Dialogue Distillation: Open-domain Dialogue Augmentation Using Unpaired Data

Rongsheng Zhang[1][*]**, Yinhe Zheng**[2,3][*]**, Jianzhi Shao**[4][†]**, Xiaoxi Mao**[1]**, Yadong Xi**[1]**, Minlie Huang**[2][‡]

[1] Fuxi AI Lab, NetEase Inc., Hangzhou, China

[2] Department of Computer Science and Technology, Institute for Artifical Intelligence, State Key Lab of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China.

[3] Samsung Research China - Beijing (SRC-B), Beijing, China [4] Alibaba Group, Hangzhou, China

zhangrongsheng@corp.netease.com, yh.zheng@samsung.com,
sjz250796@alibaba-inc.com, aihuang@tsinghua.edu.cn

Abstract

Recent advances in open-domain dialogue systems rely on the success of neural models that are trained on large-scale data. However, collecting large-scale dialogue data is usually time-consuming and labor-intensive. To address this data dilemma, we propose a novel data augmentation method for training open-domain dialogue models by utilizing unpaired data. Specifically, a data-level distillation process is first proposed to construct augmented dialogues where both post and response are retrieved from the unpaired data. A ranking module is employed to filter out low-quality dialogues. Further, a model-level distillation process is employed to distill a teacher model trained on high-quality paired data to augmented dialogue pairs, thereby preventing dialogue models from being affected by the noise in the augmented data. Automatic and manual evaluation indicates that our method can produce high-quality dialogue pairs with diverse contents, and the proposed data-level and model-level dialogue distillation can improve the performance of competitive baselines.

1 Introduction

Open-domain dialogue systems have attracted much research attention (Shum et al., 2018; Huang et al., 2020), thanks to the success of neural generation models trained with large-scale data. Existing research has been endeavored to address various aspects in dialogue systems, such as modeling persona (Qian et al., 2018; Zheng et al., 2019; Zhang et al., 2018), expressing emotion (Zhou et al., 2018a), or generating knowledge-grounded dialogues (Ghazvininejad et al., 2018; Zhou et al., 2018b, 2020).
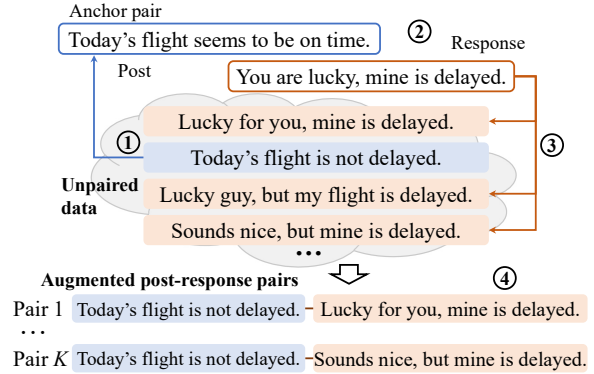

Figure 1: Process of constructing augmented post-response pairs. The sentence in blue rectangle is used to match the anchor pair and the corresponding response is then used to retrieve similar sentences in unpaired data. Each augmented pair contains two sentences both from unpaired data.

In general, training neural open-domain dialogue models requires a large amount of high-quality *paired data*, e.g., post-response pairs, which are usually labor-intensive and time consuming to collect. A feasible solution to this data dilemma is to use data augmentation techniques, which are popular in various research areas such as computer vision (Cubuk et al., 2019) or machine translation (Sennrich et al., 2016). Nevertheless, this technique is rarely investigated in the study of open-domain dialogues, and few existing approaches are specifically designed for either the generation-based dialogue models (Li et al., 2019) or the retrieval-based dialogue models (Du and Black, 2018). Moreover, existing data augmentation approaches only take a set of paired data as input without considering to utilize unpaired data.

As a matter of fact, high-quality *unpaired data*, i.e., non-conversational texts, are generally easier to collect compared to high-quality dialogue pairs. Specifically, these unpaired data provide us a rich

---

[*] Equal contribution. Order determined by swapping the one in Zheng et al. (2020b)

[†] Work performed while at Fuxi AI Lab, NetEase Inc.

[‡] Corresponding Author

bank of alternative expressions for different contents. It is thus feasible to augment the training dialogue pairs utilizing sentences extracted from the unpaired data. As shown in Figure 1, we can extract various sentences from the unpaired data that are similar to a given post-response pair (i.e., anchor pair). Augmented pairs that carry richer expressions can be then constructed by combining these extracted sentences. To the best of our knowledge, there are no previous studies for open-domain dialogues that try to construct augmented dialogue pairs by utilizing retrieved unpaired data.

In this paper, we propose a novel data augmentation method "**Dialogue Distillation**" to improve the performance of open-domain dialogue models by utilizing unpaired data. Our method involves two phases of distillation. The first phase is at the data level as it constructs (i.e., distills) post-response pairs by matching sentences retrieved from a set of unpaired data. Specifically, given a set of training pairs $\{\langle x_i, y_i \rangle\}$, a randomly selected sentence s is firstly used as a query to retrieve the most similar x_i, and then the corresponding y_i are used as queries to retrieve similar s_i from the unpaired data. Augmented pairs $\langle s, s_i \rangle$ are then constructed and filtered using a ranking module. Note that different from previous approaches, *the post and response sentences that constitute an augmented pair are both from the unpaired data*, which are human written and thereby fluent and content-rich. The second phase is at the model-level as it distills a teacher model using the augmented data. Specifically, we borrow the idea of knowledge distillation (Hinton et al., 2015) to first train a teacher model on a set of high-quality dialogue pairs, and then distill the dialogue model by mimicking the distribution produced by the teacher model on the augmented data to prevent the final dialogue models from being affected by the noise in the augmented data.

Automatic and manual evaluation results indicate that our data-level distillation process can produce high-quality post-response pairs that are content-rich, and our model-level distillation process can better utilize these augmented data to improve the performance of both retrieval-based and generation-based open-domain dialogue models.

Our contributions are summarized as follows:

1) We propose a data-level and model-level distillation method for open-domain dialogue models. The data-level distillation constructs new post-response pairs where both post and response are retrieved from unpaired data, and the model-level distillation distills a teacher model trained on high quality paired data to augmented pairs. To the best of our knowledge, this is the first attempt to augment open-domain dialogue pairs by utilizing the retrieved unpaired data.

2) Automatic and manual evaluation shows that the augmented pairs produced by our method are content-rich, and these augmented data can be used to improve the performance of both generation-based and retrieval-based dialogue models.

2 Related Work

There are two major categories of open-domain dialogue models: 1) retrieval-based models, which retrieve the best matching response from the pre-collected dialogues (Lu and Li, 2013); and 2) generation-based models, which decode responses from a learned distribution (Sutskever et al., 2014; Vinyals and Le, 2015). Recent advances in these two categories all focus on DNN-based data-driven methods (Huang et al., 2020).

Data augmentation is an effective approach to boost the performance of neural models. It has been explored in various NLP tasks, such as text classification (Wei and Zou, 2019; Zheng et al., 2020a), machine reading comprehension (Yu et al., 2018) and machine translation (Sennrich et al., 2016). Although proved to be effective, this technique is rarely investigated in open-domain dialogue models. Few existing approaches are restricted to only take the dialogue pairs as their inputs (Li et al., 2019; Zhao et al., 2017), whereas unpaired texts, i.e., sentences without replies, are not utilized.

Note that the pre-training based methods (Devlin et al., 2019; Radford et al., 2019; Golovanov et al., 2019; Zheng et al., 2020b) share a similar motivation with our study, i.e., to boost the performance of neural NLP models utilizing unlabeled (i.e., unpaired) texts. Nevertheless, the data augmentation method proposed in our study can be regarded as a supplement to these pre-training approaches. Experiments demonstrate that our method can be used to improve the performance of dialogue models even if these models are initialized with strong pre-trained models.

Our study is also related to the knowledge distillation method (Hinton et al., 2015), which also employs a teacher model and tries to minimize the KL divergence between the teacher distribu-
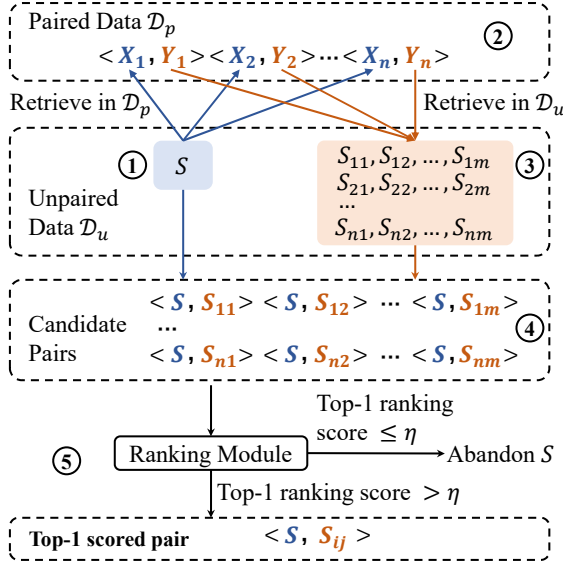
Figure 2: Framework of data-level distillation. (1) The sentence S is randomly selected in the unpaired data $\mathcal{D}_u$. (2) A set of posts $X_1, \ldots, X_n$ that are similar to S are retrieved from the paired data $\mathcal{D}_p$. (3) Each corresponding response Y_i is then used to retrieve m sentences $S_{i1}, \ldots, S_{im}$ that are similar to Y_i from $\mathcal{D}_u$. (4) Then $n \times m$ candidate pairs can be formed by grouping S with each sentence: $\langle S, S_{ij} \rangle$, $(i = 1, \ldots, n, j = 1, \ldots, m)$. (5) A ranking module is used to rank these candidate pairs.

tion and the model distribution. The most related work in this branch compared to ours was done by Kim and Rush (2016). However, their methods do not utilize unpaired data, and the augmented data are decoded from a probability model using beam search. Whereas our method tries to utilize the unpaired data, and the augmented data are generated by aligning human produced sentences.

There are also works that try to utilize retrieved non-conversational texts to improve the diversity of the dialogue model (Wu et al., 2019; Cai et al., 2019; Zhu et al., 2019; Su et al., 2020). However, most of these studies focus on extracting templates from these non-conversational texts rather than generating augmented pairs, and they typically use specifically designed model structures. Nevertheless, the data augmentation method proposed in our study can be used in combination with any dialogue models to improve the performance.

3 Data-level Distillation

The data-level distillation in our method aims at constructing a set of new post-response pairs $\mathcal{D}_a$ by matching non-parallel sentences retrieved from unpaired data $\mathcal{D}_u$. Specifically, $\mathcal{D}_p$ consists of

N post-response pairs: $\mathcal{D}_p = \{\langle X_i, Y_i \rangle\}_{i=1}^N$, in which X_i and Y_i is the post and response, respectively, and $\mathcal{D}_u$ consists of M non-parallel sentences: $\mathcal{D}_u = \{S_i\}_{i=1}^M$. Note that M is usually much larger than N because non-parallel sentences are generally easier to collect.

Further, the output of our data-level distillation process is a set of augmented post-response pairs: $\mathcal{D}_a = \{\langle X_i', Y_i' \rangle\}_{i=1}^K$, in which both the post and response come from the unpaired dataset $\mathcal{D}_u$, i.e., $X_i' \in \mathcal{D}_u$ and $Y_i' \in \mathcal{D}_u$ for $i = 1, \ldots, K$.

The data-level distillation involves two major processes: 1) constructing candidate pairs and 2) filtering low-quality candidates. The whole framework is shown in Figure 2 and detailed below.

3.1 Constructing Candidate Pairs

We first construct candidate dialogue pairs with the help of some post-response pairs $\langle X_i, Y_i \rangle$ selected from $\mathcal{D}_p$. The basic intuition is that sentences that are similar to post X_i can usually be responded with sentences that are similar to the corresponding response Y_i. Candidate dialogue pairs can be then constructed by anchoring sentences in $\mathcal{D}_u$ using $\langle X_i, Y_i \rangle$.

The construction of candidate pairs starts by randomly selecting a sentence S from the unpaired dataset $\mathcal{D}_u$. We then treat S as a candidate post, and it is used to retrieve n posts X_i $(1 \le i \le n)$ that are similar to S from the paired data $\mathcal{D}_p$. In this study, the sentence retrieval process is implemented based on the Okapi BM25 algorithm, which scores the similarity of input sentences using bag-of-words features. Then the corresponding n post-response pairs $\langle X_i, Y_i \rangle$ $(1 \le i \le n)$ are extracted from $\mathcal{D}_p$. For each response Y_i, we further retrieve m sentences S_{ij} $(1 \le j \le m)$ that are similar to Y_i from the unpaired dataset $\mathcal{D}_u$. These sentences S_{ij} can then serve as candidate responses to the original sentence S, and therefore $n \times m$ candidate pairs $\langle S, S_{ij} \rangle$, $(1 \le i \le n, 1 \le j \le m)$ are generated. Moreover, for each candidate pair $\langle S, S_{ij} \rangle$, we name the post-response pair $\langle X_i, Y_i \rangle$ in $\mathcal{D}_p$ that are used to produce $\langle S, S_{ij} \rangle$ as its *"anchor pair"* since it anchors the sentences S and S_{ij} from $\mathcal{D}_u$.

Note that we have explored other variants of the above process, such as treating the initial sentence S as a candidate response rather than a candidate post or utilizing more advanced text retrieval methods to extract similar sentences. However, we notice little difference in neither the quality of the

final augmented pairs nor the performance improvement brought to the dialogue models.

3.2 Filtering Candidate Pairs

In order to enhance the quality of the augmented data, we propose to filter out low-quality pairs using a ranking module, which calculates a score for each candidate pair obtained above. Specifically, high-quality pairs that are fluent and coherent are expected to receive high scores. In this study, we implement the score function as a text matching model, which is built by fine-tuning a pre-trained BERT model on the paired dataset $\mathcal{D}_p$. Negative samples are constructed by replacing the original responses using randomly sampled sentences from $\mathcal{D}_p$. The ranking score for each input pair is calculated as the matching score produced by the matching model.

In this study, we follow a quite rigorous policy to select the final augmented pairs in $\mathcal{D}_a$. For each sample sentence S from $\mathcal{D}_u$, we only extract the top-1 scored pair $\langle S, S_{ij} \rangle$ among all its $n \times m$ candidate pairs, and $\langle S, S_{ij} \rangle$ is added to $\mathcal{D}_a$ only when its matching score exceeds a certain threshold $\eta (0.9 \leq \eta)$. We repeat the above procedures with newly sampled sentences from $\mathcal{D}_u$ until a desired number of augmented pairs in $\mathcal{D}_a$ are obtained. The whole data-level distillation process in our method is summarized in Algorithm 1.

Note that the matching model used in the ranking process can also be directly used to align sentences from the unpaired dataset $\mathcal{D}_u$. Specifically, for a sampled sentence S from $\mathcal{D}_u$, we can treat all other sentences in $\mathcal{D}_u$ as its candidate response and select an augmented pair by ranking all these candidates. Although theoretically possible, this approach is practically infeasible considering the large amount of sentences in $\mathcal{D}_u$ and the tremendous computational load to rank these candidates. Note that previous works on effective ranking (such as Henderson et al. (2017, 2020)) can not be directly adapted to this study because our ranking model does not use dot-product scoring function.

4 Model-level Distillation

A straightforward way to improve a dialogue model with the augmented dialogue data is to directly merge the original paired data $\mathcal{D}_p$ with $\mathcal{D}_a$. However, this naive approach may lead to sub-optimal performance since the augmented pairs in $\mathcal{D}_a$ might not be as high-quality as these human-crafted pairs

---

Algorithm 1 Data-level distillation process

Input: A set of unpaired data $\mathcal{D}_u = \{S_i\}_{i=1}^M$, a set of paired data $\mathcal{D}_p = \{\langle X_i, Y_i \rangle\}_{i=1}^N$, a threshold η.
Output: Augmented dialogue pairs $\mathcal{D}_a = \{\langle X_i', Y_i' \rangle\}_{i=1}^K$.
1: $\mathcal{D}_a \leftarrow$ Empty set
2: **while** $|\mathcal{D}_a| < K$ **do**
3:     $\widetilde{\mathcal{D}}_a \leftarrow$ Empty set
4:     Sample a sentence $S \sim \mathcal{D}_u$.
5:     Retrieve n posts $\{X_i\}_{i=1}^n$ that are similar to S in $\mathcal{D}_p$.
6:     Get the responses $\{Y_i\}_{i=1}^n$ for $\{X_i\}_{i=1}^n$ from $\mathcal{D}_p$.
7:     **for** each Y_i **do**
8:         Retrieve m sentences $\{S_{ij}\}_{j=1}^m$ that are similar to Y_j in $\mathcal{D}_u$.
9:         $\widetilde{\mathcal{D}}_a \leftarrow \widetilde{\mathcal{D}}_a \bigcup \{\langle S, S_{ij} \rangle\}_{j=1}^m$
10:     **end for**
11:     Calculate the ranking score for each pair in $\widetilde{\mathcal{D}}_a$.
12:     Extract the top-1 scored pair $\langle S, S_{ij} \rangle$ from $\widetilde{\mathcal{D}}_a$.
13:     **if** The ranking score of $\langle S, S_{ij} \rangle$ exceeds η **then**
14:         $\mathcal{D}_a \leftarrow \mathcal{D}_a \bigcup \{\langle S, S_{ij} \rangle\}$
15:     **end if**
16: **end while**

---

in $\mathcal{D}_p$. In this study, we apply the model-level distillation in the training process to prevent the dialogue models from being affected by the noise in $\mathcal{D}_a$. This approach can be used in both retrieval-based and generation-based dialogue models.

4.1 Retrieval-based Dialogue Model

A retrieval-based dialogue model produces responses by retrieving a best matching sentence from the pre-collected dialogue dataset. Its key component is a matching function $\mathcal{P}_\theta(l|X, Y)$ that predicts whether a response Y matches a given post X. Specifically, $l \in \{0, 1\}$ is a matching label, where $l = 1$ means Y is a proper response for X and $l = 0$ otherwise. The model parameters θ can be learned by optimizing a negative log likelihood (NLL) loss defined as

$$\mathcal{L}_{m-nll}(\theta) = - (1 - l)\log\mathcal{P}_\theta(0|X, Y) \\ - l\log\mathcal{P}_\theta(1|X, Y) \tag{1}$$

In this study, we formalize the matching function using the BERT model (Devlin et al., 2019; Whang et al., 2020). A teacher model $\mathcal{P}_{\theta_t}(l|X, Y)$ is first obtained by optimizing the NLL loss $\mathcal{L}_{m-nll}(\theta_t)$ on the paired dataset $\mathcal{D}_p$. After the training is completed, the teacher model is fixed and used to compute a knowledge distillation (KD) loss (Kim and Rush, 2016) as

$$\mathcal{L}_{m-kd}(\theta) = - \sum_{i=0}^1 \mathcal{P}_{\theta_t}(i|X, Y) \cdot \log\mathcal{P}_\theta(i|X, Y). \tag{2}$$

The final matching model is trained on the following loss:

$$\mathcal{L}_M(\theta) = \mathcal{L}_{m-nll}(\theta) + \alpha_m \mathcal{L}_{m-kd}(\theta), \qquad (3)$$

where the loss $\mathcal{L}_M(\theta)$ is evaluated using $\mathcal{D}_p \bigcup \mathcal{D}_a$ and α_m is used to balance these two losses.

4.2 Generation-based Dialogue Model

A generation-based dialogue model tries to capture the distribution of the response sentences Y given the post sentence X, i.e., $\mathcal{P}_\phi(Y|X)$, which can be formalized as

$$\mathcal{P}_\phi(Y|X) = \prod_{i=1}^{|Y|} P_\phi(y_i|y_{<i}, X), \qquad (4)$$

where $|Y|$ is the length of Y, $y_{<i} = y_1 \cdots y_{i-1}$ is the token sequence before y_i. The model parameters ϕ can be learned by optimizing the NLL loss:

$$\mathcal{L}_{g-nll}(\phi) = -\sum_{i=1}^{|Y|} \log P_\phi(y_i|y_{<i}, X). \qquad (5)$$

In this study, we parameterize the dialogue generation model using the Transformer-based encoder-decoder framework (Vaswani et al., 2017; Golovanov et al., 2019; Zheng et al., 2020b). Similar to the retrieval-based approach, a teacher model is first obtained by optimizing the NLL loss $\mathcal{L}_{g-nll}$ on the paired dataset $\mathcal{D}_p$ and the trained teacher model is used to compute a KD loss as

$$\begin{aligned} \mathcal{L}_{g-kd}(\phi) = -\sum_{i=1}^{|Y|}\sum_{j=1}^{|\mathcal{V}|} & P_{\phi_t}(y_i = j|y_{<i}, X) \\ & \times \log P_\phi(y_i = j|y_{<i}, X), \end{aligned} \qquad (6)$$

where $|\mathcal{V}|$ denotes the size of the vocabulary and ϕ_t is the parameter of the teacher model, which is fixed.

The final loss for the generation model is:

$$\mathcal{L}_G(\phi) = \mathcal{L}_{g-nll}(\phi) + \alpha_g \mathcal{L}_{g-kd}(\phi), \qquad (7)$$

where the loss $\mathcal{L}_G(\theta)$ is evaluated using $\mathcal{D}_p \bigcup \mathcal{D}_a$ and α_g is used to balance these two losses.

5 Experiment

5.1 Dataset

The evaluation of our method is performed on a corpus collected from Weibo[1]. Specifically, the paired data $\mathcal{D}_p$ contains 300K post-response pairs, which are made up of Weibo posts and their following replies. All these pairs are manually filtered with annotators by removing ungrammatical sentences and incoherent dialogues. The unpaired data $\mathcal{D}_u$ contains about 2 million posts on Weibo that do not have replies. Non-fluent sentences in $\mathcal{D}_u$ are filtered out using a set of heuristic rules. Further, two additional sets of paired data are also prepared to validate and test the dialogue models, with 10K and 5K pairs respectively. These dialogue pairs are collected and manually filtered using the same criterion as $\mathcal{D}_p$.

5.2 Implementation Details

Data-level Distillation: We implement the retrieval module in Section 3.1 using the Lucene library [2], and set the value of both n and m to 5. The matching model used in Section 3.2 is fine-tuned with $\mathcal{D}_p$ for three epochs based on the pretrained BERT-base model (Devlin et al., 2019). The hyperparameter setting of the matching model follows the work of Devlin et al. (2019).

Model-level Distillation: For the *retrieval-based dialogue model*, the matching model used in Section 3.2 is directly used as the teacher model to calculate the KD loss (Eq. 2). The final retrieval-based dialogue model is initialized with the pretrained BERT-base weights and fine-tuned using the loss in Eq. 3 for 2 epochs on $\mathcal{D}_p \bigcup \mathcal{D}_a$. The value of α_m in Eq. 3 is set to 1.

For the *generation-based dialogue model*, the encoder and decoder share the same set of parameters, which is initialized using a pretrained GPT model (Wang et al., 2020). The teacher model uses the same architecture and it is fine-tuned using the paired dataset $\mathcal{D}_p$ for 15 epochs on the NLL loss (Eq. 5). The final generative dialogue model is first initialized using the pre-trained GPT weights and fine-tuned using the loss in Eq. 7 for 50 epochs on $\mathcal{D}_p$ and $\mathcal{D}_a$. The value of α_g in Eq. 7 is set to 1. Moreover, the GPT model used in the initialization phase is trained on a corpus collected from various Chinese novels. This corpus contains about 0.5 billion tokens and a character-level vocabulary of size 13,084.

See Appendix A for more details of the model setting and reproduction guidance. The data and code for all experiments can be downloaded from

---

[1] https://www.weibo.com

[2] https://lucene.apache.org/core/

the link[3].

5.3 Evaluating Augmented Dialogue Pairs

5.3.1 Baselines

We first evaluate the quality of the augmented pairs generated by our Data-Level (**DL**) distillation process. Three different matching thresholds η in Algorithm 1 are tested, i.e., $\eta = 0.90, 0.95, 0.99$. Several strong baselines are also compared:

CVAE: A CVAE-based model as proposed by Li et al. (2019) is trained on the paired data $\mathcal{D}_p$. Augmented pairs are generated by sampling different latent codes.

BT: Augmented pairs are generated by Back Translating (i.e., translate Chinese to English and then translate back to Chinese) the post sentences of the dialogue pairs in $\mathcal{D}_p$. The translation is done via the Google Translate API.

SP: A variant of our method is implemented by first Sampling a post-response Pair $\langle X, Y \rangle$ from $\mathcal{D}_p$, and then retrieving a best-matching post and response from the unpaired data $\mathcal{D}_u$ using X and Y as the query, respectively. An augmented pair is constructed by pairing the retrieved post and response sentence without the ranking process.

Note that there are two major differences between the baseline SP and our data-level distillation process: 1) the baseline SP starts with a dialogue pair $\langle X, Y \rangle$ sampled from $\mathcal{D}_p$ rather than a candidate post sampled from $\mathcal{D}_u$; 2) The ranking process is not used in the baseline SP to further filter the candidate pairs.

5.3.2 Metrics

The automatic evaluation of augmented dialogue pairs uses the following metrics: 1) ***Distinct*** (Li et al., 2016) is used to measure the proportion of unique n-grams in the augmented dialogue pairs (n=1,2,3,4); 2) ***Novelty*** (Wang and Wan, 2018) is used to measure the proportion of new n-grams in the augmented dialogue pairs (n=1,2,3,4), i.e., n-grams that are covered by the augmented dialogue pairs but are not shown in the paired dataset $\mathcal{D}_p$. A higher novelty score means the augmented dialogue pairs contain more "novel" contents.

Manual evaluation is also used to evaluate the quality of augmented dialogue pairs. Three annotators are employed to rate these pairs from two aspects: 1) *Fluency* (**Flu.**): whether the augmented pairs are fluent; 2) *Coherency* (**Coh.**): whether the

response is coherent with the post so that they make a plausible dialogue pair. The rating scale for each measure is of (0, 1, 2), in which 0 means worst and 2 best.

5.3.3 Results

Each data augmentation method introduced above are used to generate 300K augmented dialogue pairs, and on which automatic evaluation is performed. Further, manual evaluation is carried out on 200 dialogue pairs that are randomly sampled from these augmented data, and the inter-rater agreement between annotators is measured using the Fleiss's kappa κ (Randolph, 2005). The κ value for *Fluency* and *Coherency* is 0.69 (substantial agreement), and 0.42 (moderate agreement), respectively. Note that this evaluation is purely regarding the augmented dialogue data, without considering any dialogue model training.

The evaluation results in Table 1 demonstrate that the augmented dialogue data produced by our method outperform all the baselines in almost all the metrics. We can further observe that: **1**) Our method obtains similar scores on all the metrics compared to these human-produced and filtered dialogue pairs in $\mathcal{D}_p$. This indicates that the augmented dialogue pairs generated by our method are of high quality. We present some examples of the augmented pairs together with their associated anchor pairs in Table 2. **2**) The matching threshold η can be used to trade off between the coherency and diversity of the augmented dialogue pairs. Specifically, a higher η value improves Fluency and Coherency scores but hurts Distinct and Novelty scores of the augmented pairs.

5.4 Evaluating Dialogue Models

5.4.1 Baselines

We evaluate the benefit of the augmented dialogue data in both retrieval-based and generation-based dialogue models. Specifically, 300K augmented dialogue pairs are generated using these three baselines introduced in Section 5.3.1, and the model-level distillation process as introduced in Section 4 is used to train the dialogue models. We denote these three dialogue model baselines as **CVAE+ML**, **BT+ML**, and **SP+ML**, respectively. Note that the notation "ML" means that the Model-Level distillation is used. Moreover, besides comparing to different data augmented methods as introduced in Section 5.3.1, several other competitive dialogue model baselines are also tested:

---

[3]https://github.com/njuzrs/dialogue_distillation

Model	Distinct-1,2,3,4				Novelty-1,2,3,4				Flu.	Coh.
CVAE	0.178‡	09.40‡	34.54‡	60.73‡	00.25‡	08.47‡	25.45‡	40.62‡	1.529‡	0.862‡
BT	0.193‡	12.42‡	43.43‡	70.38‡	03.07‡	21.66‡	35.28‡	45.18‡	1.771‡	<u>1.408</u>†
SP	**0.228**	11.56‡	37.76‡	57.73‡	18.48‡	46.65‡	73.56‡	87.79‡	1.839‡	0.777‡
DL η=0.90	<u>0.226</u>‡	**13.72**	**48.24**	**76.21**	**23.76**	**55.95**	**80.64**	**92.10**	1.835‡	1.183‡
DL η=0.95	0.224‡	<u>13.44</u>‡	<u>47.51</u>‡	<u>75.55</u>‡	<u>22.81</u>‡	<u>55.51</u>‡	<u>80.37</u>‡	<u>91.97</u>‡	<u>1.856</u>†	1.358‡
DL η=0.99	0.213‡	12.61‡	45.06‡	72.87‡	21.59‡	54.40‡	79.69‡	91.62‡	**1.877**	**1.428**
$\mathcal{D}_p$(human)	0.199	13.51	47.70	75.52		N/A			1.868	1.617

Table 1: Automatic and manual evaluation on the quality of augmented pairs produced by different methods. The bottom row corresponds to the human filtered dialogue pairs in $\mathcal{D}_p$. The best results are in **bold**, and the second best results are <u>underlined</u> (except "human"). Significant tests between the best model and others were performed using t-test. † and ‡ indicates p-value < 0.01 and 0.001, respectively.

	Augmented pairs	Associated anchor pairs from $\mathcal{D}_p$
Post	I'm almost moved to cry (我已经快感动地哭了)	I am so moved today! (今天感动得快哭了！)
Resp	What happened there? (发生什么事情呢？)	What happen (发生什么事)
Post	I like it, men should be like this (这话我喜欢。男人就该这样)	I like this types of man (喜欢这样的男人)
Resp	I like it too, just as you do (我也喜欢。跟你一样)	Your taste is just like mine (怎么跟我喜欢的一样)
Post	I liked to play it when I was young (小时候很喜欢玩)	My favorite toy in kindergarten (幼儿园最喜欢玩的)
Resp	I have also played, it's so cute (表示有幸玩过，很萌哒)	I have also played, lol (我也玩过哒)

Table 2: Example pairs produced by the proposed data augmentation method. The associated anchor pairs are also shown. More examples are shown in Appendix B.

Teacher: Training the dialogue models on the paired data $\mathcal{D}_p$ with the NLL loss. Note that this setting produces the teacher models used in Section 4.

AP: Training dialogue models only on the <u>A</u>ugmented <u>P</u>airs $\mathcal{D}_a$ with the NLL loss.

UP+PreT: First fine-tuning the pre-trained GPT (with the NLL loss in Eq. 5) or BERT-base model (with the MLM loss (Devlin et al., 2019)) on the <u>UnP</u>aired Data $\mathcal{D}_u$, and then using these fine-tuned weights to initialize the dialogue models, which are further fine-tuned on $\mathcal{D}_p$ with the NLL loss.

NP+ML: Sampling 300K pairs from a set of Weibo dialogues that are not manually filtered and use these "<u>N</u>oisy <u>P</u>airs" as the augmented pairs. The model-level distillation process introduced in Section 4 is used to train this baseline.

We denote our method as **DL+ML** since it trains the dialogue model using both the data-level and model-level distillation. The threshold η in Algorithm 1 is set to 0.95 for a better trade-off between the coherency and diversity of the augmented data. Further, we also test another method to work with data-level distillation (i.e., utilizing $\mathcal{D}_a \bigcup \mathcal{D}_p$): **DL+PreT**, i.e., first pre-train the dialogue model on $\mathcal{D}_a$ and then fine-tune on $\mathcal{D}_p$ with the NLL loss.

Further, we also performed several ablation tests on our method to validate the effect of each component: 1) training dialogue models on $\mathcal{D}_p \bigcup \mathcal{D}_a$ using only the NLL loss, i.e., without the model-level distillation (**w/o ML**); 2) training dialogue models only on the paired data $\mathcal{D}_p$ using $\mathcal{L}_M(\theta)$ or $\mathcal{L}_G(\phi)$, i.e., the data-level distillation are not used (**w/o DL**); 3) training dialogue models on the augmented data $\mathcal{D}_a$ using $\mathcal{L}_M(\theta)$ or $\mathcal{L}_G(\phi)$, i.e., the paired data $\mathcal{D}_p$ are not used (**w/o PD**); 4) generating $\mathcal{D}_a$ without the ranking module (**w/o Ranking**), i.e., the candidate pairs are used as the augmented data without filtering.

Note that all the baselines and ablation models are initialized with pre-trained GPT or BERT-base weights.

5.4.2 Metrics

The retrieval-based dialogue models are evaluated using the following metrics: 1) *Mean Average Precision* (**MAP**): the average rank of the reference responses; 2) $R_{10}@k$: the recall of the reference response being in the top-k ranked candidates (k=1,2,5) when given 10 candidates in total.

The generation-based dialogue models are evaluated both automatically and manually. Specifi-

cally, the following automatic metrics are used: 1) *Perplexity* (**PPL**) which measures how the model fits the test data; 2) ***BLEU*** which evaluates the overlap of n-grams (n=1,2) between the generated and reference responses; 3) *Distinct* (**Dist.**) measures the proportion of unique n-grams in the generated responses (n=1,2). Manual evaluation is also performed for the generated dialogue responses following the same protocol as introduced in Section 5.3.2.

Model	MAP	$R_{10}@1$	$R_{10}@2$	$R_{10}@5$
Teacher	80.2	69.7	82.1	95.1
AP	76.5	65.1	78.0	92.1
UP+PreT	80.6	70.3	82.6	**95.3**
NP+ML	80.8	70.5	82.9	95.2
CVAE+ML	80.3	69.8	82.5	94.9
BT+ML	80.3	69.8	82.0	95.2
SP+ML	80.4	70.0	82.0	95.2
DL+PreT	80.7	70.2	82.7	**95.3**
DL+ML	**81.0**	**70.8**	**83.1**	**95.3**
w/o ML	80.4	69.9	82.5	95.0
w/o DL	80.5	70.1	82.3	95.1
w/o PD	79.5	68.9	81.3	94.1
w/o Ranking	80.5	70.1	82.5	95.2

Table 3: Automatic evaluation for retrieval-based dialogue models with different training and data augmentation methods.

5.4.3 Results

Automatic evaluation for each dialogue model is performed on 5K test data (see Table 3 and Table 4 for the results), and manual evaluation is performed using 200 pairs that are randomly sampled from these test data (see Table 5 for the results). The κ value for the *Fluency* and *Coherency* annotation is 0.9 (substantial agreement) and 0.56 (moderate agreement), respectively.

Our method outperforms all the baselines in almost all the metrics for both retrieval-based and generation-based dialogue models. We can further observe that: **1**) The dialogue models that utilize unpaired data $\mathcal{D}_u$ (e.g. DL+ML, DL+PreT, UP+PreT) generally outperform the models that are only trained on $\mathcal{D}_p$ (e.g., Teacher, CVAE+ML). This demonstrates that utilizing unpaired data is more effective at improving the performance of dialogue models; **2**) Training the dialogue models on the merged data $\mathcal{D}_p \bigcup \mathcal{D}_a$ without utilizing the

Model	PPL	BLEU-1,2		Dist.-1,2	
Teacher	23.9‡	12.25‡	6.61‡	3.83‡	29.69‡
AP	50.0‡	10.86‡	5.52‡	3.29‡	23.37‡
UP+PreT	24.0‡	12.60	6.81†	3.99‡	30.50‡
NP+ML	23.1‡	11.63‡	6.25‡	3.99‡	28.47‡
CVAE+ML	23.9‡	12.27‡	6.59‡	3.73‡	26.75‡
BT+ML	23.8‡	11.93‡	6.48‡	3.84‡	27.38‡
SP+ML	23.6‡	12.47‡	6.74‡	4.04	30.66‡
DL+PreT	23.7‡	**12.66**	6.92	3.95‡	30.30‡
DL+ML	**22.6**	12.42‡	**6.93**	**4.13**	**31.39**
w/o ML	23.3‡	12.30‡	6.65‡	4.06	30.89‡
w/o DL	23.5‡	12.54†	6.88	3.96‡	29.79‡
w/o PD	26.7‡	11.08‡	5.86‡	3.48‡	26.84‡
w/o Ranking	22.8‡	12.54‡	6.78‡	3.90‡	28.93‡

Table 4: Automatic evaluation results for generation-based dialogue models with different training and data augmentation methods. Significance tests between the best model and others were performed using t-test with booststrap resampling (Koehn, 2004). † and ‡ indicates p-value < 0.005 and 0.001, respectively.

model-level distillation (i.e., w/o ML) brings little or no performance improvements compared to directly training on $\mathcal{D}_p$ (i.e., Teacher). This verifies the effectiveness of the model-level distillation process proposed in our method; **3**) When the model-level distillation is employed, the augmented data produced by our data-level distillation process (i.e., DL+ML) can better improve the performance of dialogue models compared to the augmented data produced by other data augmentation methods (e.g. CVAE+ML, NP+ML, SP+ML, BT+ML). This verifies the effectiveness of the data-level distillation process proposed in our study.

Model	Flu.	Coh.
Teacher	1.968‡	1.432‡
AP	1.985	1.417‡
UP+PreT	1.957‡	1.500
NP+ML	1.967‡	1.473†
CVAE+ML	1.977†	1.475†
BT+ML	1.957‡	1.503
SP+ML	1.973†	1.453‡
DL+PreT	1.975	1.492†
DL+ML	**1.993**	**1.518**

Table 5: Manual evaluation for generation-based dialogue models. Significant tests between the best model and others were performed using t-test. † and ‡ indicate p-value < 0.05 and 0.01, respectively.

6 Conclusion

This paper presents a novel dialogue distillation method that consists of two processes, i.e., 1) a data augmentation process to construct new post-response pairs from unpaired data and 2) a model distillation process that distills a teacher model trained on the original data to the augmented data. Automatic and manual evaluation shows that our method can produce high-quality post-response pairs that are both coherent and content-rich, which can be further used to improve the performance of competitive baselines. Our method may inspire other research in low-resource NLP tasks.

7 Acknowledgments

References

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. 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 1219–1228.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. 2019. Autoaugment: Learning augmentation policies from data. In *Proceedings of CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Wenchao Du and Alan Black. 2018. Data augmentation for neural online chats response selection. In *Proceedings of EMNLP Workshop SCAI*.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of AAAI*.

Sergey Golovanov, Rauf Kurbanov, Sergey Nikolenko, Kyryl Truskovskyi, Alexander Tselousov, and Thomas Wolf. 2019. Large-scale transfer learning for natural language generation. In *Proceedings of ACL*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. Convert: Efficient and accurate conversational representations from transformers. *arXiv preprint arXiv:1911.03688*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *Proceedings of NIPS Deep Learning Workshop*.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of EMNLP*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. 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 NAACL*.

Juntao Li, Lisong Qiu, Bo Tang, Dongmin Chen, Dongyan Zhao, and Rui Yan. 2019. Insufficient data can also rock! learning to converse using smaller data with augmentation. In *Proceedings of AAAI*.

Zhengdong Lu and Hang Li. 2013. A deep architecture for matching short texts. In *Proceedings of NIPS*.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *IJCAI*, pages 4279–4285.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of ACL*.

Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. 2020. Diversifying dialogue generation with non-conversational text. *arXiv preprint arXiv:2005.04346*.

I Sutskever, O Vinyals, and QV Le. 2014. Sequence to sequence learning with neural networks. *Advances in NIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*.

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

Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of IJCAI*.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *NLPCC*.

Jason W Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of EMNLP*.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and HeuiSeok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. In *Interspeech*.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7281–7288.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of ICLR*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, pages 2204–2213.

Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskenazi. 2017. Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability. In *Proceedings of SIGdial*, pages 27–36.

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020a. Out-of-domain detection for natural language understanding in dialog systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1198–1209.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Yinhe Zheng, Rongsheng Zhang, Xiaoxi Mao, and Minlie Huang. 2020b. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of AAAI*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018a. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of AAAI*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018b. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of IJCAI*, pages 4623–4629.

Hao Zhou, Chujie Zheng, Kaili Huang, Minlie Huang, and Xiaoyan Zhu. 2020. KdConv: A chinese multi-domain dialogue dataset towards multi-turn knowledge-driven conversation. In *Proceedings of the 58th Conference of the Association for Computational Linguistics*.

Qingfu Zhu, Lei Cui, Weinan Zhang, Furu Wei, and Ting Liu. 2019. Retrieval-enhanced adversarial training for neural response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3763–3773.