1344 | 0.139 |
| ND (soft target) | 0.0145 | 0.0678 | 0.1447 | 0.158 |

Table 9: Comparison of soft targets, hard targets, and random targets for negative distillation.

**Study of soft target** To evaluate the superiority of soft targets for negative distillation, we sample responses (i.e., hard target) by greedy search on the predictions of negative teachers for comparison. The results in Table 9 show that ND with soft targets can diversify the responses more effectively, demonstrating the advantages of richer negative information (e.g., the similarity between labels) in soft targets. What's more, we randomly select responses from the negative training set $\mathcal{D}_N$ as negative targets. The sharp decline in performance proves that the negative teacher can produce targeted generic responses.

**Effect of progressive distillation** In order to verify the effectiveness of progressive negative distillation, we conduct negative distillation with fixed $\alpha$. The value is obtained by calculating the average of $\alpha$ Equation 15 integral of Equation 15 divided by the convergence steps. The results in Table 8 demonstrate that the progressive distillation policy can help the student exploit negative knowledge more effectively. Besides, note that ND with fixed $\alpha$ also outperforms the Standard model.

**Is ND adapted for beam search?** He and Glass (2020) and Li et al. (2020a) choose greedy decoding due to its simplicity and higher diversity than

| Models | Dist-1 ↑ | Dist-2 ↑ | Dist-3 ↑ | LF ↑ | KL-1 ↓ | KL-2 ↓ | BLEU-3 ↑ | BLEU-4 ↑ |
|---|---|---|---|---|---|---|---|---|
| Standard | 0.0060 | 0.0238 | 0.0455 | 0.068 | 0.92 | 0.62 | 0.375 | 0.372 |
| NT | 0.0077 | 0.0326 | 0.0640 | 0.083 | 0.76 | 0.70 | 0.349 | 0.334 |
| UL | 0.0059 | 0.0270 | 0.0570 | 0.075 | 0.82 | 0.95 | 0.321 | 0.294 |
| ND | **0.0108** | **0.0427** | **0.0799** | **0.120** | **0.72** | **0.05** | **0.386** | **0.404** |

Table 10: Automatic evaluation results using beam search (size 5) on DailyDialog. Our approach (**ND**) outperform baselines on all the metrics.

beam decoding. However, we find that both **NT** and **UL** tend to generate long but non-fluent and incoherent responses. So we conduct beam search with adding the length penalty. Table 10 summarizes the results and it shows that both two baselines get better **KL** and **BLEU** scores than using greedy search. **ND** outperform baselines on all the metrics, confirming the effectiveness of our method.

| |
|---|
| **Input:** The phone turns off by itself. |
| **Standard:** What's the problem?<br>**NT:** I see. Is a part of the most important things. And I'll be happy to keep you there. Is the most important thing for you...<br>**UL:** I see. Is not allowed to be a good thing. Birthday. Birthday. - hour are very important.....<br>**ND:** I am calling to tell you that he is in the right desk. |
| **Input:** Well, I bought this skirt here yesterday afternoon. But I got the wrong size. It's too small for my daughter. I wonder if I could refund it. |
| **Standard:** That's fine.<br>**NT:** That's fine. Your mind and address number. number. And date are available. And so on. Is very simple.....<br>**UL:** I'm sorry, sir. But you've got to work overtime before you leave the contract. Service is very important. Service. Service usually be late.<br>**ND:** I think you have to pay the money. |

Table 11: Examples of generated responses.

**Case Study** Table 11 shows some cases generated by the proposed method and baselines. **Standard** prefers generic and meaningless responses. Both **NT** and **UL** tend to generate a short generic sentence followed by a incoherent and non-fluent subsequence. In contrast, **ND** can produce diverse and coherent responses.

## 4 Related work

**Diversity Dialogue Learning** There are two lines of work for solving the generic response problem: One line promotes the diversity from positive view, which is outside of our work. Specially, previous work includes MMI (Li et al., 2016), GAN (Li et al., 2017a; Zhang et al., 2018), CVAE (Zhao et al., 2017), BT (Su et al., 2020), AdaLabel (Wang et al., 2021), IAT (Zhou et al., 2021) and Nucleus Sampling (Holtzman et al., 2020). The other line alleviates the generic response problem using negative training. He and Glass (2020) regards frequent response problem as a sub-problem of the generic response problem and conduct negative update for the high-frequency responses during training. Li et al. (2020a) focuses on high-frequency tokens rather than tokens and punishes them by using the unlikelihood objective (Welleck et al., 2020). Both of them handle the generic response problem only from the angle of reducing frequency, thus can not capture all the features of generic replies.

**Negative Training for Dialogue Learning** Negative training for retrieval-based dialogue learning has been previously extensively studied (Humeau et al., 2020; Nugmanova et al., 2019), while we focus on the dialogue generation in this work. He and Glass (2020) uses negative training to prevent generic and malicious responses in dialogue models. Li et al. (2020a) generalizes unlikelihood to dialogue generation for improving repetition, specificity and coherence. Lagutin et al. (2021) proposes implicit unlikelihood training to minimizing repetition. Our work proposes a new negative training paradigm aimed at improving the diversity of dialogue responses while avoiding the problem of poor consistency and fluency of previous work.

## 5 Conclusion

We present a novel negative training paradigm to improve the diversity of dialogue responses. It formulates the conventional negative training as a knowledge distillation process, which is rarely explored before. The negative teacher can produce the corresponding generic and dull responses given any query, which naturally avoids problems that hinder previous negative training methods. Besides, we further boost the performance of negative distillation by exploiting richer information, i.e., multi-level features. Extensive experiments validate the superiority of our proposed method compared with prior negative training work.

## References


Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Richárd Csáky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5650–5669, Florence, Italy. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Tianxing He and James Glass. 2020. Negative training for neural dialogue response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2044–2058, Online. Association for Computational Linguistics.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

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, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. 2019. NLNL: negative learning for noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 101–110. IEEE.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ilia Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.

Evgeny Lagutin, Daniil Gavrilov, and Pavel Kalaidin. 2021. Implicit unlikelihood training: Improving neural text generation with reinforcement learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1432–1441, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017a. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020a. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. 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.

Zuchao Li, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, Zhuosheng Zhang, and Hai Zhao. 2020b. Data-dependent gaussian prior objective for

language generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuanjing Huang, and Yaqian Zhou. 2021. SENT: Sentence-level distant relation extraction via negative training. 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 6201–6213, Online. Association for Computational Linguistics.

Aigul Nugmanova, Andrei Smirnov, Galina Lavrentyeva, and Irina Chernykh. 2019. Strategy of the negative sampling for training retrieval-based dialogue systems. In *IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019, Kyoto, Japan, March 11-15, 2019*, pages 844–848. IEEE.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7087–7097, Online. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society.

Jörg Tiedemann. 2009. *News from OPUS—A Collection of Multilingual Parallel Corpora with Tools and Interfaces*.

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 *NIPS*, pages 5998–6008.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. In *ICML Deep Learning Workshop*.

Pavlos Vougiouklis, Jonathon Hare, and Elena Simperl. 2016. A neural network approach for knowledge-driven response generation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3370–3380, Osaka, Japan. The COLING 2016 Organizing Committee.

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, Online. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1815–1825.

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.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2021. Learning from perturbations: Diverse and informative dialogue generation with inverse adversarial training. 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 694–703, Online. Association for Computational Linguistics.