Learning to Imagine: Distillation-Based Interactive Context Exploitation for Dialogue State Tracking

Jinyu Guo[1,2], Kai Shuang[1,2]*, Kaihang Zhang[1,2], Yixuan Liu[1,2], Jijie Li[3], Zihan Wang[4]

[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications
[2]School of Computer Science, Beijing University of Posts and Telecommunications
[3]Beijing Academy of Artificial Intelligence, Beijing, China
[4]Graduate School of Information Science and Technology, The University of Tokyo
{guojinyu, shuangk, zkh1999, liuyixuan}@bupt.edu.cn, jjli@baai.ac.cn, zwang@tkl.iis.u-tokyo.ac.jp

Abstract

In dialogue state tracking (DST), the exploitation of dialogue history is a crucial research direction, and the existing DST models can be divided into two categories: full-history models and partial-history models. Since the "select first, use later" mechanism explicitly filters the distracting information being passed to the downstream state prediction, the partial-history models have recently achieved a performance advantage over the full-history models. However, besides the redundant information, some critical dialogue context information was inevitably filtered out by the partial-history models simultaneously. To reconcile the contextual consideration with avoiding the introduction of redundant information, we propose DICE-DST, a model-agnostic module widely applicable to the partial-history DST models, which aims to strengthen the ability of context exploitation for the encoder of each DST model. Specifically, we first construct a teacher encoder and devise two contextual reasoning tasks to train it to acquire extensive dialogue contextual knowledge. Then we transfer the contextual knowledge from the teacher encoder to the student encoder via a novel turn-level attention-alignment distillation. Experimental results show that our approach extensively improves the performance of partial-history DST models and thereby achieves new state-of-the-art performance on multiple mainstream datasets while keeping high efficiency.

Introduction

Imagination is the eye of the soul.

Joseph Joubert (1754 AD - 1824 AD)

Task-Oriented Dialogue (TOD) systems have achieved substantial progress and have penetrated our daily lives much more than before. As an essential component of dialogue management, Dialogue State Tracking (DST) is in charge of utilizing multi-turn dialogues to extract the compact dialogue information which contains user goals and intentions as the dialogue state. In each turn, the dialogue state is typically in the form of a set of *(slot, value)* pairs. For example, in Fig. 1, the dialogue state at turn 1 is *("restaurant-food":"Turkish")*. The continuously updated dialogue state indicates the progress of the dialogue and is leveraged to determine the next system action.

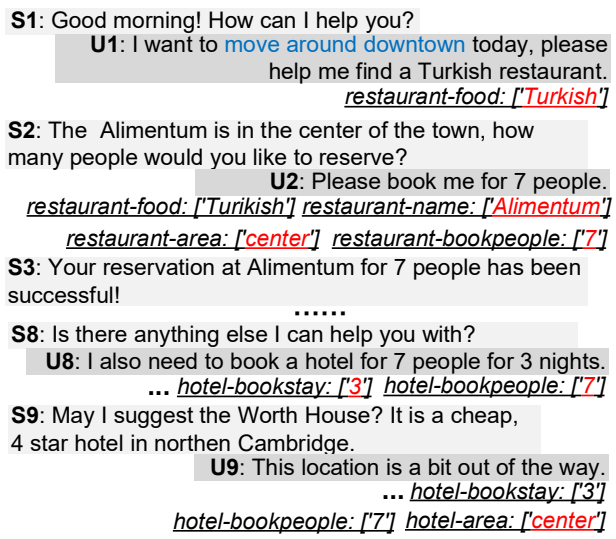

Figure 1: An example of DST. Utterances on the left and the right sides are from system and user, respectively. Each red slot value in the figure indicates that it is updated in its turn.

In dialogue state tracking, the utilization of dialogue history is a crucial research direction. The existing DST methods can generally be divided into two categories based on the use of dialogue history: full-history methods and partial-history methods. Full-history methods employ the current turn dialogue concatenated to the entire historical utterances as input to ensure the integrity of the input information (Xu and Hu 2018; Lei et al. 2018; Goel, Paul, and Hakkani-Tür 2019; Ren 2020; Shan et al. 2020; Rastogi et al. 2020; Hosseini-Asl et al. 2020). Nevertheless, this type of method usually encounters two issues: 1) The huge spatial costs and the serious efficiency issues brought by the input of all dialogues, and more seriously; 2) since the truly useful dialogues for the state update of each turn are only a small portion of all dialogue utterances, the input of all dialogues leads to the introduction of a large amount of re-

*Corresponding author.

dundant information, which may confuse the model (Yang, Huang, and Mao 2021). On the contrary, the current turn dialogue concatenated to a portion of the dialogue history is fed into the partial-history methods through either rule-based or learning-based approaches (Chao and Lane 2019; Kim et al. 2020; Guo et al. 2021, 2022; Feng et al. 2022). Initially, such methods appeared because the computational resources cannot afford the burden of feeding the whole dialogue. Subsequent models of this type aim to use only the part of the dialogue that is most relevant to the state tracking at each turn. This "select first, use later" mechanism explicitly minimizes the distracting information passed to the downstream state prediction, which achieves superior performance.

However, for the existing partial-history methods, besides the redundant information, some critical dialogue context information was inevitably filtered out simultaneously. As shown in Fig. 1, in the last turn of the dialogue (i.e., turn 9), the user thinks the location of the hotel is out of the way, but there is no explicit expression of the user's intention to find a hotel in the center of the city. We can observe that the texts explicitly indicating this requirement exist in turn 1. In light of this, if turn 1 is not input to a partial-history DST model, the mindless model can only process the contents seen without any imagination, while a human can easily imagine based on experience and thereby understand the user's requirement. Therefore, if it is possible to enable the model to imagine during encoding the partial dialogues, the obtained dialogue representations for the downstream state generation will be an essential complement to the dialogue context while avoiding introducing additional dialogues.

To achieve this goal, we propose DICE-DST, a model-agnostic module widely applicable to the partial-history DST models. It aims to optimize the encoder of each DST model from the perspective of dialogue context supplementation without introducing additional dialogues. Specifically, we first construct a teacher encoder and devise two contextual reasoning tasks to train it to acquire extensive dialogue contextual knowledge. Then we transfer the contextual knowledge from the teacher encoder to the student encoder via a novel turn-level attention-alignment distillation. The training of the student encoder will be supervised by both the distillation and the objective of dialogue state tracking in each DST model. In this training process, we input the full history of each dialogue session to the teacher encoder, while we input the dialogue turns used and unused to the student encoder separately according to each model's settings. Then we apply the attention-alignment mechanism to complement the missing attention values to transfer the crucial dialogue contextual knowledge. In the inference stage, the trained student encoder will be capable of supplementing the contextual knowledge related to the input dialogues and thereby facilitating the downstream dialogue state tracking. To the best of our knowledge, our proposed DICE-DST is the first work to reconcile the contextual consideration with avoiding the introduction of redundant information.

We extensively evaluate our proposed method[1], and experimental results show that our DICE-DST improves the performance of multiple partial-history DST models and therefore achieves new state-of-the-art performance on the mainstream benchmarks: MultiWOZ 2.1 (Eric et al. 2020) and MultiWOZ 2.2 (Zang et al. 2020). In addition, DICE-DST also achieves new state-of-the-art performance on Sim-M and Sim-R (Shah et al. 2018) and competitive performance on DSTC2 (Henderson, Thomson, and Williams 2014). Notably, since the full dialogue history is not fed into the target encoder during the inference stage, our method could show superiority in terms of performance while keeping high efficiency at the same time.

In summary, our contributions are mainly three-fold:

- We propose DICE-DST which is widely applicable to the existing partial-history DST methods, which enables the encoder to strengthen the context of dialogue input without introducing additional dialogues.
- We devise two contextual reasoning tasks to train the teacher encoder to acquire extensive dialogue contextual knowledge. We also propose a novel turn-level attention alignment mechanism to interactively bridge the gap between the teacher encoder and the student encoder.
- Experiments show that our approach widely improves the performance of partial-history DST models and achieves new state-of-the-art performance on multiple mainstream datasets while maintaining high efficiency.

Related Work

Traditional DST models usually determine dialogue states by considering only utterances at the current turn (Mrkšić et al. 2017; Zhu et al. 2020; Lee, Lee, and Kim 2019). With the prevalence of pre-trained language models (PrLMs) (Kenton and Toutanova 2019; Radford et al. 2019; Lan et al. 2019), some DST models employ the current turn dialogue concatenated to the whole historical utterances as input to ensure the integrity of the input information (Xu and Hu 2018; Lei et al. 2018; Goel, Paul, and Hakkani-Tür 2019; Ren 2020; Shan et al. 2020; Rastogi et al. 2020; Hosseini-Asl et al. 2020). Recently, granularity in DST has been proposed to quantify the utilization of dialogue history (Yang, Huang, and Mao 2021). Its experimental results demonstrate that redundant content can become distracting information to pose a hindrance. The contrasting partial-history models recently applied the mechanism of pre-use selection to explicitly minimizes the distracting information passed to the downstream state prediction (Chao and Lane 2019; Kim et al. 2020; Guo et al. 2021, 2022; Feng et al. 2022). Especially, (Guo et al. 2022) dynamically selects the relevant dialogue contents corresponding to each slot for state updating, and thereby achieves superior performance. Nevertheless, all partial-history models will inevitably filter out some critical dialogue context information simultaneously.

On the other hand, knowledge distillation (Hinton et al. 2015; Tang et al. 2019) aims to transfer knowledge from one model to another. Recently, a great variety of knowledge distillation approaches have been developed on top of the pre-trained language models (Sun et al. 2019; Sanh et al. 2019; Sun et al. 2020; Jiao et al. 2020; Sun et al. 2023). Some approaches compress BERT to a tiny structure with

[1]Code is available at https://github.com/guojinyu88/DICE_DST

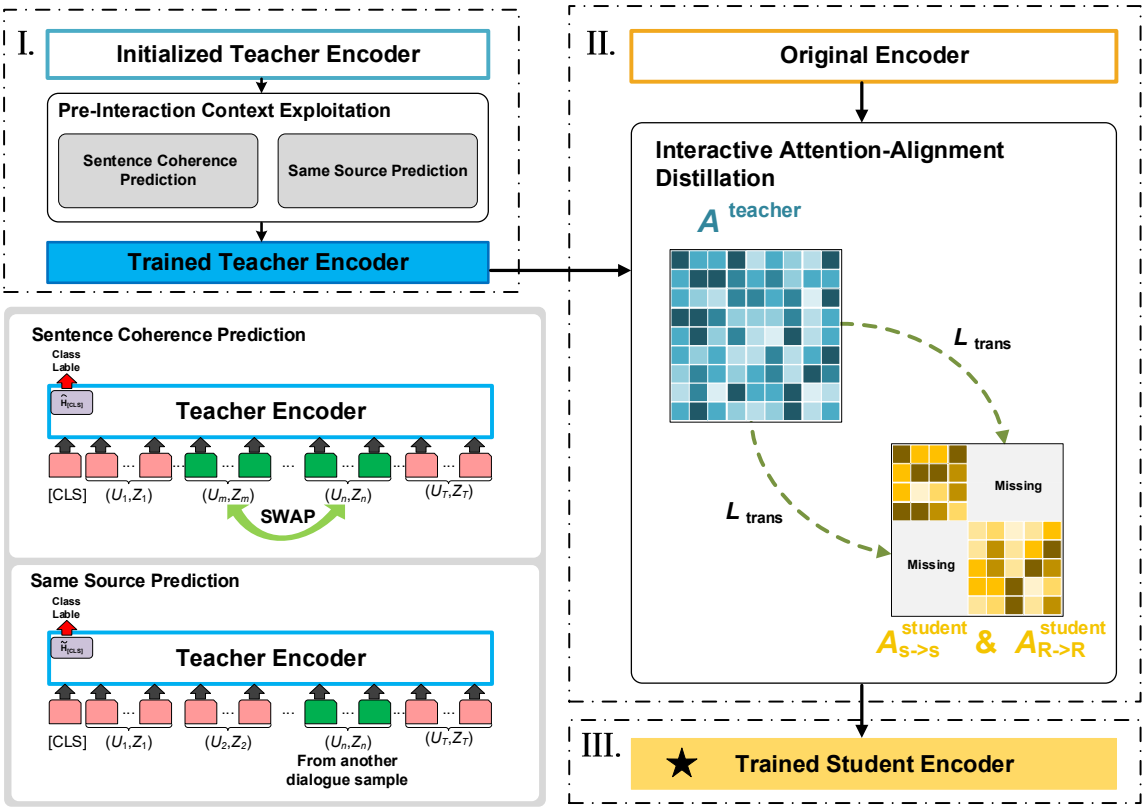

Figure 2: The architecture of the proposed DICE-DST. It consists of three stages, which are marked with dash lines respectively. Stage I is the training process of the teacher encoder. Stage II is the training of the student encoder, which is supervised by both the distillation and the objective of dialogue state tracking. Stage III is the inference stage.

fewer transformer layers and smaller hidden sizes by distilling the output of the teacher model. In addition to distilling the predicted logits and hidden states of the teacher model, the relationships between these outputs can also be captured as knowledge, which enables the student model to have the potential to surpass its teacher model. (Li et al. 2021) asks representation-based encoders to conduct virtual interactions that mimic the behaviors as interaction-based models do, and it takes the knowledge distilled from interaction-based encoders as supervised signals to promise the effectiveness of virtual interactions. (Hu et al. 2018) investigates knowledge distillation in the context of machine reading comprehension. The mechanism of attention alignment in our work is partially inspired by the distillation of such attentive information.

Approach

In this section, we first introduce the encoder of a DST model. Then we formally introduce DICE-DST, a model-agnostic module that aims to optimize the encoder of each partial-history DST model from the perspective of dialogue context supplementation. As illustrated in Fig. 2, our approach is mainly composed of two modules: a Pre-Interaction Context Exploitation (PICE) module and an Interactive Attention-Alignment Distillation (IAAD) module to conduct a two-stage process. Where PICE first constructs a teacher encoder and trains it with two contextual reasoning tasks to acquire extensive dialogue contextual knowledge, then IAAD transfers the knowledge from the teacher encoder to the student encoder via a turn-level attention-alignment distillation. The training of the student encoder will be supervised by both the distillation and the objective of dialogue state tracking in each DST model. Details of each module are given respectively in the remainder of this section.

Problem Formulation and Encoder in DST

Given a dialogue $X = \{(U_1; Z_1), (U_2; Z_2), \ldots, (U_T; Z_T)\}$ of T turns where U_t represents user utterance and Z_t represents system response of turn T, DST is tasked to extract dialogue states at each turn $t(t \leq T)$, which is defined as B_t. In general, the issue of DST consists of two objectives: 1) learning a dialogue utterance encoder $\mathcal{F}_E : D_t \rightarrow E_t$ that takes the dialogue utterances D_t as input and obtains the representation for the concatenated dialogue sequence E_t; 2) learning a $\mathcal{F}_B : E_t \rightarrow B_t$ that takes the representation E_t

as input and predicts the dialogue state B_t at each turn t. In this paper, DICE-DST aims to optimize the dialogue utterance encoder $\mathcal{F}_E$ from the perspective of dialogue context supplementation, and thereby facilitate the downstream dialogue state prediction. For the sake of simplicity, some slots that would be also input to the encoder in some DST models are omitted in the next sections.

Let $C = \{1, 2, \ldots, t\}$ denote the set of turns of the entire dialogue history up to turn t. As aforementioned, our approach applies to the partial-history DST models. We use $S = \{s_1, s_2, \ldots, s_k\}$ to denote the set of turns of dialogues that are input to the encoder of a partial-history DST model (i.e., corresponds to $D_t = \{(U_{s_1}; Z_{s_1}), (U_{s_2}; Z_{s_2}), \ldots, (U_{s_k}; Z_{s_k})\}$), and its size as $k = |S|$. Therefore, the set of turns that are not input to the encoder is denoted as $R = \complement_C S = \{r_1, r_2, \ldots, r_l\}$, and its size as $l = t - k$.

Pre-Interaction Context Exploitation

In this stage, we first construct a teacher encoder, and then we propose two contextual reasoning tasks as follows to train it to acquire extensive dialogue contextual knowledge:

Sentence Coherence Prediction Sentence coherence is a basic aspect of contextualization. Here we employ a task to train the teacher encoder's ability to determine whether all dialogue turns in one dialogue session are in the correct order. Specifically, we take all dialogue sessions from the corresponding dataset as samples. For half of these samples, we swap the position of two turns in each dialogue session. While for the others, we maintain the original sequence. After this operation, we enter each dialogue session into the teacher encoder, and then we feed the [CLS] token's hidden state to an MLP as a classifier to make a two-way prediction:

$$\breve{H} = \text{Teacher}(\breve{X}), \breve{H} \in \mathbb{R}^{len(\breve{X}) \times d} \tag{1}$$

$$P^{sc} = \text{softmax}(\text{MLP}(\breve{H}_{[\text{CLS}]})) \tag{2}$$

Same Source Prediction In addition to identifying the order of contents in one dialogue session, we also propose a task to enable the encoder to determine whether all dialogue turns belong to the same dialogue session. Since only a portion of each dialogue session is fed into the partial-history DST model during the inference phase, we believe this task will facilitate the student encoder after the knowledge has been transferred to imagine in the correct direction in the inference stage. Specifically, for half of the dialogue samples, one turn of each dialogue session will be replaced with a turn belonging to another dialogue session, while in the other half of dialogue samples, the original sequence is maintained. Similarly, we feed each dialogue session into the teacher encoder, and a binary classifier follows [CLS] token to predict:

$$\widetilde{H} = \text{Teacher}(\widetilde{X}), \widetilde{H} \in \mathbb{R}^{len(\widetilde{X}) \times d} \tag{3}$$

$$P^{ss} = \text{softmax}(\text{MLP}(\widetilde{H}_{[\text{CLS}]})) \tag{4}$$

The loss functions for both two tasks above are cross-entropy loss as follows:

$$\mathcal{L}^{sc} = \sum_{i \in [0,1]} -y_i^{sc} log P_i^{sc} \tag{5}$$

$$\mathcal{L}^{ss} = \sum_{i \in [0,1]} -y_i^{ss} log P_i^{ss} \tag{6}$$

The training goal is to minimize $\mathcal{L}_{teacher} = \mathcal{L}^{sc} + \mathcal{L}^{ss}$ so that the teacher encoder can be adapted to both two tasks.

Interactive Attention-Alignment Distillation

In this module, we take the original encoder in each partial-history DST model as the student encoder and transfer the contextual knowledge from the teacher encoder to the student encoder via a proposed turn-level attention-alignment distillation. The distillation process is performed in parallel with the training of the model on the DST dataset. Specifically, we input the full history of each dialogue session to the teacher encoder, while we input the dialogue turns used and unused to the student encoder separately according to each model's settings. In this case, the different inputs lead to a gap between these two encoders in their contextual representations. To bridge this gap, we first make a detailed analysis of it through the mechanism of pre-trained language models (PrLMs). A PrLM utilizes the multi-head self-attention (MHA) mechanism to produce each element's representation by a weighted average of the rest of the elements, and the attention score of each head between two elements can be calculated as $A_{i \to j} = \text{MHA}(i, j) = \frac{K_i Q_j^\top}{\sqrt{d}}$. In the inference stage, the teacher encoder can compute contextual representation based on the full dialogue history, and its attention score can be represented as follows:

$$A^{\text{teacher}} = \begin{bmatrix} A_{1 \to 1} & \cdots & A_{1 \to t} \\ \vdots & \ddots & \vdots \\ A_{t \to 1} & \cdots & A_{t \to t} \end{bmatrix} \tag{7}$$

where t is the number of the current turn, $A_{i \to j}$ denotes the attention map generated by the representation of turn i dialogue attending to the representation of turn j dialogue.

On the other hand, the student encoder can only focus on the interactions between the turns of the input partial history in the inference stage, which can be represented as:

$$A^{\text{student}} = \begin{bmatrix} A_{s_1 \to s_1} & \cdots & A_{s_1 \to s_k} \\ \vdots & \ddots & \vdots \\ A_{s_k \to s_1} & \cdots & A_{s_k \to s_k} \end{bmatrix} \tag{8}$$

We observe that the interaction between the input dialogue turns and the unused dialogue turns $A_{S \to R}, A_{R \to S}$, as well as the interaction between the unused dialogue turns $A_{R \to R}$ are missing in the student encoder. We consider that these gaps contain crucial dialogue contextual knowledge. Based on the above analysis, we first enable the student encoder to mimic the interaction, and then employ the teacher encoder to supervise it by the knowledge distillation to complete the missing interactions. Specifically, the student encoder first encodes the set of unused dialogue turns and the set of input dialogue turns as follows:

$$\mathcal{R} = \text{Student}\left(X_{r_1} \oplus \ldots \oplus X_{r_l}\right) \tag{9}$$

$$\mathcal{S} = \text{Student}\left(X_{s_1} \oplus \ldots \oplus X_{s_k}\right) \tag{10}$$

Then the missing $A_{S\to R}, A_{R\to S}$ and $A_{R\to R}$ mentioned above can be calculated as follows:

$$A^{\text{student}}_{S_i\to R_j} = \frac{(w_k\mathcal{S}_i)(w_q\mathcal{R}_j)^\top}{\sqrt{d}} \tag{11}$$

$$A^{\text{student}}_{R_j\to S_i} = \frac{(w_k\mathcal{R}_j)(w_q\mathcal{S}_i)^\top}{\sqrt{d}} \tag{12}$$

$$A^{\text{student}}_{R_n\to R_m} = \frac{(w_k\mathcal{R}_n)(w_q\mathcal{R}_m)^\top}{\sqrt{d}} \tag{13}$$

where w_k and w_q are the parameters of the transformer encoder at multi-head attention. $\mathcal{R}_j, \mathcal{R}_n, \mathcal{R}_m$ is extracted from $\mathcal{R}$, and $\mathcal{S}_i$ is extracted from $\mathcal{S}$. As aforementioned, the existing attention score in the teacher encoder will guide the student encoder to complete the missing attention values, which is exactly how the teacher encoder transfers the contextual knowledge to the student encoder. The goal is to minimize the MSE loss across all attention scores to be completed:

$$\mathcal{L}_{\text{trans}} = \sum \left(A^{\text{teacher}} - A^{\text{student}}\right)^2 \tag{14}$$

The student encoder will be trained to minimize $\mathcal{L} = \mathcal{L}_{\text{trans}} + \mathcal{L}_{\text{DST}}$, where $\mathcal{L}_{\text{DST}}$ are different in various methods. In the inference stage, the student encoder only takes partial history and will supplement the knowledge related to the input without introducing additional dialogues.

In addition, our approach can be easily extended to the RNN-based model. In PICE, when an RNN-based encoder acts as a student, we replace the teacher encoder with an RNN-based one as well, and we retain the same contextual reasoning tasks. In IAAD, we feed the corpus into an RNN-based encoder to get the hidden state and use the hidden state as the basis for calculating the attention map as follows:

$$A = \text{softmax}(\frac{H_R H_S^\top}{\sqrt{d}}) \tag{15}$$

where $H_R, H_S^\top$ is the hidden state of the unused dialogues and the input dialogues respectively.

Experiments

Datasets and Evaluation Measures

Our proposed method is evaluated in most of the mainstream benchmark task-oriented dialogue challenges: MultiWOZ 2.2, MultiWOZ 2.1, Sim-R, Sim-M, and DSTC2. MultiWOZ 2.2 and MultiWOZ 2.1 are two versions of the most popular task-oriented dialogue dataset nowadays. It is a fully-labeled collection of human-human written dialogues spanning multiple domains. Compared to MultiWOZ 2.1, MultiWOZ 2.2 is re-annotated with a different set of annotators and also canonicalized entity names. Sim-M and Sim-R contain human-paraphrased simulated dialogues in the movie and restaurant domains. The prevalence of out-of-vocabulary (OOV) values exists in their slots. DSTC2 is in the restaurant domain.

We exploit the widely adopted Joint Goal Accuracy (JGA) (Wu et al. 2019) on all test sets. JGA refers to the accuracy of the dialogue state in each turn, which is defined as the ratio of dialogue turns for which all slot values have been filled correctly according to the ground truth.

Experimental Settings

We employ ALBERT-large model (Lan et al. 2019) as the backbone of the teacher encoder and student encoder. For the teacher encoder, dialogues with more than 512 tokens will be truncated by retaining as many turns as possible. The truncated dialogues with more than three turns will be considered a separate dialogue session to generate training data. We group 32 samples as a batch to jointly train the teacher encoder. For the student encoder, we employ a group size of 8 to batch process dialogue turns. During training, we use ground-truth selected slots instead of the predicted ones. We set the maximum number of turns as 16 and the maximum length of tokens as 512. We use AdamW optimizer (Loshchilov and Hutter 2017) with $\beta\_1 = 0.9$, $\beta\_2 = 0.999$, $\epsilon = 1e-8$ and set the warmup proportion to 0.1. We set the learning rate of the pre-trained language model parameters to $2e-5$ and the learning rate of the other parameters to $1e-4$. We utilize dropout (Srivastava et al. 2014) with the probability of 0.1.

Baselines and Characteristics

We take into account the different characteristics of the models when selecting the partial-history DST baselines, and we will first introduce these characteristics as follows:

Update Strategy The selective-update methods first perform slot update selection, then only the slots selected are permitted to update values, while the other slots directly inherit the values from the previous turn (Kim et al. 2020; Guo et al. 2021, 2022). The equal-update methods treat all slots equally and predict the dialogue state at every turn from scratch. Typically, a selective-update method requires an additional encoder for slot update selection.

Encoder This refers to the type of backbone of the model's encoder. Despite the current dominance of BERT-based encoders, we still apply DICE-DST to the classical RNN-based generative DST models to widely evaluate the validity of DICE-DST.

Number of Dialogue Turns Involved This refers to the number of dialogue turns input to the encoder.

Dialogue Continuity This refers to whether the dialogue turns input to the encoder is continuous.

Based on these characteristics, we select the following baselines, and Table 1 shows their detailed information.

BERT-DST (Chao and Lane 2019) generates language representations suitable for scalable DST. It decodes only the slot values of the slots mentioned in the current turn of dialogue, and then uses a rule-based mechanism to update from the previous turn state to the current turn state. This is one of the early efforts of the BERT-based DST model. *TRADE* (Wu et al. 2019) encodes the whole dialogue context and decodes the value for every slot using a copy-augmented decoder. It is the first to consider knowledge transfer between domains in multi-domain DST scenarios. *SOM-DST* (Kim et al. 2020) is the first previous-based method. It takes the dialogue state as an explicit memory that can be selectively overwritten and inputs it into

Model	Update Strategy	Encoder	Dialogue Continuity	No. of Dialogue Turns Involved
BERT-DST (Chao and Lane 2019)	equal update	BERT	Yes	1
TRADE (Wu et al. 2019)†	equal update	RNN	Yes	2
SOM-DST (Kim et al. 2020)	selective update	BERT & RNN	Yes	2
DSGF-NET (Feng et al. 2022)	equal update	BERT	Yes	2
DiCoS-DST (Guo et al. 2022)‡	selective update	BERT	No	3

Table 1: Statistics on the characteristics of the 5 baselines in the paper. † means that the range of dialogue history used in TRADE can be freely set. In this paper, we use the dialogue of the current turn and previous turn as input to TRADE to evaluate the effectiveness of our approach. ‡ means that the number of dialogue turns used in DiCoS-DST can be set. In this paper, we choose the model using 3 turns of dialogue whose performance was optimal. The subsequent main experimental results on these two models are all based on these settings.

Model	MultiWOZ 2.1	MultiWOZ 2.2	Sim-M	Sim-R	DSTC2
	JGA (%)	JGA (%)	JGA (%)	JGA (%)	JGA (%)
BERT-DST (Chao and Lane 2019)	-	47.95	80.1	89.6	69.3
BERT-DST + DICE†	-	49.21 (±0.51)	81.9 (±1.2)	90.2 (±0.3)	71.5 (±0.2)
TRADE (Wu et al. 2019)	41.30	41.10	-	-	-
TRADE + DICE†	42.90 (±0.30)	42.60 (±0.30)	-	-	-
SOM-DST (Kim et al. 2020)	53.68	-	-	-	-
SOM-DST + DICE†	54.52 (±0.34)	-	-	-	-
DSGF-NET (Feng et al. 2022)	56.70	55.80	-	-	-
DSGF-NET + DICE‡	57.63 (±0.36)	56.84 (±0.42)	-	-	-
DiCoS-DST (Guo et al. 2022)	61.02	61.13	84.7	91.5	78.4
DiCoS-DST + DICE†	**61.76 (±0.29)**	**61.98 (±0.33)**	**85.3 (±0.8)**	**91.9 (±0.2)**	**79.2 (±0.2)**

Table 2: Accuracy (%) on the test sets of benchmark datasets. † means that we build our approach on the source code provided by the author of the paper. ‡ means that we build the original model ourselves and apply our approach.

PrLM	MultiWOZ 2.2
ALBERT (large)	**61.98**
ALBERT (base)	61.65 (-0.33)
BERT (large)	61.69 (-0.29)
BERT (base)	61.51 (-0.47)

Table 3: Ablation study with joint goal accuracy (%).

Method	MultiWOZ 2.2
DiCoS-DST + DICE-DST	**61.98**
-Sentence Coherence Predication	61.65 (-0.33)
-Same Source Predication	61.53 (-0.45)
20% of the samples are processed	61.37 (-0.61)
80% of the samples are processed	61.42 (-0.56)

Table 4: Ablation study with joint goal accuracy (%).

Mechanism	MultiWOZ 2.2
$A^{student}_{R_l \to S_k}$ and $A^{student}_{S_k \to R_l}$	**61.98**
$A^{student}_{R_l \to S_k}$	61.54 (-0.44)
$A^{student}_{S_k \to R_l}$	61.42 (-0.56)

Table 5: Ablation study with joint goal accuracy (%).

BERT together with the current turn dialogue. Finally, it decodes each slot value using a pointer-generator network. *DSGF-NET* (Feng et al. 2022) generates a dynamic schema graph to explicitly fuse the prior slot-domain membership relations and dialogue-aware dynamic slot relations. It also employs a schema-agnostic graph attention network to share information. *DiCoS-DST* (Guo et al. 2022) dynamically selects the relevant dialogue contents corresponding to each slot from a combination of three perspectives. Since only the selected dialogue contents are fed into the state generator, this mechanism explicitly minimizes the distracting information passed to the downstream state prediction and thereby achieves the prior state-of-the-art performance.

Main Results

Table 2 shows the results of our approach applying to the baselines. The application of our DICE-DST brings a considerable performance improvement for each baseline. In particular, DICE-DST further improves the performance of DiCoS-DST and thereby achieves new state-of-the-art performance on MultiWOZ 2.1 and MultiWOZ 2.2 with joint goal accuracy of 61.76% and 61.98%. Besides, despite the sparsity of experimental results on Sim-M and Sim-R, the combination of DiCoS-DST and DICE-DST still achieves SOTA performance on these two datasets. On DSTC2, the performance of this combination is also competitive, which is second only to that of Seq2seq-DU (Feng, Wang, and Li 2021). In general, DICE-DST improves the performance of

k	DiCoS-DST	DiCoS-DST + DICE	Difference
1	61.04	61.75	+0.71
2	61.13	61.98	+0.85
3	60.88	61.41	+0.53

Table 6: The JGA (%) of different k. k is the number of additional dialogue turns used in addition to the current turn.

	Original	+ DICE-DST	Difference
TRADE†	45.6	45.4	-0.2
TripPy	55.29	55.16	-0.13

Table 7: Accuracy (%) on the test sets of MultiWOZ 2.1. † means that we use the entire dialogue history as input.

BERT-based and RNN-based models to a similar degree. About the update strategy, the performance improvement of DICE-DST for the two selective-update models is relatively lower than that for the other equal-update models. We conjecture that this is due to the slot update signal usually exists in the current turn rather than in the dialogue history, so the effect of contextual supplementation is limited. For the effectiveness on the different number of dialogue turns involved, we discuss it in detail in Analysis section.

Ablation Study

To understand the effect of different proposed techniques, we take the combination of DiCoS-DST + DICE-DST as an example to evaluate these techniques separately.

Different PrLMs We employ different pre-trained language models as the backbone of the teacher encoder for training and testing on MultiWOZ 2.2. Table 3 shows that the JGA of other PrLMs decreases in varying degrees compared with ALBERT-large. We can observe that the teacher encoder based on each PrLMs improves the performance of DiCoS-DST by participating in distillation. This demonstrates that our mechanism can achieve consistent performance gain based on various representation foundations.

Effect of Contextual Reasoning Tasks We conduct an ablation study of the proposed two contextual reasoning tasks on MultiWOZ 2.2. As shown in Table 4, we observe that the performance degrades by 0.33% for JGA when the Sentence Coherence Prediction is removed. Likewise, removing the Same Source Prediction brings a 0.45% performance degradation, which is a little more than the drop from removing the former task. In addition, we also attempt to adjust the proportion of samples performing the operation in both tasks. As shown in rows 4 to 5 in Table 4, performing operations on 20% or 80% of the samples reduces the performance of the trained teacher encoder. We conjecture that this is due to processing 50% of the samples may better balance the positive and negative samples.

Unidirectional versus Bidirectional Interaction Given that the goal of our approach is to supplement the context of the used dialogues with the unused dialogues, is it possible that performing unidirectional interactions during attention alignment can yield better results? To test this conjecture, we separately perform two unidirectional interactive attention alignments (i.e., only $A^{student}_{R_l \to S_k}$ or $A^{student}_{S_k \to R_l}$). As reported in Table 5, the JGA of each unidirectional interactive attention alignment decreases markedly compared to the bidirectional interaction. This indicates that the bidirectional attention alignment enables more adequate interaction. In addition, the performance of $A^{student}_{R_l \to S_k}$ is better than that of $A^{student}_{S_k \to R_l}$, which is consistent with the unidirectional complement goal of our approach.

Analysis

Effectiveness on Different Number of Turns

Intuitively, DICE-DST improves more for partial-history DST models with fewer input dialogue turns. To verify this, we utilize DICE-DST to optimize DiCoS-DST on different numbers of selected dialogue turns k to compare their performance gains. As shown in Table 6, the improvement from DICE-DST is greatest when $k = 2$. We believe that this is because the less input information cannot provide enough "material" for the imagination of the module, while the excessive input narrows the imagination space and introduces more noise simultaneously.

Break the Limitation of Partial-History?

All of the previous sections explore the effect of DICE-DST on the partial-history DST methods. What is the effect of DICE-DST on the full-history DST methods? To investigate it, we take TRADE and TripPy (Heck et al. 2020) as examples and optimize them by DICE-DST to observe the performance change. As reported in Table 7, the application of DICE-DST caused a slight degradation in the performance of both models. This experimentally shows that since the full-history DST models have already input the complete dialogue history information, it is difficult to have further improvements by additional imagination, which is to some extent consistent with intuition.

Conclusion

We introduce an effective DICE-DST that is widely applicable to the partial-history DST models. It aims to optimize the encoder of each DST model from the perspective of dialogue context supplementation without introducing additional dialogues. It first constructs a teacher encoder and trains it with two contextual reasoning tasks to acquire extensive dialogue contextual knowledge, then it transfers the contextual knowledge from the teacher encoder to the student encoder via a novel turn-level attention-alignment distillation. Experimental results show that DICE-DST widely improves the performance of partial-history DST models and achieves new SOTA performance on multiple mainstream datasets while maintaining high efficiency. We believe the combination of the "select first, use later" mechanism and the distillation opens a door to a promising area of long text, and we will explore it for more than DST in the future.

Acknowledgements

This work was supported by Beijing Natural Science Foundation (Grant No. 4222032) and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (Grant No. 61921003). This work was also supported by BUPT Excellent Ph.D. Students Foundation (Grant no. CX2022225) and the China National Key R&D Program (Grant No.2020AAA0105200).

References

Chao, G.-L.; and Lane, I. 2019. BERT-DST: Scalable End-to-End Dialogue State Tracking with Bidirectional Encoder Representations from Transformer. *Proc. Interspeech 2019*, 1468–1472.

Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; Kumar, A.; Goyal, A.; Ku, P.; and Hakkani-Tur, D. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 422–428.

Feng, Y.; Lipani, A.; Ye, F.; Zhang, Q.; and Yilmaz, E. 2022. Dynamic Schema Graph Fusion Network for Multi-Domain Dialogue State Tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 115–126.

Feng, Y.; Wang, Y.; and Li, H. 2021. A Sequence-to-Sequence Approach to Dialogue State Tracking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1714–1725.

Goel, R.; Paul, S.; and Hakkani-Tür, D. 2019. Hyst: A hybrid approach for flexible and accurate dialogue state tracking. *arXiv preprint arXiv:1907.00883*.

Guo, J.; Shuang, K.; Li, J.; and Wang, Z. 2021. Dual Slot Selector via Local Reliability Verification for Dialogue State Tracking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 139–151.

Guo, J.; Shuang, K.; Li, J.; Wang, Z.; and Liu, Y. 2022. Beyond the Granularity: Multi-Perspective Dialogue Collaborative Selection for Dialogue State Tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2320–2332.

Heck, M.; van Niekerk, C.; Lubis, N.; Geishauser, C.; Lin, H.-C.; Moresi, M.; and Gasic, M. 2020. TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 35–44.

Henderson, M.; Thomson, B.; and Williams, J. D. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, 263–272.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hosseini-Asl, E.; McCann, B.; Wu, C.-S.; Yavuz, S.; and Socher, R. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33: 20179–20191.

Hu, M.; Peng, Y.; Wei, F.; Huang, Z.; Li, D.; Yang, N.; and Zhou, M. 2018. Attention-Guided Answer Distillation for Machine Reading Comprehension. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2077–2086.

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 the Association for Computational Linguistics: EMNLP 2020*, 4163–4174.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Kim, S.; Yang, S.; Kim, G.; and Lee, S.-W. 2020. Efficient Dialogue State Tracking by Selectively Overwriting Memory. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 567–582.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Lee, H.; Lee, J.; and Kim, T.-Y. 2019. SUMBT: Slot-Utterance Matching for Universal and Scalable Belief Tracking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5478–5483.

Lei, W.; Jin, X.; Kan, M.-Y.; Ren, Z.; He, X.; and Yin, D. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1437–1447.

Li, D.; Yang, Y.; Tang, H.; Wang, J.; Xu, T.; Wu, W.; and Chen, E. 2021. VIRT: Improving Representation-based Models for Text Matching through Virtual Interaction. *arXiv preprint arXiv:2112.04195*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Mrkšić, N.; Séaghdha, D. Ó.; Wen, T.-H.; Thomson, B.; and Young, S. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1777–1788.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rastogi, A.; Zang, X.; Sunkara, S.; Gupta, R.; and Khaitan, P. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8689–8696.

Ren, L. 2020. *Scalable and accurate dialogue state tracking via hierarchical sequence generation*. University of California, San Diego.

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

Shah, P.; Hakkani-Tür, D.; Tür, G.; Rastogi, A.; Bapna, A.; Nayak, N.; and Heck, L. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Shan, Y.; Li, Z.; Zhang, J.; Meng, F.; Feng, Y.; Niu, C.; and Zhou, J. 2020. A contextual hierarchical attention network with adaptive objective for dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6322–6333.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Sun, L.; Ye, J.; Peng, H.; Wang, F.; and Yu, P. S. 2023. Self-Supervised Continual Graph Learning in Adaptive Riemannian Spaces. In *Proceedings of AAAI*.

Sun, S.; Cheng, Y.; Gan, Z.; and Liu, J. 2019. Patient Knowledge Distillation for BERT Model Compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4323–4332.

Sun, Z.; Yu, H.; Song, X.; Liu, R.; Yang, Y.; and Zhou, D. 2020. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2158–2170.

Tang, R.; Lu, Y.; Liu, L.; Mou, L.; Vechtomova, O.; and Lin, J. 2019. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*.

Wu, C.-S.; Madotto, A.; Hosseini-Asl, E.; Xiong, C.; Socher, R.; and Fung, P. 2019. Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 808–819.

Xu, P.; and Hu, Q. 2018. An End-to-end Approach for Handling Unknown Slot Values in Dialogue State Tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1448–1457.

Yang, P.; Huang, H.-Y.; and Mao, X.-L. 2021. Comprehensive Study: How the Context Information of Different Granularity Affects Dialogue State Tracking? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2481–2491.

Zang, X.; Rastogi, A.; Sunkara, S.; Gupta, R.; Zhang, J.; and Chen, J. 2020. MultiWOZ 2.2: A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 109–117.

Zhu, S.; Li, J.; Chen, L.; and Yu, K. 2020. Efficient Context and Schema Fusion Networks for Multi-Domain Dialogue State Tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 766–781.