Heterogeneous-Branch Collaborative Learning for Dialogue Generation

Yiwei Li, Shaoxiong Feng, Bin Sun, Kan Li*

School of Computer Science, Beijing Institute of Technology
{liyiwei,shaoxiongfeng,binsun,likan}@bit.edu.cn

Abstract

With the development of deep learning, advanced dialogue generation methods usually require a greater amount of computational resources. One promising approach to obtaining a high-performance and lightweight model is knowledge distillation, which relies heavily on the pre-trained powerful teacher. Collaborative learning, also known as online knowledge distillation, is an effective way to conduct one-stage group distillation in the absence of a well-trained large teacher model. However, previous work has a severe branch homogeneity problem due to the same training objective and the independent identical training sets. To alleviate this problem, we consider the dialogue attributes in the training of network branches. Each branch learns the attribute-related features based on the selected subset. Furthermore, we propose a dual group-based knowledge distillation method, consisting of positive distillation and negative distillation, to further diversify the features of different branches in a steadily and interpretable way. The proposed approach significantly improves branch heterogeneity and outperforms state-of-the-art collaborative learning methods on two widely used open-domain dialogue datasets.

Introduction

Open-domain Neural dialogue generation (Sordoni et al. 2015; Vinyals and Le 2015; Shang, Lu, and Li 2015), aiming to generate diverse and coherent responses, has gained increasing attention and achieved impressive performance. It is important to recognize, however, that these considerable improvements typically come at the expense of over-parameterized networks, inhibiting their development on real-world resource-limited scenarios such as mobile chatbot applications. Knowledge distillation is an appropriate knowledge-transfer methodology to resolve this issue, which uses predicted distributions (Hinton, Vinyals, and Dean 2015), hidden states (Sun et al. 2019), or attention matrices (Jiao et al. 2020), etc. of a teacher model as targets to induce the student to imitate. A conventional distillation process involves two stages that begin with a cumbersome pre-trained teacher model and then distill the knowledge to the compact student model. Unfortunately, training such a complex teacher model is time-consuming and a high-capacity model may not always be available.

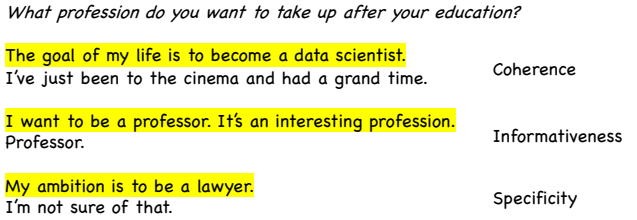

Figure 1: Dialogue examples with two levels related to three dialogue attributes. The response quality can be assessed by multiple perspectives.

With a view to overcoming traditional limitations, online knowledge distillation (Anil et al. 2018; Lan, Zhu, and Gong 2018), also called collaborative learning (Zhang et al. 2018c; Song and Chai 2018), is currently receiving considerable attention. Instead of pre-training a high-capacity teacher, collaborative learning conducts a single-stage group-based knowledge distillation that transfers the knowledge between less-parameterized student branches simultaneously. Aside from accelerating model learning efficiency over conventional KD, another major advantage of collaborative learning is the ability to find a more robust local minimum when compared to a single model learning method. It is important to note, however, that the homogeneity problem among branches along with training will lead to early saturation, thereby affecting the effectiveness of group distillation. To alleviate this problem, Chen et al. (2020) impeded homogenization by equipping a diversity holding mechanism; Wu and Gong (2021) randomly enhanced the input to guarantee the discrepancy between branches; Feng et al. (2021a) proposed random routing to improve the diversity of features.

Even though the aforementioned approaches have demonstrated their superiority, one major drawback remains to limit further branch heterogenization: previous work only focused on the classification task, resulting in the same training objective (i.e. classification accuracy) for all branches of the framework with independent identical distribution (i.i.d.) training data. It will make different branches tend to converge to similar feature representations (Li et al. 2016c; Lan,

*Corresponding author.

Zhu, and Gong 2018; Chen et al. 2020). Consequently, there is barely any intuitive approach to allow branches to develop in various training directions, which is a significant obstacle to fostering diversity among them.

As opposed to classification task, goals of dialogue generation model focus more on dialogue attributes than the accuracy with the references. As demonstrated in See et al. (2019), inadequate modeling of conversational aspects such as coherence and specificity results in inferior model performance and low response quality. Taking Figure 1 as an example, response quality can be affected by multiple perspectives. As a result, directly applying the collaborative learning to the dialogue generation task will lead to sub-optimal performance. A variety of dialogue attributes can provide a natural insight into improving branch heterogeneity. With this in mind, it is possible to develop more effective and interpretable techniques to further enhance branch diversity.

In this work, we propose a heterogeneous attribute-aware collaborative learning paradigm for response generation, comprising two types of branches: auxiliary and master. To achieve the goals of the dialogue system in a fine-grained way, we train each auxiliary branch on the corresponding aspect-specific sub-set to capture features along with some dialogue attributes. Each sub-set is collected according to the corresponding scoring method. Unlike auxiliary branches, the master branch is trained with the entire dataset to learn features roughly but comprehensively. In previous work, each branch learns from all the other branches, which is prone to homogenize different branches as the training continues. To further improve the diversity of auxiliary branches and integrate multi-view knowledge steadily, we propose the dual group-based knowledge distillation, consisting of positive distillation (Hinton, Vinyals, and Dean 2015) and negative distillation (Li et al. 2022a).

Specifically, positive distillation is conducted from all the auxiliary branches to the master branch, transferring the attribute-specific knowledge effectively, whilst negative distillation is performed within the attribute-related branches to enforce them to learn different dialogue properties. Furthermore, negative distillation is implemented on hierarchical feature representation to use multi-level negative knowledge. However, due to some common features shared by different auxiliary branches, blindly maximizing the distance of hidden states among auxiliary branches will harm the model performance and the training stability. To this end, we design a novel distillation approach called orthogonal negative distillation. It only strengthens the features of each auxiliary branch orthogonal to the other branches, avoiding disturbing the learning of common knowledge.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to propose the collaborative dialogue learning approach that transfers attribute-aware knowledge in a one-stage manner.
- Dual group-based knowledge distillation is proposed for better guiding auxiliary branches to learn attribute-specific knowledge. Orthogonal negative distillation can incentivize branches to capture biased features while avoiding harming the common knowledge.
- Extensive Evaluations on two widely used open-domain dialogue datasets demonstrate that the proposed approach significantly improves the branch heterogeneity and outperforms the state-of-the-art collaborative learning methods.

Method

Approach Overview

Taking $C = \{c_1, c_2, ..., c_{T_c}\}$ as context, the objective of dialogue generation task is to generate the response $R = \{r_1, r_2, ..., r_{T_r}\}$, where T_c and T_r represent the length of context and response, respectively. Instead of training a complicated and huge model, we build a collaborative dialogue learning framework to obtain a less-parameterized but effective model for inference. The overview of the proposed framework is illustrated in Figure 2. In consideration of diverse dialogue attributes, we split the training dataset to several sub-sets according to scoring methods measuring the sample quality from multiple perspectives. Each attribute-related sub-set guides one branch to learn the corresponding specific knowledge. After that, we propose dual knowledge distillation in which positive distilling occurs between the master branch and all of the auxiliary branches, while negative distillation occurs within the attribute-related branches to encourage them to learn different dialogue properties. The orthogonal negative distillation is designed to identify biased features without interfering with knowledge.

Dialogue Attribute Learning

The generative dialogue model aims to learn a conditional probability distribution $p_\theta(R|C)$. The maximum likelihood estimation (MLE) is usually used to guide the model to generate the target responses:

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

where r_i is the ground-truth tokens. Therefore, the performance of the dialogue model largely depends on the distribution characteristics of the training set. Recently, a line of work introduces a data manipulation strategy, to boost the model performance with the corresponding dialogue attributes. They first measure the quality of samples in terms of a certain dialogue attribute by a scoring method, and then discard the low-score samples. The selection data can induce the model to learn attribute-related features more effectively for the generation of high-quality responses. Specifically, the raw training samples $\mathcal{D}$ are reorganized into multiple view-specific training sub-sets $(\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_M)$ based on the scores of $\mathcal{S}_m$ and a certain selection proportion. Note that each sample can be assigned to multiple sub-sets as it may obtain high scores from more than one scoring method. Then, each branch m is trained with corresponding sub-set $\mathcal{D}_m$ with Equation 1. Three dialogue attributes are considered in this paper and the following is the details of their corresponding scoring methods:

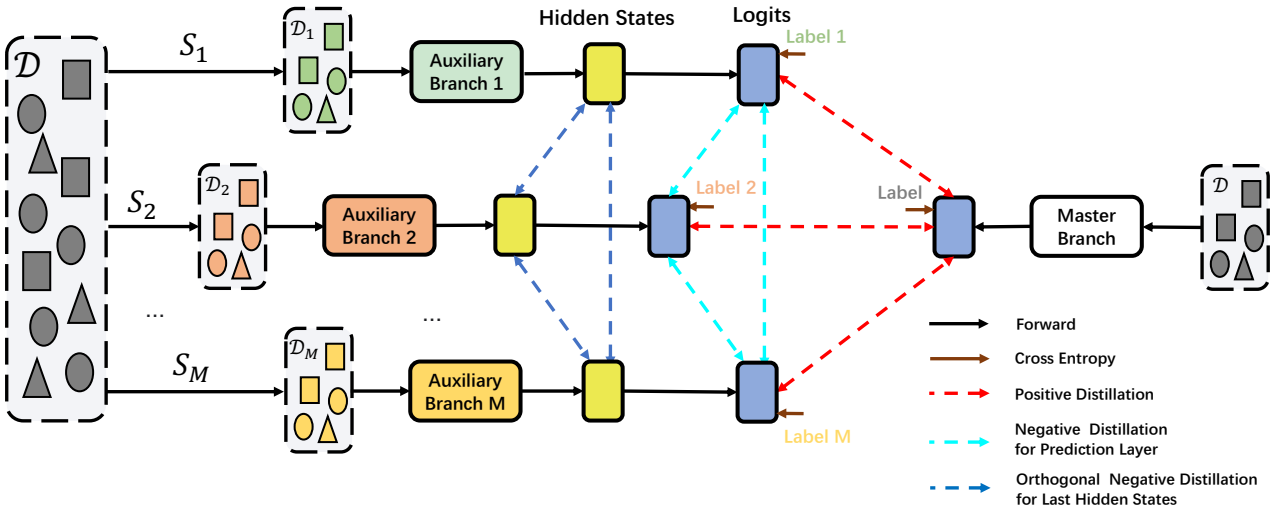

Figure 2: An overview of the proposed heterogeneous attribute-aware collaborative dialogue learning.

Coherence reflects how well a dialogue response semantically relates to its context. A joint score (Akama et al. 2020):

$$S_{C+R}(c, r) = \alpha S_C + \beta S_R \tag{2}$$

that contains two parts: connectivity S_C and content relatedness S_R. The S_C is evaluated by the co-occurrence of keyphrases ($p \in q$, $h \in r$):

$$S_C = \sum_{(p,h)} \frac{\max(nPMI(p, h), 0) \cdot |p| \cdot |h|}{|c| \cdot |r|}, \tag{3}$$

where $|\cdot|$ means the number of words and the $nPMI$ represents the normalized pointwise mutual information (Bouma 2009). In addition, S_R is evaluated by the cosine of the context and its response:

$$S_R = \max(cos(c_{emb}, r_{emb}), 0) \tag{4}$$

The c_{emb} and the r_{emb} are vector representations of the context and response.

Informativeness reflects how much the information related to the query is contained in the generated response, which is evaluated by Entropy_Src (Csaky, Purgai, and Recski 2019): This score is the entropy of a response utterance:

$$H_{src}(r|D) = -\sum_{(c_i,r)\in D} p(c_i|r) \log p(c_i|r), \tag{5}$$

where r represents the response, D represents the dialogue dataset, and c_i means a context of r in D. By using this scoring method, the dialogue pair with many-to-one problem will be filtered, thereby alleviating the phenomenon of general response.

Specificity (See et al. 2019) reflects how much the generated response is good at word usage:

$$Spe(t) = \frac{idf(t) - min\_idf}{max\_idf - min\_idf}, \tag{6}$$

where t is a token of the response, and $idf(t) = \log(\frac{R}{R_t})$. R is the number of responses in the dataset, and R_t is the number of those responses that contain t. Using this scoring method, specific tokens can be identified in the response.

Dual Knowledge Distillation

The distillation objective is employed to alter the representation of two models, denoted as $f^A(x)$ and $f^B(x)$:

$$\mathcal{L}_{\text{KD}} = \sum_{x\in\mathcal{D}} L\left(f^A(x), f^B(x)\right), \tag{7}$$

where $L(\cdot)$ provides a measurement function for calculating distances between representations in multi-levels.

A conventional collaborative learning process only distills positive knowledge, where $L(\cdot)$ is aiming to minimize the distance between branches. However, when it comes to attribute-related branches in dialogue learning, there are different directions in which they tend to converge. It is not appropriate to directly apply positive knowledge distillation to the collaborative dialogue learning framework. In this paper, we propose dual knowledge distillation consisting both positive and negative distillation (where $L(\cdot)$ seeks to maximize the distance between auxiliary branches), as a means of transferring attribute-specific knowledge in a reasonable manner.

Positive Distillation In order to transfer the attribute-aware knowledge to master branch, positive distillation (PD) is performed on the prediction layer:

$$\mathcal{L}_{PD}(\boldsymbol{A}, \boldsymbol{B}) = -\sum_{i=1}^{T_r}\sum_{k=1}^{|\mathcal{V}|} p_{\boldsymbol{A}}\left(r_i = k \mid r_{<i}, C\right) \cdot \log p_{\boldsymbol{B}}\left(r_i = k \mid r_{<i}, C\right), \tag{8}$$

where $\boldsymbol{A}$, $\boldsymbol{B}$ refers to two branches and $p_{\boldsymbol{A}}$, $p_{\boldsymbol{B}}$ are calculated by:

$$p_i = \frac{\exp\left(z_i/T\right)}{\sum_j \exp\left(z_j/T\right)}, \tag{9}$$

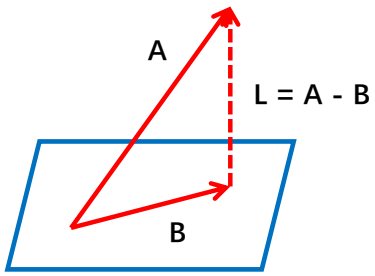

Figure 3: Orthogonal Projection for Hidden States.

where the probability distribution over words is softened with a temperature coefficient T. Positive Distillation is carried out in a bidirectional manner between the master branch and the auxiliary branches. On the one hand, the attribute-specific knowledge can be absorbed by the master branch. On the other hand, the consolidated knowledge from the master branch needs to be transferred to the auxiliary branches in order to facilitate the generation of higher quality responses from them.

Negative Distillation for Prediction Layer For the purpose of encouraging auxiliary branches to better obtain its own specific knowledge, we use the soft unlikelihood loss from Li et al. (2022a) to achieve the negative distillation (ND) within them for the prediction layer first:

$$\mathcal{L}_{ND_{pred}}(\boldsymbol{A}, \boldsymbol{B}) = -\sum_{i=1}^{T_r}\sum_{k=1}^{|\mathcal{V}|} p_{\boldsymbol{B}}\left(r_i = k \mid r_{<i}, C\right) \cdot \log\left(1 - p_{\boldsymbol{A}}\left(r_i = k \mid r_{<i}, C\right)\right), \quad (10)$$

Through this function, the distance between token prediction probabilities becomes larger, resulting in different branches producing different responses reflecting their own dialogue attributes and improving branch heterogeneity.

Negative Distillation with Orthogonal Projection Besides the explicit knowledge from the prediction layer, implicit knowledge embedded in the hidden states can also help the negative distillation process. In spite of the fact that different auxiliary branches acquire different attribute-specific knowledge, there should be some shared features in hidden states to support the basic abilities of sentence generation. Directly increasing the distance of hidden states between branches by negative distillation will damage the common knowledge for dialogue generation.

Therefore, we propose orthogonal negative distillation to protect the common features from interference inspired by Wang et al. (2019). Specifically, as shown in Figure 3, we project the hidden state $\mathbf{H}_A$ to the orthogonal space of hidden state $\mathbf{H}_B$ in order to get $\mathbf{H}_L$:

$$\mathbf{H}_L = \left(\mathbf{I} - \mathbf{H}_B\left(\mathbf{H}_B^T\mathbf{H}_B\right)^{-1}\mathbf{H}_B^T\right)\mathbf{H}_A \quad (11)$$

$\mathbf{H}_L$ contains the biased features of $\mathbf{H}_A$ which reflects its attribute-specific knowledge comparing with $\mathbf{H}_B$, getting rid of the shared features within them. On this basis, we conduct negative distillation with mean reverse square error (MRSE) (Li et al. 2022a) between $\mathbf{H}_L$ and $\mathbf{H}_B$, which is conducive to dialogue attribute learning for branch $\boldsymbol{A}$ while avoiding the common knowledge interface. The loss function is then defined as:

$$\mathcal{L}_{ND_{hidden}}(\mathbf{H}_L, \mathbf{H}_B) = \frac{1}{n}\sum_{i=1}^{n}\exp^{-SE(\mathbf{H}_L, \mathbf{H}_B)}, \quad (12)$$

where SE refers to square error. Note that we only perform ND on the last hidden states of decoder for training efficiency.

Optimization For the proposed collaborative dialogue learning framework, the overall objective function consists of two terms: a conventional cross entropy loss for dialogue generation and online knowledge distillation loss for collaborative learning. Specifically, the loss for master branch is:

$$\mathcal{L} = \mathcal{L}_{MLE} + \frac{1}{|M|}\sum^{m}\mathcal{L}_{PD}^{m}, \quad (13)$$

where $|M|$ is the number of auxiliary branches. While for each auxiliary branch:

$$\mathcal{L} = \mathcal{L}_{MLE} + \mathcal{L}_{PD}^{m} + \frac{1}{|M|-1}\sum^{m}\mathcal{L}_{ND_{pred}}^{m} + \frac{1}{|M|-1}\sum^{m}\mathcal{L}_{ND_{hidden}}^{m}. \quad (14)$$

All branches are trained simultaneously at each epoch until the master branch converges.

Experiment

Datasets

We evaluate the proposed method using two widely used dialogue datasets: **DailyDialog**, a collection of conversations that represent human daily communication (Li et al. 2017), and **OpenSubtitles**, which consists of large-scale dialogues extracted from movie subtitles (Tiedemann 2009). After data preprocessing, the number of context-response pairs in training/validation/test set is 68,066/6,820/6,841 for DailyDialog, and 200,000/20,000/10,000 for OpenSubtitles.

Implementation Details

All approaches are based on the Transformer-based sequence-to-sequence model (Vaswani et al. 2017). Each branch is built on the lightweight model architecture (Small Transformer): the encoder and decoder contain only 2 layers, in which the self-attention module has 4 attention heads and 1024 feed-forward units. The size of hidden states is set to 256. 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. The batch size is set to 64. The selection ratio for attribute-specific subset is 70%. For the temperature coefficient t, we simply set it to 1. Beam search with a size of 5 is used for decoding. We implement all approaches with Pytorch 1.11, and conduct all experiments on NVIDIA TITAN RTX.

Models	Dist-1	Dist-2	Dist-3	BLEU-1	BLEU-4	AVE	COH	H-1	H-2	H-3	KL	LF
Transformer	.0080	.0345	.0748	.2963	.4113	.8151	.7058	6.77	7.46	9.96	0.81	.0825
DML	.0167	.0669	.1296	.3154	.4221	.8164	.7069	6.97	7.87	10.50	0.50	.1427
CL-ILR	.0167	.0686	.1369	.3223	.4241	.8179	.7078	6.95	7.87	10.50	0.51	.1355
ONE	.0120	.0489	.0995	.3248	.4082	.8170	.7072	6.95	7.76	10.42	0.66	.1174
OKDDip	.0141	.0581	.1168	.3097	.4212	.8188	.7100	6.90	7.75	10.31	0.55	.1376
CDL-CI	.0191	.0815	.1679	.3139	**.4283**	.8182	.7074	6.89	7.83	10.45	0.41	.1514
CDL-CS	.0186	.0785	.1561	**.3317**	.4108	**.8198**	**.7177**	7.07	7.99	10.70	0.42	.1603
CDL-IS	**.0252**	**.1081**	**.2143**	.3184	.4261	.8179	.7121	**7.12**	**8.13**	**10.80**	**0.32**	**.1778**
Transformer	.0031	.0140	.0302	.3552	.3062	.7891	.7048	6.71	7.64	10.99	1.31	.0349
DML	.0044	.0171	.0344	.3494	.3248	.7907	.6801	6.41	7.11	10.16	1.58	.0363
CL-ILR	.0044	.0179	.0368	.3310	.3151	.7804	.6648	6.48	7.26	10.32	1.49	.0513
ONE	.0043	.0175	.0369	.3510	.3140	.7922	.6921	6.56	7.43	10.60	1.40	.0410
OKDDip	.0035	.0141	.0300	.3487	.3244	.7886	.6743	6.49	7.19	10.30	1.55	.0356
CDL-CI	**.0057**	**.0239**	**.0523**	.3474	.3254	**.7983**	**.7156**	6.73	**7.71**	**11.03**	**1.18**	**.0555**
CDL-CS	.0050	.0197	.0419	**.3552**	.3146	.7924	.6996	6.71	7.63	10.93	1.29	.0426
CDL-IS	.0050	.0211	.0460	.3443	**.3258**	.7893	.6923	**6.78**	7.68	10.95	1.29	.0524

Table 1: Automatic evaluation results on DailyDialog (Up) and OpenSubtitles (Down). The best/second-best results are bold/underlined. The branch number is 3. C refers to coherence, I for informativeness and S for specificity.

Comparison Methods

We compare our proposed collaborative dialogue learning (CDL) framework with following established collaborative learning approaches:

- DML (Zhang et al. 2018c) uses a pool of network-based students, where each student is an individual network and they asynchronously collaborate.
- CL-ILR (Song and Chai 2018) distills knowledge among multiple branches of a hierarchical network.
- ONE (Lan, Zhu, and Gong 2018) automatically generates gated ensemble logit from each branch as a soft target.
- OKDDip (Chen et al. 2020) proposes a two-level distillation strategy with multiple auxiliary peers and a group leader, while utilizing an attention module to construct inter-branch diversity.

Following previous work, we set the branch number is 3 for all the comparison models. For the proposed framework, it contains one master branch and two auxiliary branches with different dialogue attributes, i.e., coherency (C), informativeness (I) and specificity (S).

Automatic Evaluation

Metrics We first used automatic metrics to evaluate our method: **Dist-{1,2,3}** (distinct) (Li et al. 2016a) is a widely used metric that reflects the lexical diversity of the generated responses by calculating the proportion of unique unigrams/bigrams/trigrams. **BLEU** (Chen and Cherry 2014) measures n-gram overlap between the generated and the ground-truth responses. **AVE** (Embedding Average) (Liu et al. 2016) evaluates the semantic relationship of generated responses and ground-truth responses. **COH** (coherence) (Xu et al. 2018b) measures the cosine similarity between pairs of input and response. **H-{1,2,3}** (word entropy) (Serban et al. 2017b) measures the unigrams/bigrams/trigrams' non-genericness of responses. **KL** (KL divergence) (Csaky, Purgai, and Recski 2019) measures the distribution distance between the generated and the ground-truth response sets to reflect how well a model can approximate the ground-truth distribution. Note that the lower KL is better. **LF** (low-frequency token ratio) (Li et al. 2020) further 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.

Results Table 1 shows the results obtained at the lowest point of the validation loss. It illustrates that our framework outperforms all baselines by a significant margin on both datasets. Note that four collaborative learning baselines perform better than vanilla Transformer model, which proves that the group-base distillation can improve model performance greatly. On this basis, the proposed approach can further enhance the performance by introducing dialogue attributes learning and dual knowledge distillation. And the improvement of different dialogue attributes can be reflected by the corresponding metrics.

Human Evaluation

To further verify the effectiveness of our method in comparison to previous collaborative learning methods, we also conduct human evaluations apart from automatic evaluations. We randomly select 50 samples from the test set of DailyDialog, and three well-educated annotators are invited to judge which of the overall response quality generated by CDL and baselines is better (i.e., win, tie or loss) in terms of coherence, informativeness and fluency.

Table 2 summarizes the human evaluation results. In our experience, we have noticed that a dialogue model trained using our proposed learning framework is more capable

vs. Models	Win	Tie	Loss	Kappa
Transformer	0.82	0.15	0.03	0.5487
DML	0.47	0.42	0.11	0.6651
CL-ILR	0.43	0.53	0.05	0.5393
ONE	0.50	0.39	0.11	0.6177
OKDDip	0.45	0.43	0.11	0.5743

Table 2: Human evaluations results on DailyDialog. Our framework has a higher win rate than baselines.

Models	Dist-1	Dist-2	LF	KL	H-1
w/o Attributes	.0177	.0756	.1367	0.38	6.94
w/o OP	.0203	.0869	.1629	0.37	6.99
w/o $\mathcal{L}_{ND_{hidden}}$	.0222	.0931	.1621	0.41	7.03
w/o $\mathcal{L}_{neg}$	.0215	.0890	.1776	0.43	7.05
Full Version	.0252	.1081	.1778	0.32	7.12

Table 3: Ablation study results of the proposed collaborative dialogue learning framework.

of producing responses that are human-preferred. We use Fleiss's kappa (Fleiss 1971) to measure the inter-annotator agreement, which indicates that the annotators came to a fair agreement in the judgment.

Analysis

In order to better understand the effectiveness of the collaborative dialogue learning, we carry out extensive analysis of DailyDialog.

Ablation study We study the effects of different parts of proposed framework by ablating the dialogue attribute learning (w/o attributes), the orthogonal projection (w/o OP), the hidden state distillation (w/o $\mathcal{L}_{ND_{hidden}}$), and the whole negative distillation (w/o $\mathcal{L}_{Neg}$). The results in Table 3 show that all proposed techniques are useful for improving the response quality. The significant decline in w/o attributes indicates that the knowledge of specific dialogue property is very important for CDL. w/o $\mathcal{L}_{Neg}$ is better than w/o OP, indicating that orthogonal projection is a key technique to capture biased features without harming common knowledge.

Comparison with traditional KD Traditional knowledge distillation is an efficient method to obtain a small but effective model. The results from Table 4 show that CDL outperform KD (Teacher is Base Transformer) and Small with the same inference cost and the relative heavy Base model without a well-trained teacher.

Branch Number Study We explore the performance of proposed CDL in other number of branches. The results from Table 5 shows that, regardless of the number of branches, performance of CDL is better than baselines. The inferior performance with branch number 4 (compared with Table 1) is that the influence of positive distillation from master branch is much lower than with branch number 3, given that the more auxiliary branches the more negative

Models	Dist-1	Dist-2	LF	KL	H-1
Small	0.0080	0.0345	0.0825	0.81	6.77
Base	0.0101	0.0471	0.1084	0.56	6.83
KD	0.0124	0.0564	0.1336	0.47	6.94
CDL-IS	0.0252	0.1081	0.1778	0.32	7.12

Table 4: Comparison results with traditional knowledge distillation.

Models	Dist-1	Dist-2	LF	KL	H-1
DML	0.0154	0.0644	0.1368	0.51	6.95
CL-ILR	0.0154	0.0625	0.1232	0.52	6.90
ONE	0.0104	0.0432	0.1022	0.69	6.88
OKDDip	0.0132	0.0576	0.1348	0.48	6.96
CDL-C	0.0195	0.0833	0.1492	0.39	6.97
CDL-S	0.0182	0.0730	0.1515	0.54	**7.03**
CDL-I	**0.0207**	**0.0889**	**0.1834**	**0.37**	6.93

Models	Dist-1	Dist-2	LF	KL	H-1
DML	0.0198	0.0779	0.1457	0.44	6.93
CL-ILR	0.0159	0.0637	0.1362	0.56	6.97
ONE	0.0128	0.0523	0.1194	0.61	6.92
OKDDip	0.0167	0.0655	0.1274	0.56	6.99
CDL-CSI	**0.0211**	**0.0869**	**0.1672**	**0.41**	**6.99**

Table 5: Evaluation Results with branch number 2 (Up) and 4 (Down).

distillation will be conducted. The promising way to solve this problem is to increase the weight of positive KD loss and decrease negative KD's, and we leave it as future work.

Model Diversity Analysis we show the diversity that our method brings in a more intuitive way. We use the Euclidean distance (L2) between branches as a quantitative criterion for diversity, as performed in OKDDip (Chen et al. 2020). Table 6 shows the averaged L2 distances when each model reaches to convergence. Our mutual diversity is significantly greater than other methods, which suggests that the homogenization problem between branches have greatly relieved through multi-attribute learning and negative distillation.

Case Study Table 7 presents some responses generated by the proposed framework and baselines. Transformer prefers generic and meaningless responses. Other baselines lack of concerning one perspective of dialogue. In contrast, our CDL comprehensively consider the multiple perspectives, thus resulting in diverse and coherent responses. The results demonstrate the effectiveness of CDL.

Related Work

Dialogue Models

There are three major categories of previous work on enhancing the quality of responses. The first redesigns the model structure to facilitate the modeling of the dialogue pairs (Serban et al. 2017a; Tao et al. 2018; Gao et al. 2019).

	DML	CL-ILR	ONE	OKDDip	CDL
L2	0.163	0.141	0.198	0.165	0.308

Table 6: Branches diversity for CDL and other collaborative baselines.

Input: Here are all kinds of jades. Choose whatever you like, please.
Transformer: Ok. **DML:** Thank you. **CL-ILR:** Thank you very much. **ONE:** Thank you. I have a good idea. **OKDDip:** Thank you very much. **CDL:** Thank you, sir. I 'll be glad to do that.
Input: What do you recommend? We'd like some of your local specialities.
Transformer: That's good. **DML:** That's a good idea. What do you want to do? **CL-ILR:** I don't know what you mean. **ONE:** Well, I think I'd like to have a cup of coffee. **OKDDip:** Well, I've heard about it, but I don't think it's a good idea. **CDL:** That's great. I'd like to have some cheese.

Table 7: Examples of generated responses.

The second further proposes advanced objectives aligning with the goals of the conversation more effectively, such as MMI (Li et al. 2016a), CVAE (Serban et al. 2017b; Zhao, Zhao, and Eskénazi 2017; Gu et al. 2019; Sun et al. 2021), RL (Li et al. 2016b; Zhang et al. 2018a; Liu et al. 2020), and GAN (Xu et al. 2017, 2018a; Feng et al. 2020a). The third tries to endow the responses with topic (Xing et al. 2017; Feng et al. 2020b), emotion (Zhou et al. 2018; Rashkin et al. 2019), and persona (Qian et al. 2017; Zhang et al. 2018b; Song et al. 2020). Recently, data filtering has been introduced for dialogue learning, which discards samples regarded as low-quality by a scoring method to reflect corresponding dialogue attributes. Csaky, Purgai, and Recski (2019) proposes an entropy-based scoring method to remove generic utterances from the training data. See et al. (2019) designs a scoring method to measure the specificity of samples. Akama et al. (2020) combines the cosine distance and the keyword co-occurrence of the dialogue pairs to evaluate the coherence. Shen et al. (2021) presents a fusing approach to data filtering and Li et al. (2022b) utilizes the scoring methods to enhance rather than filter data.

Knowledge Distillation

In recent years, knowledge distillation (Hinton, Vinyals, and Dean 2015; Freitag, Al-Onaizan, and Sankaran 2017) has been widely adopted by researchers to accelerate and compress models (Jiao et al. 2020; Sanh et al. 2019). As these predicted distributions contain ranking information on similarities among categories, it treats the predictions as knowledge learned by the teacher network. As a result, it enforces similar predictions on the student network in order to transfer this knowledge. By providing more knowledge to the student network from different sources, the work follows this idea. To supervise the student network, FitNets (Romero et al. 2015) uses both predictions and intermediate representations learned by the teacher network. Kim and Rush (2016) propose using sequence-level knowledge generated from the generated sequences to guide student network training in the Seq2Seq model. Furthermore, self-knowledge distribution (Hahn and Choi 2019) demonstrates that students are able to improve performance by using their own knowledge. When it comes to dialogue generation, Feng et al. (2021b) guide the dialogue model towards better generalization by introducing bidirectional distillation and Li et al. (2022a) propose negative distillation to enhance the diversity of responses. Rather than pre-training a large teacher, we use collaborative learning and distill knowledge from a group of branches.

Collaborative Learning

The concept of collaborative learning (Anil et al. 2018; Lan, Zhu, and Gong 2018; Feng et al. 2021a; Song and Chai 2018; Chen et al. 2020) is more lightweight in terms of the stages of learning compared to that of conventional knowledge distillation. By doing so, it facilitates finding a robust local minimum for each student, resulting in greater generalization performance. Student networks are currently being implemented in two mainstream settings. One is network-based (Zhang et al. 2018c), in which students form independent networks and parameter capacity increases linearly with the number of students; the other is branch-based (CL-ILR (Song and Chai 2018) and ONE (Lan, Zhu, and Gong 2018)), where the bottom layers of students are shared. Feng et al. (2021a) enable more flexible representation sharing with random routing mechanism, where layers at any level can be shared by different involved students. However, previous work focused only on classification, resulting in the same training objective for all branches of the framework with independent identical distributions (i.i.d.). As a result, different branches will tend to converge to similar feature representations (Li et al. 2016c; Lan, Zhu, and Gong 2018; Chen et al. 2020). Different from them, we propose attrbute-related branch learning strategy and dual knowledge distillation to solve the homogenization problem.

Conclusion

We present a novel collaborative dialogue learning paradigm to improve the quality of generated responses in terms of three major dialogue attributes. CDL replaces traditional knowledge distillation with collaborative group-based distillation for lightweight knowledge interaction, and the attribute-aware knowledge is captured and transferred through auxiliary branches. Dual group-based knowledge distillation is proposed for better guiding auxiliary branches to learn attribute-specific knowledge. Besides, we further boost the performance of negative distillation by utilizing orthogonal projection to avoid harming the common knowledge. Extensive experiments validate the superiority of our proposed method over prior collaborative learning work.

Acknowledgments

This research was supported by the Beijing Natural Science Foundation (No.4222037, L181010) and the BIT Research and Innovation Promoting Project (Grant No.2022YCXY021).

References

Akama, R.; Yokoi, S.; Suzuki, J.; and Inui, K. 2020. Filtering Noisy Dialogue Corpora by Connectivity and Content Relatedness. In *EMNLP*, 941–958.

Anil, R.; Pereyra, G.; Passos, A.; Ormándi, R.; Dahl, G. E.; and Hinton, G. E. 2018. Large scale distributed neural network training through online distillation. In *ICLR*.

Bouma, G. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. In *(GSCL)*, 31–40.

Chen, B.; and Cherry, C. 2014. A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU. In *Ninth Workshop on Statistical Machine Translation*, 362–367.

Chen, D.; Mei, J.; Wang, C.; Feng, Y.; and Chen, C. 2020. Online Knowledge Distillation with Diverse Peers. In *AAAI*, 3430–3437.

Csaky, R.; Purgai, P.; and Recski, G. 2019. Improving Neural Conversational Models with Entropy-Based Data Filtering. In *ACL (1)*, 5650–5669.

Feng, S.; Chen, H.; Li, K.; and Yin, D. 2020a. PosteriorGAN: Towards Informative and Coherent Response Generation with Posterior Generative Adversarial Network. In *AAAI*, 7708–7715.

Feng, S.; Chen, H.; Ren, X.; Ding, Z.; Li, K.; and Sun, X. 2021a. Collaborative Group Learning. In *AAAI*, 7431–7438.

Feng, S.; Ren, X.; Chen, H.; Sun, B.; Li, K.; and Sun, X. 2020b. Regularizing Dialogue Generation by Imitating Implicit Scenarios. In *EMNLP*, 6592–6604.

Feng, S.; Ren, X.; Li, K.; and Sun, X. 2021b. Multi-View Feature Representation for Dialogue Generation with Bidirectional Distillation. In *AAAI*, 12812–12820.

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

Freitag, M.; Al-Onaizan, Y.; and Sankaran, B. 2017. Ensemble Distillation for Neural Machine Translation. *CoRR*.

Gao, X.; Lee, S.; Zhang, Y.; Brockett, C.; Galley, M.; Gao, J.; and Dolan, B. 2019. Jointly Optimizing Diversity and Relevance in Neural Response Generation. In *NAACL-HLT (1)*, 1229–1238.

Gu, X.; Cho, K.; Ha, J.; and Kim, S. 2019. DialogWAE: Multimodal Response Generation with Conditional Wasserstein Auto-Encoder. In *ICLR (Poster)*.

Hahn, S.; and Choi, H. 2019. Self-Knowledge Distillation in Natural Language Processing. In *RANLP*, 423–430.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

Jiao, X.; Yin, Y.; Shang, L.; Jiang, X.; Chen, X.; Li, L.; Wang, F.; and Liu, Q. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of EMNLP*, 4163–4174.

Kim, Y.; and Rush, A. M. 2016. Sequence-Level Knowledge Distillation. In *EMNLP*, 1317–1327.

Lan, X.; Zhu, X.; and Gong, S. 2018. Knowledge Distillation by On-the-Fly Native Ensemble. In *NeurIPS*, 7528–7538.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016a. A Diversity-Promoting Objective Function for Neural Conversation Models. In *HLT-NAACL*, 110–119.

Li, J.; Monroe, W.; Ritter, A.; Jurafsky, D.; Galley, M.; and Gao, J. 2016b. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*, 1192–1202.

Li, Y.; Feng, S.; Sun, B.; and Li, K. 2022a. Diversifying Neural Dialogue Generation via Negative Distillation. In *NAACL*, 407–418.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *IJCNLP(1)*, 986–995.

Li, Y.; Sun, B.; Feng, S.; and Li, K. 2022b. Stop Filtering: Multi-View Attribute-Enhanced Dialogue Learning. *CoRR*, abs/2205.11206.

Li, Y.; Yosinski, J.; Clune, J.; Lipson, H.; and Hopcroft, J. E. 2016c. Convergent Learning: Do different neural networks learn the same representations? In *ICLR*.

Li, Z.; Wang, R.; Chen, K.; Utiyama, M.; Sumita, E.; Zhang, Z.; and Zhao, H. 2020. Data-dependent Gaussian Prior Objective for Language Generation. In *ICLR*.

Liu, C.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In Su, J.; Carreras, X.; and Duh, K., eds., *]EMNLP*, 2122–2132.

Liu, Q.; Chen, Y.; Chen, B.; Lou, J.; Chen, Z.; Zhou, B.; and Zhang, D. 2020. You Impress Me: Dialogue Generation via Mutual Persona Perception. In *ACL*, 1417–1427.

Qian, Q.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *CoRR*, abs/1706.02861.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *ACL*, 5370–5381.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. In Bengio, Y.; and LeCun, Y., eds., *ICLR*.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

See, A.; Roller, S.; Kiela, D.; and Weston, J. 2019. What makes a good conversation? How controllable attributes affect human judgments. In *NAACL-HLT*, 1702–1723.

Serban, I. V.; Klinger, T.; Tesauro, G.; Talamadupula, K.; Zhou, B.; Bengio, Y.; and Courville, A. C. 2017a. Multiresolution Recurrent Neural Networks: An Application to Dialogue Response Generation. In *AAAI*, 3288–3294. AAAI Press.

Serban, I. V.; Sordoni, A.; Lowe, R.; Charlin, L.; Pineau, J.; Courville, A. C.; and Bengio, Y. 2017b. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*, 3295–3301.

Shang, L.; Lu, Z.; and Li, H. 2015. Neural Responding Machine for Short-Text Conversation. In *ACL (1)*, 1577–1586.

Shen, L.; Zhan, H.; Shen, X.; Chen, H.; Zhao, X.; and Zhu, X. 2021. Identifying Untrustworthy Samples: Data Filtering for Open-domain Dialogues with Bayesian Optimization. In *CIKM*, 1598–1608.

Song, G.; and Chai, W. 2018. Collaborative Learning for Deep Neural Networks. In *NeurIPS*, 1837–1846.

Song, H.; Wang, Y.; Zhang, W.; Liu, X.; and Liu, T. 2020. Generate, Delete and Rewrite: A Three-Stage Framework for Improving Persona Consistency of Dialogue Generation. In *ACL*, 5821–5831.

Sordoni, A.; Galley, M.; Auli, M.; Brockett, C.; Ji, Y.; Mitchell, M.; Nie, J.; Gao, J.; and Dolan, B. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *HLT-NAACL*, 196–205.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1): 1929–1958.

Sun, B.; Feng, S.; Li, Y.; Liu, J.; and Li, K. 2021. Generating Relevant and Coherent Dialogue Responses using Self-Separated Conditional Variational AutoEncoders. In *ACL/IJCNLP*, 5624–5637.

Sun, S.; Cheng, Y.; Gan, Z.; and Liu, J. 2019. Patient Knowledge Distillation for BERT Model Compression. In *EMNLP-IJCNLP*, 4322–4331.

Tao, C.; Gao, S.; Shang, M.; Wu, W.; Zhao, D.; and Yan, R. 2018. Get The Point of My Utterance! Learning Towards Effective Responses with Multi-Head Attention Mechanism. In *IJCAI*, 4418–4424.

Tiedemann, J. 2009. News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, 237–248.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *NIPS*, 5998–6008.

Vinyals, O.; and Le, Q. V. 2015. A Neural Conversational Model. In *ICML Deep Learning Workshop*.

Wang, H.; He, Z.; Lipton, Z. C.; and Xing, E. P. 2019. Learning Robust Representations by Projecting Superficial Statistics Out. In *ICLR*.

Wu, G.; and Gong, S. 2021. Peer Collaborative Learning for Online Knowledge Distillation. In *AAAI*, 10302–10310.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W. 2017. Topic Aware Neural Response Generation. In *AAAI*, 3351–3357.

Xu, J.; Ren, X.; Lin, J.; and Sun, X. 2018a. Diversity-Promoting GAN: A Cross-Entropy Based Generative Adversarial Network for Diversified Text Generation. In *EMNLP*, 3940–3949.

Xu, X.; Dusek, O.; Konstas, I.; and Rieser, V. 2018b. Better Conversations by Modeling, Filtering, and Optimizing for Coherence and Diversity. In *EMNLP*, 3981–3991.

Xu, Z.; Liu, B.; Wang, B.; Sun, C.; Wang, X.; Wang, Z.; and Qi, C. 2017. Neural Response Generation via GAN with an Approximate Embedding Layer. In *EMNLP*, 617–626.

Zhang, H.; Lan, Y.; Guo, J.; Xu, J.; and Cheng, X. 2018a. Reinforcing Coherence for Sequence to Sequence Model in Dialogue Generation. In *IJCAI*, 4567–4573.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018b. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *ACL*, 2204–2213.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018c. Deep Mutual Learning. In *CVPR*, 4320–4328.

Zhao, T.; Zhao, R.; and Eskénazi, M. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL (1)*, 654–664.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *AAAI-18*, 730–739.