Multi-View Feature Representation for Dialogue Generation with Bidirectional Distillation

Shaoxiong Feng,[1] **Xuancheng Ren,**[2] **Kan Li,**[1] **Xu Sun**[2,3]
[1]School of Computer Science & Technology, Beijing Institute of Technology
[2]MOE Key Laboratory of Computational Linguistics, School of EECS, Peking University
[3]Center for Data Science, Peking University
{shaoxiongfeng, likan}@bit.edu.cn, {renxc, xusun}@pku.edu.cn

Abstract

Neural dialogue models suffer from low-quality responses when interacted in practice, demonstrating difficulty in generalization beyond training data. Recently, knowledge distillation has been used to successfully regularize the student by transferring knowledge from the teacher. However, the teacher and the student are trained on the same dataset and tend to learn similar feature representations, whereas the most general knowledge should be found through differences. The finding of general knowledge is further hindered by the unidirectional distillation, as the student should obey the teacher and may discard some knowledge that is truly general but refuted by the teacher. To this end, we propose a novel training framework, where the learning of general knowledge is more in line with the idea of reaching consensus, i.e., finding common knowledge that is beneficial to different yet all datasets through diversified learning partners. Concretely, the training task is divided into a group of subtasks with the same number of students. Each student assigned to one subtask not only is optimized on the allocated subtask but also imitates multi-view feature representation aggregated from other students (i.e., student peers), which induces students to capture common knowledge among different subtasks and alleviates the over-fitting of students on the allocated subtasks. To further enhance generalization, we extend the unidirectional distillation to the bidirectional distillation that encourages the student and its student peers to co-evolve by exchanging complementary knowledge with each other. Empirical results and analysis demonstrate that our training framework effectively improves the model generalization without sacrificing training efficiency.

Introduction

Neural dialogue generation has drawn increasing attention, but current dialogue models still struggle with generalization, e.g., frequently producing generic and meaningless responses in inference (Mou et al. 2016; Li et al. 2016a; Serban et al. 2017b). Unlike machine translation or summarization, dialogue generation has more freedom and diversity in the semantic and linguistic aspects of responses. Without specific training guidance, they are prone to over-fitting certain aspects of corpora (e.g., naive target sequence prediction) that show distinct distributions between training and test data. Therefore, it is usually hard for these models to learn generalizable features and they may easily get stuck in a narrow local minimum that is fragile to data perturbation (Chaudhari et al. 2017; Keskar et al. 2017).

To alleviate this problem, one line of work introduces prior *common knowledge* of the real world to facilitate the model generalization, such as redesigning objective functions (e.g., maximize mutual information or coherence) instead of only fitting target sequence (Li et al. 2016b; Feng et al. 2020a), and modifying generation order (e.g., hierarchical or syntactic-based generation) instead of naive left-to-right generation (Su et al. 2018; Welleck et al. 2019). Intuitively, common knowledge is a class of knowledge that benefits both the training and the test data, as it is reflected generally in the whole corpus and not merely only works for the training data. With the common knowledge as constraint, models can be guided to learn towards a better direction that can bridge the gap of training and test data distributions more easily. Dubey et al. (2018) also verified that common knowledge from the real world plays an important role in models' quickly learning unfamiliar video games.

However, prior common knowledge is hard to manually define, since it varies with tasks and domains and can be a limiting factor if defined wrong. Recently, another line of work (Arora, Khapra, and Ramaswamy 2019; Tahami, Ghajar, and Shakery 2020; Chen et al. 2020), using *knowledge distillation* (KD; Ba and Caruana 2014; Hinton, Vinyals, and Dean 2015), has successfully extracted knowledge from a pre-trained teacher model to regularize the student model for better generalization. The student model aims to achieve a balance of using raw knowledge from the training data and distilled knowledge from the teacher model, which makes the student capture more common or generalizable knowledge and perform better in testing. Compared with previous work on introducing common knowledge, KD is more straightforward and extensible. However, conventional KD still faces two drawbacks:

- Lack of feature diversity: Because both the student and the teacher are trained on the same dataset, they may learn similar feature representations, which means the knowledge is not sufficiently diverse to conduct an effective regularization on the feature learning of the student.
- Lack of student feedback: Previous work (Romero et al.

2015; Yim et al. 2017; Furlanello et al. 2018) has proved that the student can obtain better generalization performance than the teacher, but KD still runs unidirectionally, which may damage the more generalizable knowledge learned by the student and hinder the performance improvement of the teacher.

In this work, we propose a novel training framework to tackle these problems by generating multi-view feature representations and co-evolution via bidirectional distillation. Figure 1 illustrates the proposed framework. To obtain multi-view feature representations, the training task is divided into a group of subtasks, i.e., subsets of training data, and each subtask is assigned a corresponding student model, which learns knowledge specific to different subtasks. The students are also enforced to perform as well as possible in the unseen subtasks by imitating the predictions of other students, so that common knowledge can be drawn collaboratively, which also prevents aggressive overfitting. In addition, a bidirectional knowledge distillation is further applied, which encourages the student and its student peers to exchange complementary knowledge and together evolve towards better generalization, which further eliminates the need of pre-training in conventional KD. Furthermore, the student peers for a student in knowledge distillation are randomly selected at each iteration to prevent the degeneration and homogenization (Kuncheva and Whitaker 2003; Schwenker 2013) of multi-view feature representations due to the bidirectional learning settings. In such way, the proposed training framework enables students to jointly learn diverse yet generalizable knowledge from multi-view feature representations from different data compositions.

Our main contributions are as follows:

- We propose a novel training framework that reconstructs the training task as a group of subtasks and aggregates multi-view feature representations from randomly-selected student peers to regularize students for more generalizable knowledge.
- The framework is enhanced by bidirectional knowledge distillation that allows the student to provide feedback to its student peers and makes both ends able to co-evolve.
- We conducted extensive experiments and analysis to validate the effectiveness of multi-view feature representations and bidirectional distillation and demonstrate why these mechanisms work well.

Method

In this section, we describe how to effectively capture common knowledge for improving the generalization of dialogue models. We first introduce multi-view feature representation for diverse knowledge, then propose bidirectional distillation to regularize both students and their partners, and finally present the optimization objective.

Multi-View Feature Representation

For generative conversation models, given a training example $(x, y) \in \mathcal{D}$, where x is the source sequence, i.e., the dialogue history, y is the target sequence, i.e., the corresponding response, and $\mathcal{D}$ is the whole training dataset consisting of M examples, i.e., $|\mathcal{D}| = M$, the learning objective is to minimize the following negative log-likelihood:

$$\mathcal{L}_{\mathrm{NLL}}(x, y; \theta) = -\sum\nolimits_{j=1}^{|y|} \log p(y_j | y_{<j}, x; \theta), \qquad (1)$$

where θ is the learnable parameters.

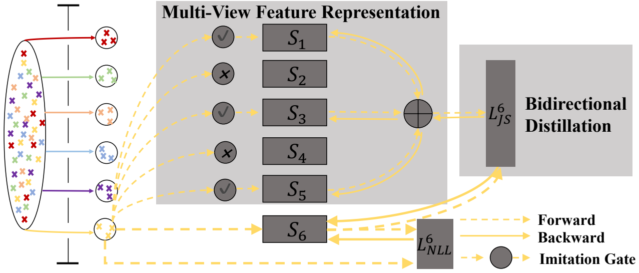

Figure 1: *An overview of multi-view feature representation and bidirectional distillation.*

In conventional knowledge distillation (Furlanello et al. 2018; Kim et al. 2020), a teacher model is first pre-trained on the whole dataset, whose parameters we denote as θ_t, and another model, parameterized as θ_s, is taken as the student that further aligns its prediction with the teacher prediction using the Kullback-Leibler divergence (Kullback and Leibler 1951):

$$\mathcal{L}_{\mathrm{KL}}(x, y, \theta_t; \theta_s) = \sum_{j=1}^{|y|} \sum_{w \in \mathcal{V}} p(w | X_j; \theta_t) \log \frac{p(w | X_j; \theta_t)}{p(w | X_j; \theta_s)},$$

where w is a word in the vocabulary $\mathcal{V}$ and X_j is defined as $(y_{<j}, x)$. The student is tasked to cover all the knowledge from the teacher due to the mean-seeking behaviour of the KL divergence, while the teacher is kept fixed in the distillation.

As the teacher and the student are trained using exactly the same data, they intend to learn similar feature representations or knowledge (Li et al. 2016c; Morcos, Raghu, and Bengio 2018). However, it should be crucial for the students to obtain sufficiently diverse sources of knowledge to extract common knowledge that generalizes to unseen examples, which is not available in such distillation settings and limits the effect of the regularization from the teacher. To address this problem, we propose to learn multi-view feature representations for the students to find more common knowledge by aligning their predictions with diverse partners. Essentially, the training dataset is broken down into N subsets $\{\mathcal{D}^n\}_{n=1}^N$, where $\cup_{n=1}^N \mathcal{D}^n = \mathcal{D}$ and $\cap_{n=1}^N \mathcal{D}^n = \varnothing$, that compose varied subtasks, each assigned an individual student. Each student is trained using supervised examples solely from its corresponding subset so that for the macro task we can get diverse representations from micros views. The supervised learning of a student is conducted as follows:

$$\mathcal{L}^n_{\mathrm{NLL}}(x^k, y^k; \theta_n) = -\sum_{j=1}^{|y^k|} \log p(y^k_j | y^k_{<j}, x^k; \theta_n), \qquad (2)$$

where $(x^k, y^k) \in \mathcal{D}^n$ and n identifies the student and the subset.

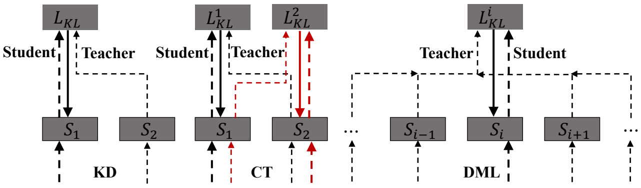

Figure 2: Comparison of knowledge aggregation and transfer among vanilla KD, Co-Teaching (CT), and deep mutual learning (DML). The dashed and solid lines represent forward and backward propagation, respectively. The differently colored lines in CT represent students S_1 and S_2 are trained on independent batches of training data.

In turn, we aggregate the corresponding multi-view feature representation for each student to imitate by averaging over predictions of all other students. However, aggregation with naive averaging will lead multi-view feature representations of all students to be similar, which can also cause students to homogenize (Kuncheva and Whitaker 2003; Schwenker 2013). Therefore, we further introduce the imitation gate $g(\cdot)$ (shown in Figure 1) to maintain the diversity of multi-view feature representations. Specifically, each student randomly imitates a subgroup of students at each iteration, which subjects to $g(p) \sim \text{Bernoulli}(p)$, and the number of the imitated students is decided by the imitation probability p. During training, students are regularized by different and dynamic multi-view feature representations, which alleviates the homogenization of students. Besides, the imitation gate also reduces the computational cost (i.e., forward and backward propagation). For example, we can keep the computational cost constant by adjusting the imitation probability when the number of students increases. The distillation loss of each student is computed as:

$$\mathcal{L}^n_{\text{KL}}(x^k, y^k, \theta_{/n}; \theta_n) = \sum_{j=1}^{|y^k|} \sum_{w \in \mathcal{V}} p(w|X^k_j; \theta_{/n}) \frac{p(w|X^k_j; \theta_{/n})}{p(w|X^k_j; \theta_n)},$$

where $p(w|X^k_j; \theta_{/n})$ is the aggregated probability distribution from other students, $/n$ denotes the students other than the student n, and X^k_j denotes $(y^k_{<j}, x^k) \in \mathcal{D}^n$, which ensures that the student n never observes data directly from other subsets. $p(w|X^k_j; \theta_{/n})$ is calculated as:

$$p(w|X^k_j; \theta_{/n}) \triangleq \frac{1}{H} \sum_{i=1\ldots N, i \neq n} g^i(p) p(w|X^k_j; \theta_i), \tag{3}$$

where $H = \sum_{i=1\ldots N, i \neq n} g^i(p)$ is the number of the imitated students. By randomly selecting distilled students at each iteration, a single student is kept from the access of a global view of all the data and may maintain its specialty rather than homogenize.

Bidirectional Distillation

Previous work, such as vanilla KD (Hinton, Vinyals, and Dean 2015), Co-Teaching (CT) (Feng et al. 2019), and deep mutual learning (DML) (Zhang et al. 2018c), only conducts the unidirectional knowledge distillation from the teacher to the student, illustrated in Figure 2. However, according to prior research (Romero et al. 2015; Yim et al. 2017; Furlanello et al. 2018), the student can achieve better performance than the teacher. It suggests the unidirectional knowledge distillation has the risk of damaging the more generalizable knowledge that the student has learned.

To alleviate this problem, we propose the bidirectional knowledge distillation that induces both sides to reach a consensus by simultaneously regularizing each other, rather than one side monotonously imitating the other side. Similar to feature fusion methods (Hou, Liu, and Wang 2017), we first fuse the prediction $p(w|X^k_j; \theta_n)$ of the student n and the prediction $p(w|X^k_j; \theta_{/n})$ of corresponding multi-view feature representation to construct more generalizable knowledge $p(w|X^k_j; \theta_n, \theta_{/n})$:

$$p(w|X^k_j; \theta_n, \theta_{/n}) \triangleq \frac{1}{2} p(w|X^k_j; \theta_n) + \frac{1}{2} p(w|X^k_j; \theta_{/n}) \tag{4}$$

Then we enforce the student to imitate the fused knowledge $p(w|X^k_j; \theta_n, \theta_{/n})$, which can be expressed as:

$$\begin{aligned} \mathcal{L}^n_{\text{KL}}(x^k, y^k, \theta_{/n}; \theta_n) = & \sum_{j=1}^{|y^k|} \sum_{w \in \mathcal{V}} p(w|X^k_j; \theta_n, \theta_{/n}) \cdot \\ & \log \frac{p(w|X^k_j; \theta_n, \theta_{/n})}{p(w|X^k_j; \theta_n)} \end{aligned} \tag{5}$$

Besides, we also allow the teacher (i.e., the partners) to be regularized and conduct parameter updates. The final distillation loss is formulated as the Jensen–Shannon (JS) divergence (Dagan, Lee, and Pereira 1997):

$$\begin{aligned} \mathcal{L}^n_{\text{JS}}(x^k, y^k; \theta_n, \theta_{/n}) = & \frac{1}{2} \mathcal{L}^n_{\text{KL}}(x^k, y^k, \theta_{/n}; \theta_n) + \\ & \frac{1}{2} \mathcal{L}^n_{\text{KL}}(x^k, y^k, \theta_n; \theta_{/n}) \end{aligned} \tag{6}$$

In CT and DML, students can also provide knowledge for the imitated student due to iterative parameter updates. However, iterative parameter updates have two disadvantages: 1) It obviously slows down the training speed. 2) Compared with it, simultaneous parameter updates can speed up the convergence and obtain better performance (Mescheder, Nowozin, and Geiger 2017; Nagarajan and Kolter 2017).

Optimization

In this section, combining the NLL loss in Equation 2 with the distillation loss in Equation 6, we give the final optimization objective as follows:

$$\mathcal{L} = \sum_{n=1}^{N} (\mathcal{L}^n_{\text{NLL}} + T^2 * \mathcal{L}^n_{\text{JS}}), \tag{7}$$

where T^2 is a scalar coefficient. It is used to maintain an equilibrium between the NLL loss $\mathcal{L}^n_{\text{NLL}}$ and the distillation loss $\mathcal{L}^n_{\text{JS}}$ because we use $\frac{1}{T}$ to soften the probability distribution when calculating KL divergence.

Unlike vanilla KD and CT, we do not pre-train any student model as DML, which significantly saves the training time. Once the whole training set is divided into a group of subtasks, each student learns the assigned subtask and conducts the bidirectional distillation simultaneously. All students update parameters in parallel until convergence.

Experiment

We conduct experiments on two high-quality open-domain dialogue datasets, DailyDialog and PersonaChat, compared with state-of-the-art methods, and provide extensive analysis to examine the effect of the proposed method.

Datasets

We adopt two commonly-used dialogue datasets:

- **DailyDialog** (Li et al. 2017b) covers a variety of daily scenarios, such as work, health, and politics. We first extract the *(dialogue history, response)* pairs from the raw dataset. Each pair consists of two consecutive dialogue turns, in which the first turn and the second turn represent dialogue history and response, respectively. Then we limit the length of dialogue turns to $[5, 25]$ by discarding the pairs whose response is shorter than 5 words and truncating the turns whose length is longer than 25 words. Finally, the processed dataset contains 50K, 4.5K, and 4.3K pairs for training, validation, and testing, respectively.
- **PersonaChat** (Zhang et al. 2018a) is collected by two crowdsourced workers chit-chatting with each other, conditioned on the assigned personas. In our experiments, we only use the conversation text and process it as DailyDialog. The processed dataset contains 106K, 13K, and 12.5K pairs for training, validation, and testing.

Baselines

We re-implemented the following four methods and compared them with the proposed **MRBD** (**M**ulti-View Feature **R**epresentation and **B**idirectional **D**istillation):

- **Seq2Seq+Att** uses a vanilla Seq2Seq model (Sutskever, Vinyals, and Le 2014) with attention mechanism (Bahdanau, Cho, and Bengio 2015). The encoder and the decoder are based on a 2-layer bidirectional GRU (Cho et al. 2014) and a 2-layer unidirectional GRU, respectively. The size of hidden units is 500.
- **KD** uses two dialogue models as the student and the teacher, similar to Tahami, Ghajar, and Shakery (2020). The student learns from both the ground-truth responses and the probability distributions of the teacher.
- **CT** stands for the Co-Teaching training framework (Feng et al. 2019), in which two students are trained on independent training sets, and they provide complementary knowledge for each other. In practice, one student can still access the data assigned to the other student due to the whole training set shuffled once in each epoch.
- **DML** means Deep Mutual Learning (Zhang et al. 2018c), which constructs a group of students trained on the same training set. Each student learns from both the ground-truth responses and the knowledge equally aggregated from all other students. All students update the parameters iteratively, which means one student needs to recalculate a new prediction for the next students to imitate after updating its parameters.

In our experiments, for a fair comparison, models in baselines and our method comprise the Seq2Seq-based generative dialogue models with the same settings as Seq2Seq+Att. Besides, we use the KL divergence to calculate the distance of probability distributions for all baselines.

DailyDialog	Dist-1	Dist-2	Ent-1	Ent-2	Dis-1	Dis-2
Seq2Seq+Att	4.054	27.962	7.689	12.773	0.148	0.454
KD	4.219	29.007	7.732	12.892	0.142	0.395
CT	4.591	29.387	7.864	13.087	0.323	0.477
DML	4.316	29.193	8.027	12.932	0.146	0.374
MRBD	**4.762**	**30.592**	**8.232**	**13.257**	**0.136**	**0.357**

PersonaChat	Dist-1	Dist-2	Ent-1	Ent-2	Dis-1	Dis-2
Seq2Seq+Att	0.854	5.122	7.136	11.294	0.601	1.138
KD	0.862	5.343	7.159	11.718	0.502	0.964
CT	1.093	7.161	7.233	12.038	0.641	1.246
DML	0.952	6.399	7.196	11.824	0.435	0.891
MRBD	**1.745**	**12.391**	**7.419**	**12.246**	**0.300**	**0.578**

Table 1: Results of the automatic evaluation.

Experimental Settings

According to the performance on the validation set, including loss and metrics, we set the hyper-parameters of the proposed method and baselines as follows: We set the embedding size to 500, the vocabulary size for both DailyDialog and PersonaChat to 20K. The dropout probability and the temperature T are 0.1 and 3, respectively. We use Adam optimizer (Kingma and Ba 2015), with a learning rate of 0.0001, gradient clipping at 5.0, and a mini-batch size of 64. Following the settings of Feng et al. (2019), CT needs to pretrain students on the whole training set before Co-Teaching. We set the number of students to 6 for DML and MRBD. The imitation probability in MRBD is 0.5. The training set is randomly divided into six non-overlapping subsets with the same number of pairs. For CT, DML, and MRBD, we choose the student model that achieves the best performance on the validation set for the final evaluation.

Experimental Results

It is challenging to assess the quality of the generated responses, especially in semantics (e.g., coherence and fluency). In this work, we conduct two kinds of evaluations, automatic evaluation and human evaluation. The automatic evaluation focuses on the diversity, specificity, and distribution of responses that can be well reflected by the statistics of words. The human evaluation considers the coherence, similarity, and fluency of responses. Both BLEU (Papineni et al. 2002) and EmbSim (Liu et al. 2016) are adopted to measure the similarity of the generated response with reference, but they show a poor correlation with human evaluation.

Automatic Evaluation **Dist-{1,2}** (Distinct) are widely employed to evaluate the diversity of the generated responses (Li et al. 2016a; Zhang et al. 2018b; Feng et al. 2020a), which represent the percentage (%) of unique unigrams and bigrams. We use **Ent-{1,2}**[1] (Word Entropy) and

[1] Ent $= -\frac{1}{|U|}\sum_{w\in U}\log_2 p_g(w)$, where p_g is estimated based on the training set.

DailyDialog	Coherence	Similarity	Fluency	Average
Seq2Seq+Att	3.36	2.91	3.64	3.303
KD	2.55	2.09	2.45	2.363
CT	2.45	2.18	1.82	2.150
DML	2.00	1.72	2.00	1.906
MRBD	**1.45**	**1.55**	**1.09**	**1.363**
PersonaChat	Coherence	Similarity	Fluency	Average
Seq2Seq+Att	3.09	2.63	3.26	2.993
KD	2.25	2.13	3.01	2.463
CT	2.38	1.87	2.38	2.210
DML	1.37	1.75	2.12	1.746
MRBD	**1.12**	**1.38**	**1.25**	**1.250**

Table 2: Results of the human evaluation. Lower is better.

Model	Dist-1	Dist-2	Ent-1	Ent-2	Dis-1	Dis-2
w/o Subtask	5.317	32.149	7.265	12.476	0.205	0.486
w/o Subgroup	4.692	30.385	8.101	12.889	0.138	0.404
w/o BiDistill	3.836	25.480	8.225	13.449	0.134	0.348

Table 3: Results of the ablation study.

Ratio	Dist-1	Dist-2	Ent-1	Ent-2	Dis-1	Dis-2
0%	4.762	30.592	8.232	13.257	0.136	0.357
25%	4.808	31.585	8.351	13.343	0.133	0.308
50%	5.115	31.191	7.311	12.550	0.167	0.498
100%	5.317	32.149	7.265	12.476	0.205	0.486

Table 4: Results of different ratios of subtask overlap.

Dis-{1,2}[2] (KL divergence) to measure the specificity and distribution distance of the generated responses (Csaky, Purgai, and Recski 2019). The responses with higher word entropy contain more meaningful and low-frequency words. A lower KL divergence represents a more similar response distribution. We report both unigrams and bigrams versions of word entropy and KL divergence. The results are shown in Table 1. We can see that our training framework significantly outperforms all state-of-the-art baselines in terms of diversity, specificity, and distribution distance on all datasets, especially on PersonaChat. Compared with other baselines, CT also obtains more diverse and more specific responses as MRBD but shows a dramatic decline in the distribution distance of responses. We argue that the diversified multi-view knowledge has better regularization effects than the diversified single-view knowledge for fitting the distribution of the real-world responses. In practice, the performance of CT will weaken once we do not pre-train the students before Co-Teaching. Moreover, DML gains more performance improvements than KD in comparison to Seq2Seq+Att, which also validates multi-view knowledge is beneficial for regularizing the feature learning of students. Finally, we conducted the significant test on both DailyDialog and PersonaChat, and the results demonstrate that the performance improvements of MRBD are significant (i.e., $p < 0.01$).

Human Evaluation For all datasets, we randomly extracted 200 pairs from the test sets. Then we invited three well-educated annotators to rank the responses generated by different models in terms of **coherence** (how much information in the generated response is relevant to dialogue history), **similarity** (how much information in the generated response is related to reference), and **fluency** (how likely the generated response is from human). Ties are allowed. Table 2 reports the evaluation results. We can see that MRBD achieves consistent improvements across all metrics. Especially in the coherence and fluency, MRBD shows substantive gains. CT and DML have greater advantages than KD with respect to fluency. We also calculate the spearman's rank correlation coefficient (Zar 2014) to evaluate the inter-annotator agreement. The results are 0.542 and 0.602 on DailyDialog and PersonaChat, respectively, with $p < 0.001$.

Experimental Analysis

In this section, we provide extensive analysis to validate the effectiveness of multi-view feature representation and bidirectional distillation, and further discuss why the proposed framework works better. Unless otherwise stated, the following results are based on the test set of DailyDialog.

Ablation Study We first conduct the ablation study to analyze the contributions of different mechanisms quantitatively. Then we further investigate the impact of the overlapping ratio of subtasks and the imitation probability of students on model performance.

Table 3 shows the results of MRBD w/o Subtask (i.e., students are trained on the same training set), w/o Subgroup (each student imitates all other students), w/o BiDistill (i.e., students adopt unidirectional distillation). As we can see, MRBD w/o Subtask improves the diversity of responses but yields a sharp decline in terms of specificity and distribution distance. It is because students can not provide diversified multi-view knowledge to regularize each other for common knowledge without the subtask mechanism. The generated responses are more diverse but limited in the training set, which is in line with observations in Csaky, Purgai, and Recski (2019), i.e., the diversity of responses still increases after over-fitting the training set. MRBD w/o Subgroup shows a slight decrease in all metrics compared with MRBD, indicating that the subgroup mechanism conducts a positive effect on maintaining the diversity of knowledge. The specificity and distribution distance of MRBD w/o BiDistill obtain slight improvements, but the diversity declines dramatically, which demonstrates that the unidirectional distillation causes students only to imitate the aggregated knowledge and may lose knowledge learned from the assigned subtask.

Impact of Subtask Overlap Table 4 gives the results of MRBD with the overlapping ratios of 0%, 25%, 50%, 100%. We can discover that MRBD (25%) achieves better performance than other variants. After the overlapping ratio of 25%, the performance of MRBD represents a gradual decline in specificity and distribution distance, which is consistent with the observation in MRBD w/o subtask. It suggests that allowing subtasks to overlap appropriately is beneficial for students to gain more performance improvements.

[2] Dis $= \frac{1}{|U_r|}\sum_{w\in U_r}\log_2\frac{p_r(w)}{p(w)}$, where p_r and p are estimated based on references and the generated responses, respectively.

Probability	Dist-1	Dist-2	Ent-1	Ent-2	Dis-1	Dis-2
0.2	4.610	29.913	8.032	12.926	0.162	0.400
0.5	4.762	30.592	8.232	13.257	0.136	0.357
0.8	4.841	31.219	8.138	13.013	0.136	0.364
1.0	4.692	30.385	8.101	12.889	0.138	0.404

Table 5: Results of different imitation probabilities.

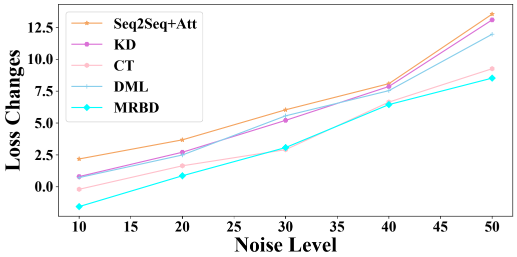

Figure 3: Robustness against noisy data.

Impact of Imitation Probability Table 5 presents the results of MRBD with the imitation probabilities of 0.2, 0.5, 0.8, and 1.0. The performance of MRBD first ascends and then slowly declines as the imitation probability gradually increases, which means that if students share too many views, the aggregated multi-view knowledge will be more similar, exacerbating the homogenization of students. Besides, MRBD consumes less computational cost compared with DML due to the adjustable imitation probability.

Model Generalization Analysis We first validate the robustness of MRBD against noisy data, and then investigate why it achieves better generalization than baselines.

Robustness against Noisy Data It is challenging to collect a large-scale and high-quality dialogue dataset. Moreover, identifying data noise in the raw dialogue dataset is labor-consuming. Knowledge distillation is beneficial for the model to resist noisy data because the student not only learns from reference but also considers the prediction from the teacher. To evaluate the robustness of models against data noise, we first add noisy data into the training set by replacing the correct responses with randomly selected responses, and then observe the changes of the test loss with respect to the proportion of noisy data. We report the results in Figure 3. Our method and CT achieve better robustness than other baselines, attributed to the independent training sets in CT and the subtask mechanism in MRBD.

Robustness against Parameter Perturbation Previous research (Chaudhari et al. 2017; Keskar et al. 2017) has proved that a wider local minimum generally represents better generalization. Specifically, with a wide local minimum, the accuracy of model prediction will not change dramatically under small perturbations in inference. To measure the width of local minima reached by baselines and our method, we add independent Gaussian noise with variable standard deviation σ to the parameters of the learned models, and then observe the changes of the test loss. Figure 4 plots the loss changes with respect to the perturbation level (i.e., the magnitude of σ). We can see that the losses of baselines change more drastically than our method after adding perturbations, which means MRBD finds a wider local minimum than baselines, indicating better generalization.

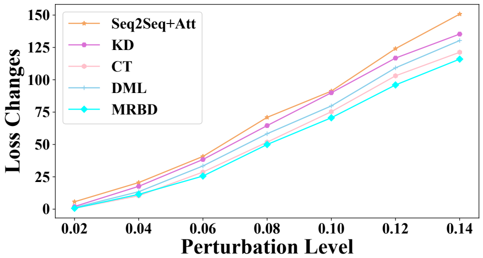

Figure 4: Robustness against parameter perturbation.

Metrics	KD	CT	DML	MRBD
Entropy	1.360	5.475	4.362	**5.765**
Diversity	0.580	0.747	0.706	**0.798**

Table 6: Entropy and diversity of predictions generated by baselines and our method.

Effect of Multi-View Feature Representation Finally, we discuss why multi-view feature representation performs better regularization on the feature learning of students.

Entropy Analysis of Prediction According to previous work (Pereyra et al. 2017; Chaudhari et al. 2017), the softened target with high entropy is beneficial for models to reach a wider local minimum. Therefore, we compare our training framework with baselines (except for Seq2Seq+Att) in terms of entropy of predictions of the teacher (or the student peers). Note that for MRBD, we choose the first three students to aggregate the predictions. The results are reported in Table 6. Our method obtains the highest entropy, which suggests that multi-view feature representation can enhance the entropy of predictions for better regularization.

Diversity Analysis of Prediction We further evaluate the diversity of predictions that reflects the degree of homogenization of students. The diversity is calculated by the average Euclidean distance between the predictions of each pair of students. As shown in Table 6, MRBD outperforms all baselines, which indicates that multi-view feature representation effectively alleviates the homogenization of students. Besides, the results also demonstrate that the diversity of predictions is related to the entropy of predictions. The predictions with more diversity can avoid the centralization of probability distribution in the aggregated prediction.

Effect of Knowledge for Regularization To validate the effectiveness of knowledge in predictions, we further introduce another kind of regularization approach, including weight decay (L_2) and label smoothing, to compare with the above four knowledge-based regularization approaches. Based on results in Table 1 and Table 7, we can find that the knowledge-based regularization has significant advantages compared with the weight decay-based regularization. The label smoothing-based regularization achieves better performance than KD and is slightly weaker than both CT and DML, but MRBD still outperforms it in all metrics. Thus

Model	Dist-1	Dist-2	Ent-1	Ent-2	Dis-1	Dis-2
w/ L_2	4.114	28.538	7.737	12.827	0.146	0.423
w/ LS	4.307	29.297	7.771	13.026	0.152	0.411

Table 7: Results of Seq2Seq+Att w/ L_2 (weight decay), w/ LS (label smoothing).

Dialogue history: I'd like to apply for a credit card. Can you help me with my application?
S2S+Att: Then what would you like for a computer? KD: Yes, which size do you recommend? CT: Sir, your account certainly is? You need a credit card. DML: Yes, it's very simple. Would you have an e-mail? MRBD: Certainly, madam. Which one would you like to apply for?
Dialogue history: What subjects are you studying?
S2S+Att: I am working in Germany and I am studying studying. KD: I am studying in a company. I am studying with you. CT: I am not interested in English. DML: I have to work abroad and work in English. MRBD: I am studying languages. I am going to work at university.

Table 8: Examples generated by baselines and MRBD.

knowledge in predictions indeed conducts a more informed guide than the conventional regularization approaches.

Case Study Table 8 shows several examples that consist of dialogue history and responses generated by different models. We can see that the responses generated by our method show more relevance with dialogue history than baselines. Although the knowledge-based baselines also generate some responses that contain words or phrases related to conversation topics, the semantics of these responses are still contradictory to dialogue history. Besides, these baselines represent low fluency, especially S2S+Att (Seq2Seq+Att).

Related Work

Previous seq2seq-based dialogue models (Vinyals and Le 2015; Shang, Lu, and Li 2015; Sordoni et al. 2015; Serban et al. 2016) tend to generate dull and meaningless responses when interacted, although they usually perform well in the training set. To tackle this problem, one line of work introduces common knowledge of the real world to constrain the feature learning of models for better generalization. Li et al. (2016a) first proposed to use mutual information maximization as the training objective instead of only predicting target sequence. Li et al. (2016b) considered the conversation task as a reinforcement learning problem and used rewards as the training objective. Some work (Li et al. 2017a; Zhang et al. 2018b; Feng et al. 2020a) further proposed a variety of manually or automatically defined rewards for more effective and comprehensive constraints. Beside, modifying the generation process with task-related inductive biases is also a promising attempt, such as syntactic-based generation (Dusek and Jurcícek 2016; Welleck et al. 2019), hierarchical generation (Serban et al. 2017a; Su et al. 2018), and latent variable based generation (Serban et al. 2017b; Zhao, Zhao, and Eskénazi 2017; Gu et al. 2019; Shen, Feng, and Zhan 2019). Several research even incorporated more specific knowledge into the dialogue task, such as topics (Xing et al. 2017), personas (Qian et al. 2018; Zhang et al. 2018a), emotions (Zhou et al. 2018), implicit scenarios (Feng et al. 2020b), and structure knowledge (Ghazvininejad et al. 2018; Young et al. 2018; Zhan et al. 2020).

Our work belongs to another line of work that aims to capture common knowledge from the training data by knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015). Tahami, Ghajar, and Shakery (2020) introduced KD into the retrieval-based dialogue model where knowledge from a better performing teacher is used to regularize a lower-performance but much faster student. (Feng et al. 2019) proposed a Co-Teaching retrieval-based dialogue model where two students are optimized on independent training sets, and imitate and guide each other with their own knowledge from the assigned training sets. However, both KD and Co-Teaching limit the teacher knowledge to single-view feature representation. As the training data is shuffled once in each epoch, each student in Co-Teaching can still access the data allocated to another student, which means two students may learn similar feature representations. Our work is more related to deep mutual learning (DML) (Zhang et al. 2018c) where a group of students with different initializations even architectures are trained on the same dataset and try to learn multi-view knowledge. Each student aggregates the teacher knowledge from all other students equally. Unfortunately, students tend to learn similar feature representations due to optimized on the same dataset, and will further homogenize based on similar teacher knowledge. These problems hinder students from learning diverse views. Meanwhile, the computation cost will increase dramatically as the number of views grows. More importantly, all of the above methods still conduct unidirectional knowledge distillation, which is not beneficial for continuous performance improvement.

The existing NLG tasks using KD mainly contain neural machine translation and text generation (Kim and Rush 2016; Tang, Lu, and Lin 2019; Wei et al. 2019; Chen et al. 2020). To the best of our knowledge, our method, including multi-view feature representation and bidirectional distillation, is the first work that applies knowledge distillation to generative dialogue systems.

Conclusion

In this work, we propose a novel training framework, multi-view feature representation with bidirectional distillation (MRBD), to guide the dialogue model towards better generalization. The students in MRBD not only learn from the assigned subtasks but also imitate diversified multi-view knowledge from the randomly selected student peers trained on different unseen subtasks. Besides, we further construct bidirectional distillation that allows the student peers to exchange knowledge simultaneously and find common parts together. Therefore, the proposed method can automatically capture common knowledge by maintaining a balance between the diversity and consistency of feature representation. The experimental results and analysis validate the superiority of the knowledge-based regularization and demonstrate the effectiveness of multi-view feature representation and bidirectional distillation.

Acknowledgements

This research is supported by Beijing Natural Science Foundation (No. L181010 and 4172054), National Key R&D Program of China (No. 2016YFB0801100), National Basic Research Program of China (No. 2013CB329605), and Beijing Academy of Artificial Intelligence (BAAI). Xu Sun and Kan Li are the corresponding authors.

References

Arora, S.; Khapra, M. M.; and Ramaswamy, H. G. 2019. On Knowledge distillation from complex networks for response prediction. In *NAACL-HLT (1)*, 3813–3822.

Ba, J.; and Caruana, R. 2014. Do Deep Nets Really Need to be Deep? In *NIPS*, 2654–2662.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

Chaudhari, P.; Choromanska, A.; Soatto, S.; LeCun, Y.; Baldassi, C.; Borgs, C.; Chayes, J. T.; Sagun, L.; and Zecchina, R. 2017. Entropy-SGD: Biasing Gradient Descent Into Wide Valleys. In *ICLR (Poster)*.

Chen, Y.; Gan, Z.; Cheng, Y.; Liu, J.; and Liu, J. 2020. Distilling Knowledge Learned in BERT for Text Generation. In *ACL*, 7893–7905.

Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*, 1724–1734.

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

Dagan, I.; Lee, L.; and Pereira, F. C. N. 1997. Similarity-Based Methods for Word Sense Disambiguation. In *ACL*, 56–63.

Dubey, R.; Agrawal, P.; Pathak, D.; Griffiths, T.; and Efros, A. A. 2018. Investigating Human Priors for Playing Video Games. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, 1348–1356.

Dusek, O.; and Jurcícek, F. 2016. Sequence-to-Sequence Generation for Spoken Dialogue via Deep Syntax Trees and Strings. In *ACL (2)*.

Feng, J.; Tao, C.; Wu, W.; Feng, Y.; Zhao, D.; and Yan, R. 2019. Learning a Matching Model with Co-teaching for Multi-turn Response Selection in Retrieval-based Dialogue Systems. In *ACL (1)*, 3805–3815.

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

Feng, S.; Ren, X.; Chen, H.; Sun, B.; Li, K.; and Sun, X. 2020b. Regularizing Dialogue Generation by Imitating Implicit Scenarios. In *EMNLP (1)*, 6592–6604.

Furlanello, T.; Lipton, Z. C.; Tschannen, M.; Itti, L.; and Anandkumar, A. 2018. Born-Again Neural Networks. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, 1602–1611.

Ghazvininejad, M.; Brockett, C.; Chang, M.; Dolan, B.; Gao, J.; Yih, W.; and Galley, M. 2018. A Knowledge-Grounded Neural Conversation Model. In *AAAI*, 5110–5117.

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

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

Hou, S.; Liu, X.; and Wang, Z. 2017. DualNet: Learn Complementary Features for Image Recognition. In *ICCV*, 502–510.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. T. P. 2017. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *ICLR*.

Kim, K.; Ji, B.; Yoon, D.; and Hwang, S. 2020. Self-Knowledge Distillation: A Simple Way for Better Generalization. *CoRR* abs/2006.12000.

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

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics* 22(1): 79–86.

Kuncheva, L. I.; and Whitaker, C. J. 2003. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Mach. Learn.* 51(2): 181–207.

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, J.; Monroe, W.; Shi, T.; Jean, S.; Ritter, A.; and Jurafsky, D. 2017a. Adversarial Learning for Neural Dialogue Generation. In *EMNLP*, 2157–2169.

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

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*.

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 *EMNLP*, 2122–2132.

Mescheder, L. M.; Nowozin, S.; and Geiger, A. 2017. The Numerics of GANs. In *NIPS*, 1825–1835.

Morcos, A. S.; Raghu, M.; and Bengio, S. 2018. Insights on representational similarity in neural networks with canonical correlation. In *NeurIPS*, 5732–5741.

Mou, L.; Song, Y.; Yan, R.; Li, G.; Zhang, L.; and Jin, Z. 2016. Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text Conversation. In *COLING*, 3349–3358.

Nagarajan, V.; and Kolter, J. Z. 2017. Gradient descent GAN optimization is locally stable. In *NIPS*, 5585–5595.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*, 311–318.

Pereyra, G.; Tucker, G.; Chorowski, J.; Kaiser, L.; and Hinton, G. E. 2017. Regularizing Neural Networks by Penalizing Confident Output Distributions. In *ICLR (Workshop)*.

Qian, Q.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Assigning Personality/Profile to a Chatting Machine for Coherent Conversation Generation. In *IJCAI*, 4279–4285.

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

Schwenker, F. 2013. Ensemble Methods: Foundations and Algorithms [Book Review]. *IEEE Comput. Intell. Mag.* 8(1): 77–79.

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.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A. C.; and Pineau, J. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*, 3776–3784.

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.; Feng, Y.; and Zhan, H. 2019. Modeling Semantic Relationship in Multi-turn Conversations with Hierarchical Latent Variables. In *ACL (1)*, 5497–5502.

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.

Su, S.; Lo, K.; Yeh, Y. T.; and Chen, Y. 2018. Natural Language Generation by Hierarchical Decoding with Linguistic Patterns. In *NAACL-HLT (2)*, 61–66.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*, 3104–3112.

Tahami, A. V.; Ghajar, K.; and Shakery, A. 2020. Distilling Knowledge for Fast Retrieval-based Chat-bots. In *SIGIR*, 2081–2084.

Tang, R.; Lu, Y.; and Lin, J. 2019. Natural Language Generation for Effective Knowledge Distillation. In *DeepLo@EMNLP-IJCNLP*, 202–208.

Vinyals, O.; and Le, Q. V. 2015. A Neural Conversational Model. *CoRR* abs/1506.05869.

Wei, H.; Huang, S.; Wang, R.; Dai, X.; and Chen, J. 2019. Online Distilling from Checkpoints for Neural Machine Translation. In *NAACL-HLT (1)*, 1932–1941.

Welleck, S.; Brantley, K.; III, H. D.; and Cho, K. 2019. Non-Monotonic Sequential Text Generation. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 6716–6726.

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.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In *CVPR*, 7130–7138.

Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018. Augmenting End-to-End Dialogue Systems With Commonsense Knowledge. In *AAAI*, 4970–4977.

Zar, J. H. 2014. Spearman rank correlation: overview. *Wiley StatsRef: Statistics Reference Online* .

Zhan, H.; Zhang, H.; Chen, H.; Shen, L.; Lan, Y.; Ding, Z.; and Yin, D. 2020. User-Inspired Posterior Network for Recommendation Reason Generation. In *SIGIR*, 1937–1940.

Zhang, S.; Dinan, E.; Urbanek, J.; Szlam, A.; Kiela, D.; and Weston, J. 2018a. Personalizing Dialogue Agents: I have a dog, do you have pets too? In *ACL (1)*, 2204–2213.

Zhang, Y.; Galley, M.; Gao, J.; Gan, Z.; Li, X.; Brockett, C.; and Dolan, B. 2018b. Generating Informative and Diverse Conversational Responses via Adversarial Information Maximization. In *NeurIPS*, 1815–1825.

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*, 730–739.