Beyond the Granularity: Multi-Perspective Dialogue Collaborative Selection for Dialogue State Tracking

Jinyu Guo[1], **Kai Shuang**[1]*, **Jijie Li**[1], **Zihan Wang**[2] and **Yixuan Liu**[1]

[1]State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
[2]Graduate School of Information Science and Technology, The University of Tokyo
{guojinyu, shuangk, lijijie, liuyixuan}@bupt.edu.cn
zwang@tkl.iis.u-tokyo.ac.jp

Abstract

In dialogue state tracking, dialogue history is a crucial material, and its utilization varies between different models. However, no matter how the dialogue history is used, each existing model uses its own consistent dialogue history during the entire state tracking process, regardless of which slot is updated. Apparently, it requires different dialogue history to update different slots in different turns. Therefore, using consistent dialogue contents may lead to insufficient or redundant information for different slots, which affects the overall performance. To address this problem, we devise DiCoS-DST to dynamically select the relevant dialogue contents corresponding to each slot for state updating. Specifically, it first retrieves turn-level utterances of dialogue history and evaluates their relevance to the slot from a combination of three perspectives: (1) its explicit connection to the slot name; (2) its relevance to the current turn dialogue; (3) Implicit Mention Oriented Reasoning. Then these perspectives are combined to yield a decision, and only the selected dialogue contents are fed into State Generator, which explicitly minimizes the distracting information passed to the downstream state prediction. Experimental results show that our approach achieves new state-of-the-art performance on MultiWOZ 2.1 and MultiWOZ 2.2, and achieves superior performance on multiple mainstream benchmark datasets (including Sim-M, Sim-R, and DSTC2).[1]

1 Introduction

Task-oriented dialogue systems have recently attracted growing attention and achieved substantial progress. Dialogue state tracking (DST) is a core component, where it is responsible for interpreting user goals and intents and feeding downstream policy learning in dialogue management. The common practice treats it as a problem of compacting the dialogue content into a series of slot-value pairs that represent information about the user goals updated until the current turn. For example, in Figure 1, the dialogue state at turn 2 is {("$hotel-type$", "$guesthouse$"), ("$hotel-pricerange$", "$cheap$")}.

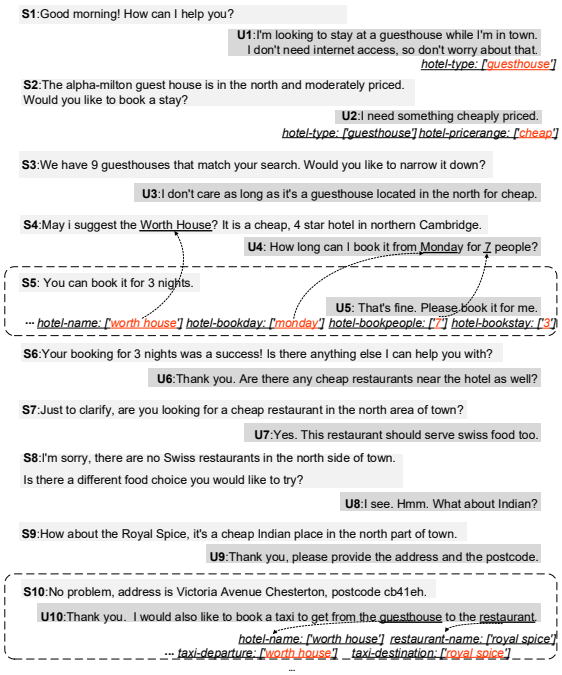


Figure 1: An example of multi-domain dialogues. Utterances at the left and the right sides are from system and user, respectively. Each red slot value in the figure indicates that it is updated in its turn.

In dialogue state tracking, dialogue history is a crucial source material. Recently, granularity has been proposed to quantify the utilization of dialogue history(Yang et al., 2021). In DST, the definition of granularity is the number of dialogue turns spanning from a certain dialogue state in the dialogue to the current dialogue state. Traditional DST models usually determine dialogue

*Corresponding author.
[1]Code is available at https://github.com/guojinyu88/DiCoS-master

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2320 - 2332

states by considering only utterances at the current turn (i.e., $\text{granularity} = 1$), while recent researches attempt to utilize partial history (i.e., $\text{granularity} = k,\ k < T$) or introduce all dialogue history information into the prediction (i.e., $\text{granularity} = T$). However, no matter what granularity is used, we find that each model uses a constant granularity it determines, regardless of which slot is being updated. Apparently, it requires different granularity for different slots in different turns. For example, in Figure 1, the granularity required for slot "$hotel-name$", "$hotel-bookday$", and "$hotel-bookpeople$" in turn 5 is 2, while slot "$hotel-bookstay$" in turn 5 requires a granularity of 1. Therefore, using a constant granularity may lead to insufficient input for updating some slots, while for others, redundant while confusing contents can become distracting information to pose a hindrance, which affects the overall performance.

Furtherly, granularity means directly working on all dialogue contents from a particular turn to the current turn, regardless of the fact that there are still dialogue contents that are not relevant to the slot. Therefore, if it is possible to break the limitation of granularity and to dynamically select relevant dialogue contents corresponding to each slot, the selected dialogue contents as input will explicitly minimize distracting information being passed to the downstream state prediction.

To achieve this goal, we propose a DiCoS-DST to fully exploit the utterances and elaborately select the relevant dialogue contents corresponding to each slot for state updating. Specifically, we retrieve turn-level utterances of dialogue history and evaluate their relevance to the slot from a combination of three perspectives. First, we devise an SN-DH module to touch on the relation of the dialogue and the slot name, which straightforward reflects the relevance. Second, we propose a CT-DH module to explore the dependency between each turn in the dialogue history and the current turn dialogue. The intuition behind this design is that the current turn dialogue is crucial. If any previous turn is strongly related to the current turn dialogue, it can be considered useful as dependency information for slot updating. Third, we propose an Implicit Mention Oriented Reasoning module to tackle the implicit mention (i.e., coreferences) problem that commonly exists in complex dialogues. Specifically, we build a novel graph neural network (GNN) to explicitly facilitate reasoning over the turns of dialogue and all slot-value pairs for better exploitation of the coreferential relation information. After the evaluation of these three modules, we leverage a gate mechanism to combine these perspectives and yield a decision. Finally, the selected dialogue contents are fed into State Generator to enhance their interaction, form a new contextualized sequence representation, and generate a value using a hybrid method.

We evaluate the effectiveness of our model on most mainstream benchmark datasets on task-oriented dialogue. Experimental results show that our proposed DiCoS-DST achieves new state-of-the-art performance on both two versions of the most actively studied dataset: MultiWOZ 2.1 (Eric et al., 2019) and MultiWOZ 2.2 (Zang et al., 2020) with joint goal accuracy of 61.02% and 61.13%. In particular, the joint goal accuracy on MultiWOZ 2.2 outperforms the previous state-of-the-art by 3.09%. In addition, DiCoS-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 et al., 2014).

Our contributions in this work are three folds:

- We propose a Multi-Perspective Dialogue Collaborative Selector module to dynamically select relevant dialogue contents corresponding to each slot from a combination of three perspectives. This module can explicitly filter the distracting information being passed to the downstream state prediction.
- We propose Implicit Mention Oriented Reasoning and implement it by building a GNN to explicitly facilitate reasoning and exploit the coreferential relation information in complex dialogues.
- Our DiCoS-DST model achieves new state-of-the-art performance on the MultiWOZ 2.1, MultiWOZ 2.2, Sim-M, and Sim-R datasets.

2 Related Work

There has been a plethora of research on dialogue state tracking. Traditional dialogue state trackers relied on a separate Spoken Language Understanding (SLU) module (Thomson and Young, 2010; Wang and Lemon, 2013) to extract relevant information. In recent years, neural network models are proposed for further improvements. One way to classify DST models is whether they use dialogue

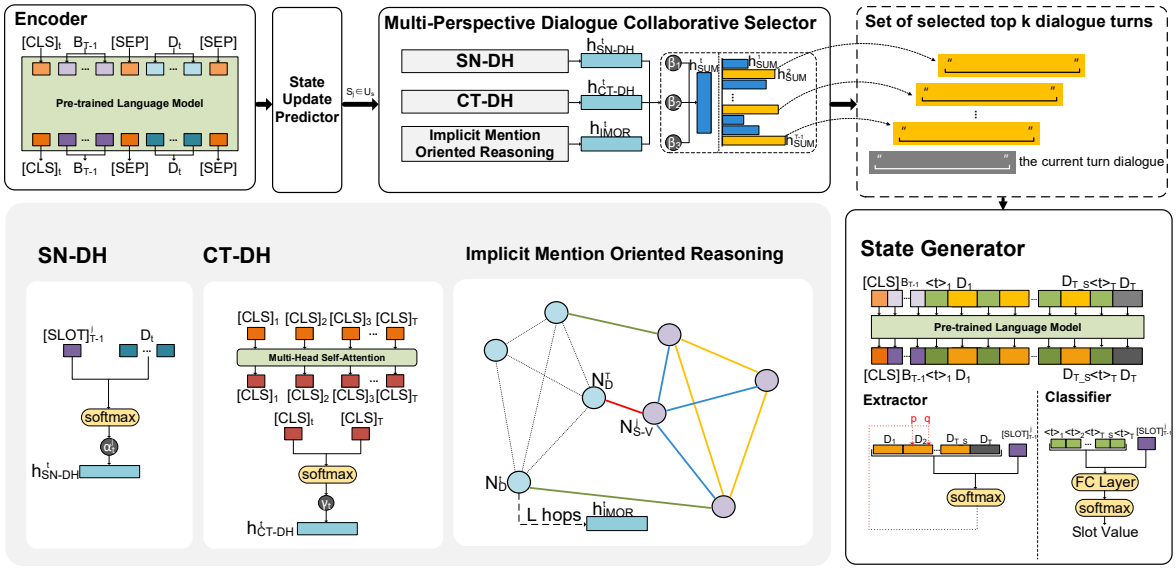


Figure 2: The architecture of the proposed DiCoS-DST model. The gray area in the lower left part of the figure shows the internal structure of the three modules in Multi-Perspective Dialogue Collaborative Selector.

history. Some DST models obtain each slot value in the dialogue state by inquiring about a part or all of the dialogue history (Xu and Hu, 2018; Lei et al., 2018; Goel et al., 2019; Ren et al., 2019; Shan et al., 2020; Zhang et al., 2020; Chen et al., 2020; Guo et al., 2021), while the others use the current turn dialogue to predict the dialogue state (Mrkšić et al., 2017; Kim et al., 2020; Heck et al., 2020; Zhu et al., 2020). Recently, (Yang et al., 2021) first proposed the granularity in DST to quantify the use of dialogue history. Its experimental results show that different models on different datasets have different optimal granularity (not always using the entire dialogue history). However, no matter what granularity is used, each model uses a constant granularity it determines, regardless of which slot is updated.

On the other hand, dialogue state tracking and machine reading comprehension (MRC) have similarities in many aspects (Gao et al., 2020). Recently, Multi-hop Reading Comprehension (MHRC) has been a challenging topic. For cases in MHRC datasets, one question is usually provided with several lexically related paragraphs, which contain many confusing contexts. To deal with this situation, cascaded models (Qiu et al., 2019; Groeneveld et al., 2020; Tu et al., 2020; Wu et al., 2021) that are composed of a reader and a retriever are often used. They retrieve the most relevant evidence paragraphs first and perform multi-hop reasoning on retrieved contexts thereafter. The mechanism of dialogue selection before state generation in our work is partially inspired by the paragraph retrieval in multi-hop reading comprehension.

3 Approach

The architecture of DiCoS-DST is illustrated in Figure 2. DiCoS-DST consists of Encoder, State Update Predictor, Multi-Perspective Dialogue Collaborative Selector, and State Generator. Here we first define the problem setting in our work. We define the number of the current turn as T. The task is to predict the dialogue state at each turn t ($t \leq T$), which is defined as $\mathcal{B}_t = \{(S^j, V_t^j) | 1 \leq j \leq J\}$, where S^j is the slot name, V_t^j is the corresponding slot value, and J is the total number of slots. For the sake of simplicity, we omit the superscript T in the variables in the next sections.

3.1 Encoder

We employ the representation of the previous turn dialogue state B_{T-1} concatenated to the representation of each turn dialogue utterances D_t as input: $E_t = [\text{CLS}]_t \oplus B_{T-1} \oplus [\text{SEP}] \oplus D_t$, $(1 \leq t \leq T)$, where $[\text{CLS}]_t$ is a special token added in front of every turn input. The representation of the previous turn dialogue state is $B_{T-1} = B_{T-1}^1 \oplus \ldots \oplus B_{T-1}^J$. The representation of each slot's state $B_{T-1}^j = [\text{SLOT}]_{T-1}^j \oplus S_j \oplus [\text{VALUE}]_{T-1}^j \oplus V_{T-1}^j$, where $[\text{SLOT}]_{T-1}^j$ and $[\text{VALUE}]_{T-1}^j$ are special tokens that represent the slot name and the slot value

at turn $T-1$, respectively. We donate the representation of the dialogue at turn t as $D_t = R_t \oplus ; \oplus U_t \oplus [\text{SEP}]$, where R_t is the system response and U_t is the user utterance. ; is a special token used to mark the boundary between R_t and U_t, and [SEP] is a special token used to mark the end of a dialogue turn.

Then a pre-trained language model (PrLM) will be adopted to obtain contextualized representation for the concatenated input sequence E_t.

3.2 State Update Predictor

We attach a two-way classification module to the top of the Encoder output. It predicts which slots require to be updated in the current turn. The subsequent modules will only process the selected slots, while the other slots will directly inherit the slot values from the previous turn.

We inject this module because whether a slot requires to be updated indicates whether the current turn dialogue is significant for this slot. For CT-DH of the subsequent Multi-Perspective Collaborative Selector, the great importance of the current turn dialogue is a prerequisite. A more detailed explanation will be given in Section 3.3.

We employ the same mechanism as (Guo et al., 2021) to train the module and to predict the state operation. We sketch the prediction process as follows:

$$\text{SUP}(S_j) = \begin{cases} \text{update}, & \text{if Total\_}score_j > \delta \\ \text{inherit}, & \text{otherwise} \end{cases} \quad (1)$$

We define the set of the selected slot indices as $\mathbf{U}_s = \{j | \text{SUP}(S_j) = \text{update}\}$.

3.3 Multi-Perspective Dialogue Collaborative Selector

For each slot S_j ($j \in \mathbf{U}_s$) selected to be updated, SN-DH, CT-DH, and Implicit Mention Oriented Reasoning modules are proposed to evaluate dialogue relevance and aggregate representations from three perspectives. Then a gated fusion mechanism is implemented to perform the dialogue selection.

SN-DH SN-DH (Slot Name - Dialogue History) aims to explore the correlation between slot names and each turn of the dialogue history. For slot S_j, the slot name is straightforward explicit information. Therefore, the correlation with the slot name directly reflects the importance of the dialogue turn. We take the slot name presentation $[\text{SLOT}]^j_{T-1}$ as the attention to the t-th turn dialogue representation D_t. The output $\boldsymbol{\alpha}^j_t = \text{softmax}(D_t([\text{SLOT}]^j_{T-1})^\intercal)$ represents the correlation between each position of D_t and the j-th slot name at turn t. Then we get the aggregated dialogue representation $h^t_{\text{SN-DH}} = (\boldsymbol{\alpha}^j_t)^\intercal D_t$, which will participate in the subsequent fusion as the embedding of the t-th turn dialogue in this perspective.

CT-DH As aforementioned, a slot that needs to be updated in the current turn means that the current turn dialogue is most relevant to this slot. In this case, if the dialogue content of any other turn contains the information that the current turn dialogue highly depends on, it can also be considered useful. Based on this consideration, we devise a CT-DH (Current Turn - Dialogue History) module to explore this association. Specifically, we build a multi-head self-attention (MHSA) layer on top of the [CLS] tokens generated from different turns of dialogue to enhance inter-turn interaction. The MHSA layer is defined as:

$$head_i = \text{Attention}(QW^Q_i, KW^K_i, VW^V_i) \quad (2)$$

$$Multihead = (head_i \oplus \ldots \oplus head_n)W^O \quad (3)$$

$$I = \text{MHSA}([\text{CLS}]_1 \oplus \ldots \oplus [\text{CLS}]_T) \quad (4)$$

where Q, K, and V are linear projections from [CLS] embeddings of each turn of dialogue, representing attention queries, key and values.

We then append an attention layer between the output representation of the current turn dialogue and each turn of dialogue history to capture interactions between them:

$$\boldsymbol{\gamma}_t = \text{Attention}([\text{CLS}]_t, [\text{CLS}]_T) \quad (5)$$

$$h^t_{\text{CT-DH}} = \boldsymbol{\gamma}_t [\text{CLS}]_T + [\text{CLS}]_t \quad (6)$$

$h^t_{\text{CT-DH}}$ will participate in the subsequent fusion as an aggregated representation of the t-th dialogue in this perspective.

Implicit Mention Oriented Reasoning Handling a complex dialogue usually requires addressing implicit mentions (i.e., coreferences). As shown in Figure 1, in turn 10, the restaurant is not referred to explicitly upon ordering a taxi within the same dialogue turn. Instead, it is present in the value of another slot. Therefore, SN-DH and CT-DH are difficult to deal with this case due to their mechanisms. To tackle this problem, we build a graph neural network (GNN) model to explicitly facilitate reasoning over the turns of dialogue and all slot-value pairs for better exploitation of the

coreferential relation. As illustrated in Figure 3, the nodes in the graph include two types: N_D for each turn dialogue and N_{S-V} for each slot-value pair. They are initialized with the MHSA output representation $[\text{CLS}]_t$ and $W_{S-V}([\text{SLOT}]^z_{T-1} \oplus [\text{VALUE}]^z_{T-1})$ $(1 \leq z \leq J)$, respectively. Then we design four types of edges to build the connections among graph nodes:

1) Add an edge between N^j_{S-V} and N^T_D (red line in Figure 3). As aforementioned, the slot S_j will be updated. This edge is to establish the connection between the slot to be updated and the current turn dialogue;

2) Add an edge between N^j_{S-V} and N^z_{S-V} $(z \neq j)$ (blue line in Figure 3). These edges are to establish connections between the slot to be updated and other slots;

3) Add an edge between N^z_{S-V} $(z \neq j)$ and $N^{t_z}_D$. t_z is the turn when the most up-to-date value of S_z is updated (green line in Figure 3). These edges are to establish connections between each slot and the turn of dialogue in which its latest slot value was updated;

4) Add an edge between $N^{z_1}_{S-V}$ and $N^{z_2}_{S-V}$ (S_{z_1} and S_{z_2} belong to the same domain) (yellow line in Figure 3). These edges are to establish connections between slots that belong to the same domain.

The motivation for this design is that we first explore the relation between the slot to be updated and other slot-value pairs based on the current turn dialogue. Then we use other slot-value pairs as media to establish relations to their corresponding dialogue turns. We add the fourth type of edges to represent the auxiliary relationship of slots that belong to the same domain.

We use multi-relational GCN with gating mechanism as in (De Cao et al., 2019; Tu et al., 2019). We define h^0_i represents initial node embedding from N_D or N_{S-V}. The calculation of node embedding after one hop can be formulated as:

$$h^{l+1}_i = \sigma(u^l_i) \odot g^l_i + h^l_i \odot (1 - g^l_i) \quad (7)$$

$$u^l_i = f_s(h^l_i) + \sum_{r \in \mathcal{R}} \frac{1}{|\mathcal{N}^r_i|} \sum_{n \in \mathcal{N}^r_i} f_r(h^l_n) \quad (8)$$

$$g^l_i = \text{sigmoid}(f_g([u^l_i; h^l_i])) \quad (9)$$

$\mathcal{N}^r_i$ is the neighbors of node i with edge type r, $\mathcal{R}$ is the set of all edge types, and h^l_n is the node representation of node n in layer l. $|\cdot|$ indicates the size of the neighboring set. Each of f_r, f_s, f_g can be implemented with an MLP. Gate control g^l_i is a vector consisting of values between 0 and 1 to control the amount information from computed update u^l_i or from the original h^l_i. Function σ denotes a non-linear activation function.

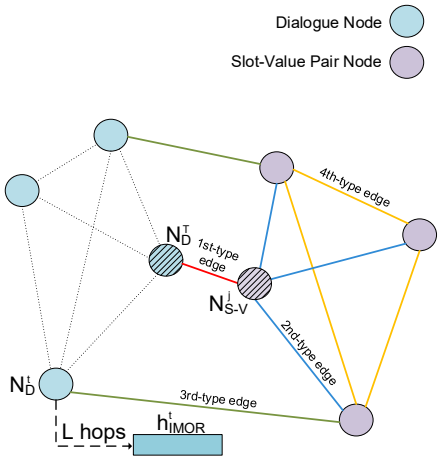


Figure 3: Diagram of the graph neural network. The dashed connection between the dialogue nodes does not actually exist. We draw them to show that using the dialogue representation output by MHSA already includes the contextual interactions between the dialogues.

After the message passes on the graph with L hops, we take the final representation of the t-th turn dialogue node N^t_D as the aggregated representation h^t_{IMOR} in this perspective.

Gating Fusion and Collaborative Selection The representations $h^t_{\text{SN-DH}}$, $h^t_{\text{CT-DH}}$, and h^t_{IMOR} of the t-th turn dialogue enter this module for fusion and ranking. To balance the information from multiple perspectives, we leverage a gate mechanism to compute a weight to decide how much information from each perspective should be combined. It is defined as follows:

$$\beta_1 = \sigma_1(W_{\beta_1} \tanh(W_1 h^t_{\text{SN-DH}})) \quad (10)$$

$$\beta_2 = \sigma_2(W_{\beta_2} \tanh(W_2 h^t_{\text{CT-DH}})) \quad (11)$$

$$\beta_3 = \sigma_3(W_{\beta_3} \tanh(W_3 h^t_{\text{IMOR}})) \quad (12)$$

$$h^t_{\text{sum}} = \beta_1 h^t_{\text{SN-DH}} + \beta_2 h^t_{\text{CT-DH}} + \beta_3 h^t_{\text{IMOR}} \quad (13)$$

After the fusion, an MLP layer is followed, and then we take the dialogues of the top k ranked turns as the selected dialogue contents.

It is worth mentioning that, unlike the state update predictor, since there is no ground-truth label of the dialogue turns that should be selected corresponding to each slot, we take this module and the following state generator as a whole and

train it under the supervision of the final dialogue state label. We mark each selected dialogue turn to make the gradient of the state generator losses only backpropagate to the marked turns to ensure the effectiveness of supervision.

3.4 State Generator

The selected dialogue content will be utilized to jointly update the dialogue state.

Cascaded Context Refinement After acquiring a nearly noise-free set $\mathbf{U}_D$ of selected dialogue turns, we consider that directly using their representations as inputs may ignore the cross attention between them since they are used as a whole. As a result, we concatenate these dialogue utterances together to form a new input sequence $C = [\text{CLS}] \oplus B_{T-1} \oplus \langle t \rangle_1 \oplus D_1 \oplus \ldots \oplus \langle t \rangle_{T\_S} \oplus D_{T\_S} \oplus \langle t \rangle_T \oplus D_T$ ($T\_S = |\mathbf{U}_D|$).

Especially, we inject an indicator token "$\langle t \rangle$" before each turn of dialogue utterance to get aggregated turn embeddings for the subsequent classification-based state prediction. Then we feed this sequence into a single PrLM to obtain the contextualized output representation.

Slot Value Generation We first attempt to obtain the value using the extractive method from representation $C_{\text{E}} = D_1 \oplus D_2 \oplus \ldots \oplus D_{T\_S} \oplus D_T$:

$$p = \text{softmax}(W_s C_{\text{E}}([\text{SLOT}]_{T-1}^{j})^{\intercal}) \quad (14)$$

$$q = \text{softmax}(W_e C_{\text{E}}([\text{SLOT}]_{T-1}^{j})^{\intercal}) \quad (15)$$

The position of the maximum value in p and q will be the start and end predictions of the slot value. If this prediction does not belong to the candidate value set of S_j, we use the representation of $C_{\text{C}} = \langle t \rangle_1 \oplus \langle t \rangle_2 \oplus \ldots \oplus \langle t \rangle_{T\_S} \oplus \langle t \rangle_T$ to get the distribution and choose the candidate slot value corresponding to the maximum value:

$$y = \text{softmax}(W_{\text{C}} C_{\text{C}}([\text{SLOT}]_{T-1}^{j})^{\intercal}) \quad (16)$$

We define the training objectives of two methods as cross-entropy loss:

$$L_{\text{ext}} = -\frac{1}{|\mathbf{U}_s|} \sum_{j}^{|\mathbf{U}_s|} (p \log \hat{p} + q \log \hat{q}) \quad (17)$$

$$L_{\text{cls}} = -\frac{1}{|\mathbf{U}_s|} \sum_{j}^{|\mathbf{U}_s|} y \log \hat{y} \quad (18)$$

where $\hat{p}$ and $\hat{q}$ are the targets indicating the proportion of all possible start and end, and $\hat{y}$ is the target indicating the probability of candidate values.

4 Experiments

4.1 Datasets and Metrics

We conduct experiments on most of the mainstream benchmark datasets on task-oriented dialogue, including MultiWOZ 2.1, MultiWOZ 2.2, Sim-R, Sim-M, and DSTC2. MultiWOZ 2.1 and MultiWOZ 2.2 are two versions of a large-scale multi-domain task-oriented dialogue dataset. It is a fully-labeled collection of human-human written dialogues spanning over multiple domains and topics. Sim-M and Sim-R are multi-turn dialogue datasets in the movie and restaurant domains, respectively. DSTC2 is collected in the restaurant domain.

We use joint goal accuracy and slot accuracy as evaluation metrics. Joint goal accuracy refers to the accuracy of the dialogue state in each turn. Slot accuracy only considers slot-level accuracy.

4.2 Baseline Models

We compare the performance of DiCoS-DST with the following baselines: TRADE encodes the dialogue and decodes the value using a copy-augmented decoder (Wu et al., 2019). BERT-DST generates language representations suitable for scalable DST (Chao and Lane, 2019). DST+LU presents an approach for multi-task learning of language understanding and DST (Rastogi et al., 2018). TripPy extracts values from the dialogue context by three copy mechanisms (Heck et al., 2020). DSS-DST consists of the slot selector based on the current turn dialogue, and the slot value generator based on the dialogue history (Guo et al., 2021). Seq2Seq-DU employs two BERT-based encoders to respectively encode the utterances and the descriptions of schemas (Feng et al., 2021). Pegasus-DST applies a span prediction-based pre-training objective designed for text summarization to DST (Zhao et al., 2021). DST-as-Prompting uses schema-driven prompting to provide task-aware history encoding (Lee et al., 2021).

4.3 Implementation Details

We employ a pre-trained ALBERT-large-uncased model (Lan et al., 2019) for the encoder. The hidden size of the encoder d is 1024. We use AdamW optimizer (Loshchilov and Hutter, 2018) and set the warmup proportion to 0.01 and L2 weight decay of 0.01. We set the peak learning rate of State Update Predictor the same as in DSS-DST and the peak learning rate of the other modules to 0.0001. We set the dropout (Srivastava et al., 2014) rate

Model	MultiWOZ 2.1		MultiWOZ 2.2		Sim-M	Sim-R	DSTC2
	Joint	Slot	Joint	Slot	Joint	Joint	Joint
TRADE	45.60	-	45.40	-	-	-	-
DST+LU	-	-	-	-	46.0	84.9	-
BERT-DST	-	-	-	-	80.1	89.6	69.3
TripPy	55.29	-	-	-	83.5	90.0	-
Pegasus-DST	54.40	-	57.60	-	-	-	73.6
DST-as-Prompting	56.66	-	57.60	-	83.3	90.6	-
Seq2seq-DU	56.10	-	54.40	-	-	-	**85.0**
DSS-DST	60.73	98.05	58.04	97.66	-	-	-
DiCoS-DST ($k = 1$)	60.89 ($\pm$0.47)	98.05 ($\pm$0.02)	61.04 ($\pm$0.56)	98.05 ($\pm$0.04)	84.5 ($\pm$1.2)	91.2 ($\pm$0.3)	77.7 ($\pm$0.2)
DiCoS-DST ($k = 2$)	**61.02** ($\pm$0.41)	**98.05** ($\pm$0.02)	**61.13** ($\pm$0.54)	**98.06** ($\pm$0.03)	**84.7** ($\pm$1.1)	**91.5** ($\pm$0.3)	78.4 ($\pm$0.2)
DiCoS-DST ($k = 3$)	60.85 ($\pm$0.24)	98.05 ($\pm$0.01)	60.88 ($\pm$0.33)	98.05 ($\pm$0.03)	83.8 ($\pm$1.1)	91.0 ($\pm$0.2)	77.3 ($\pm$0.2)

Table 1: Accuracy (%) on the test sets of benchmark datasets vs. various approaches as reported in the literature.

PrLM	MultiWOZ 2.2
ALBERT (large)	**61.13**
ALBERT (base)	60.05(-1.08)
BERT (large)	60.16(-0.97)
BERT (base)	59.51(-1.62)

Table 2: Ablation study with joint goal accuracy (%).

Model	MultiWOZ 2.2
DiCoS-DST	**61.13**
-State Update Predictor	58.48 (-2.65)
-Multi-Perspective Dialogue Collaborative Selector	54.94 (-6.19)
-Cascaded Context Refinement	59.75 (-1.38)

Table 3: Ablation study with joint goal accuracy (%). Each performance in this table represents the test results after the model was retrained with the corresponding module removed. "- State Update Predictor" means that all slots are updated in each turn. "-Multi-Perspective Dialogue Collaborative Selector" means that using the entire dialogue history without selection. "-Cascaded Context Refinement" means that directly using the representation of selected turns from the dialogue selector without context refinement.

to 0.1. We utilize word dropout (Bowman et al., 2016) with the probability of 0.1. We set L to 3. The max sequence length for all inputs is fixed to 256. During training the Multi-Perspective Dialogue Collaborative Selector, we use the ground truth selected slots instead of the predicted ones. We report the mean joint goal accuracy over 10 different random seeds to reduce statistical errors.

4.4 Main Results

Table 1 shows the performance of our DiCoS-DST and other baselines. Our model achieves state-of-the-art performance on MultiWOZ 2.1 and MultiWOZ 2.2 with joint goal accuracy of 61.02% and 61.13%. In particular, the joint goal accuracy on MultiWOZ 2.2 outperforms the previous state-of-the-art by 3.09%. Besides, despite the sparsity of experimental results on Sim-M and Sim-R, our model still achieves state-of-the-art performance on these two datasets. On DSTC2, the performance of our model is also competitive. Among our models, DiCoS-DST ($k = 2$) performs the best on all datasets. Especially, DiCoS-DST ($k = 2$) and DiCoS-DST ($k = 1$) perform better than DiCoS-DST ($k = 3$). We conjecture that selecting two turns from the dialogue history may be sufficient, and introducing more turns may confuse the model.

4.5 Ablation Study

Different PrLMs We employ different pre-trained language models with different scales as the backbone for training and testing on MultiWOZ 2.2. Table 2 shows that the joint goal accuracy of other encoders decreases in varying degrees compared with ALBERT (large). The joint goal accuracy of BERT(base) decreases by 1.62%, but still outperforms the previous state-of-the-art performance on MultiWOZ 2.2. This demonstrates that our model achieves consistent performance gain in all fair

Perspective(s)	MultiWOZ 2.2
SN-DH	57.73 (-3.40)
CT-DH	55.47 (-5.66)
IMOR	55.11 (-6.02)
SN-DH + CT-DH	59.56 (-1.57)
SN-DH + IMOR	58.68 (-2.45)
CT-DH + IMOR	56.79 (-4.34)
SN-DH + CT-DH + IMOR	**61.13**

Table 4: Ablation study with joint goal accuracy (%). IMOR stands for Implicit Mention Oriented Reasoning.

Graph	MultiWOZ 2.2
Original Graph (DiCoS-DST)	**61.13**
-1st type of edges	59.70 (-1.43)
-2nd type of edges	59.62 (-1.51)
-3rd type of edges	59.78 (-1.35)
-4th type of edges	60.65 (-0.48)
+fully connecting all dialogue nodes	61.01 (-0.12)
+3rd type of edges between each N^z_{S-V} and all N^t_D	60.04 (-1.09)

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

comparison environments with other methods.

Effect of Core Components To explore the effectiveness of core components, we conduct an ablation study of them on MultiWOZ 2.2. As shown in Table 3, we observe that the performance degrades by 2.65% for joint goal accuracy when the State Update Predictor is removed. It is worth mentioning that this performance still outperforms the previous state-of-the-art performance, which demonstrates that the large performance gain of DiCoS-DST over other baselines comes from its dialogue selection. This is also supported by the observation that the performance of the model without the Multi-Perspective Dialogue Collaborative Selection module drops drastically (degrades by 6.19% for joint goal accuracy). In addition, when we remove the Cascaded Context Refinement module, we lose 1.38%, indicating the usefulness of interaction between different dialogue turns.

Separate Perspective and Combinations We explore the performance of each separate perspective and their various combinations. When a perspective needs to be masked, we set their corresponding gating weights to 0. It can be observed in Table 4 that the SN-DH module has the greatest impact on performance, and the most effective

MultiWOZ 2.2		
k	DiCoS-DST	Granularity-Based
1	**61.04**	59.58 (-1.46)
2	**61.13**	59.88 (-1.25)
3	**60.88**	59.91 (-0.97)

Table 6: The joint goal accuracy (%) of different k. The state generator is re-trained with the corresponding selected turns as input for granularity-based methods.

MultiWOZ 2.2			
Domain	$k=0$	$k=1$	$k=2$
Attraction	**79.15**	79.04	78.79
Hotel	56.95	**58.07**	58.02
Restaurant	73.81	74.73	**75.14**
Taxi	53.50	55.12	**56.33**
Train	75.13	76.89	**77.26**

Table 7: Domain-specific results on MultiWOZ 2.2.

combination of perspectives is the combination of SN-DH and CT-DH. Despite the simplicity of the mechanism of SN-DH, the association with the slot name straightforward reflects the importance of the dialogue. To solve the common problem of coreferences in complex dialogues, the Implicit Mention Oriented Reasoning module improves the performance close enough to the CT-DH.

Graph Edges Ablation We investigate the effect of the different edges in the GNN. As shown in Table 5, the performance degradation is relatively obvious when the first, second, and third types of edges are removed separately. It indicates that the majority of the connections are indeed to construct the reasoning logic, while the correlation of the same domain's slots plays an auxiliary role. In addition, we design two comparative experiments. First, we start naively by fully connecting all dialogue nodes to enhance the interaction among dialogue turns. However, this change does not give a clear benefit. This is mostly because the initialization of the dialogue nodes using the dialogue representation output by MHSA already includes the contextual interactions between the dialogues. Second, we add a third type of edges between each slot-value pair node and all dialogue nodes without distinguishing the correspondence. We observe that this change does harm to the performance (degrades by 1.09%). This reflects the importance of using other slots to explore their corresponding turns of dialogues when dealing with coreferences.

5 Analysis

5.1 Is It Beyond the Granularity?

DiCoS-DST filters out some distracting information by selecting relevant dialogues, but is it really beyond the granularity? To investigate it, we simulate the granularity and compare it with DiCoS-DST. Specifically, we use the maximum granularity (i.e., the number of dialogue turns spanning from the selected furthest dialogue turn to the current turn) and capture the corresponding dialogue contents as input to State Generator. As shown in Table 6, DiCoS-DST outperforms the granularity-based method by 1.46% ($k = 1$), 1.25% ($k = 2$), and 0.97% ($k = 3$), indicating that there is still redundant information in the dialogue contents determined by the granularity that confuses the model.

5.2 Domain-Specific Dialogue Requirements

Table 7 shows the domain-specific results when we set different values for k ($k = 0, 1, 2$). In $taxi$ and $train$ domains, the performance of the model decreases significantly when $k = 0$ compared to $k = 2$, implying that acquiring the values of the slots in these domains is highly dependent on the dialogue history. Nevertheless, there is no significant difference in the performance in $attraction$ domain when we set different values for k. This indicates that the values of the slots in this domain can usually be simply obtained from the current turn dialogue, instead of using the dialogue history or resolving coreferences.

6 Conclusion

We introduce an effective DiCoS-DST that dynamically selects the relevant dialogue contents corresponding to each slot from a combination of three perspectives. The dialogue collaborative selector module performs a comprehensive selection for each turn dialogue based on its relation to the slot name, its connection to the current turn dialogue, and the implicit mention oriented reasoning. Then only the selected dialogue contents are fed into State Generator, which explicitly minimizes the distracting information passed to the downstream state prediction. Our DiCoS-DST model achieves new state-of-the-art performance on the MultiWOZ benchmark, and achieves competitive performance on most other DST benchmark datasets. The potential relationship among the above perspectives is a promising research direction, and we will explore it for more than dialogue selection in the future.

Acknowledgements

This work was supported by Beijing Natural Science Foundation(Grant No. 4222032) and BUPT Excellent Ph.D. Students Foundation. We thank the anonymous reviewers for their insightful comments.

Ethical Considerations

The claims in this paper match the experimental results. This work focuses on DST in task-oriented dialogue systems, and the improvements could have a positive impact on helping humans to complete goals more effectively in a more intelligent way of communication.

References

Samuel Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

Guan-Lin Chao and Ian Lane. 2019. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. *arXiv preprint arXiv:1907.03040*.

Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020. Schema-guided multi-domain dialogue state tracking with graph attention neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7521–7528.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2019. Question answering by reasoning across documents with graph convolutional networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2306–2317.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Yue Feng, Yang Wang, and Hang Li. 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)*, pages 1714–1725. Association for Computational Linguistics.

Shuyang Gao, Sanchit Agarwal, Di Jin, Tagyoung Chung, and Dilek Hakkani-Tur. 2020. From machine reading comprehension to dialogue state tracking: Bridging the gap. In *Proceedings of the 2nd*

Workshop on Natural Language Processing for Conversational AI, pages 79–89.

Rahul Goel, Shachi Paul, and Dilek Hakkani-Tür. 2019. Hyst: A hybrid approach for flexible and accurate dialogue state tracking. *arXiv preprint arXiv:1907.00883*.

Dirk Groeneveld, Tushar Khot, Ashish Sabharwal, et al. 2020. A simple yet strong pipeline for hotpotqa. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8839–8845.

Jinyu Guo, Kai Shuang, Jijie Li, and Zihan Wang. 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)*, pages 139–151.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 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*, pages 35–44.

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

Sungdong Kim, Sohee Yang, Gyuwan Kim, and Sang-Woo Lee. 2020. Efficient dialogue state tracking by selectively overwriting memory. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 567–582.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Chia-Hsuan Lee, Hao Cheng, and Mari Ostendorf. 2021. Dialogue state tracking with a language model using schema-driven prompting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4937–4949.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 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)*, pages 1437–1447.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 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)*, pages 1777–1788.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6140–6150.

Abhinav Rastogi, Raghav Gupta, and Dilek Hakkani-Tur. 2018. Multi-task learning for joint language understanding and dialogue state tracking. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 376–384.

Liliang Ren, Jianmo Ni, and Julian McAuley. 2019. Scalable and accurate dialogue state tracking via hierarchical sequence generation. 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)*, pages 1876–1885.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Yong Shan, Zekang Li, Jinchao Zhang, Fandong Meng, Yang Feng, Cheng Niu, and Jie Zhou. 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*, pages 6322–6333.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Blaise Thomson and Steve Young. 2010. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, Xiaodong He, and Bowen Zhou. 2020. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9073–9080.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2704–2713.

Zhuoran Wang and Oliver Lemon. 2013. A simple and generic belief tracking mechanism for the dialog state tracking challenge: On the believability of observed information. In *Proceedings of the SIGDIAL 2013 Conference*, pages 423–432.

Bohong Wu, Zhuosheng Zhang, and Hai Zhao. 2021. Graph-free multi-hop reading comprehension: A select-to-guide strategy. *arXiv preprint arXiv:2107.11823*.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. *arXiv preprint arXiv:1905.08743*.

Puyang Xu and Qi Hu. 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)*, pages 1448–1457.

Puhai Yang, Heyan Huang, and Xian-Ling Mao. 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)*, pages 2481–2491. Association for Computational Linguistics.

Xiaoxue Zang, Abhinav Rastogi, and Jindong Chen. 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*, pages 109–117.

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, S Yu Philip, Richard Socher, and Caiming Xiong. 2020. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 154–167.

Jeffrey Zhao, Mahdis Mahdieh, Ye Zhang, Yuan Cao, and Yonghui Wu. 2021. Effective sequence-to-sequence dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7486–7493.

Su Zhu, Jieyu Li, Lu Chen, and Kai Yu. 2020. Efficient context and schema fusion networks for multi-domain dialogue state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 766–781.

Appendices

A Visualization

Dialogue Example

S1:Good morning! How can I help you?

U1:I'm looking to stay at a guesthouse while I'm in town. I don't need internet access, so don't worry about that.

hotel-type: ['guesthouse']

S2:The alpha-milton guest house is in the north and moderately priced. Would you like to book a stay?

U2:I need something cheaply priced.

hotel-type: ['guesthouse'] hotel-pricerange: ['cheap']

S3:We have 9 guesthouses that match your search. Would you like to narrow it down?

U3:I don't care as long as it's a guesthouse located in the north for cheap.

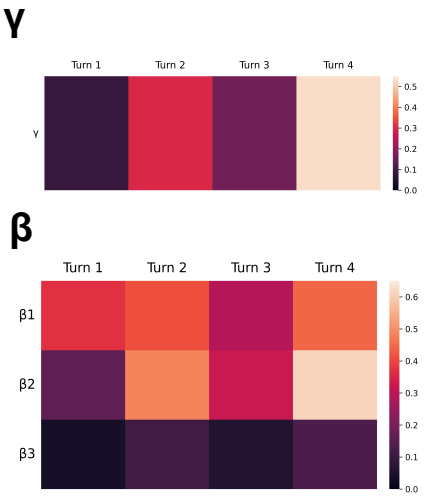
S4:May i suggest the Worth House? It is a cheap, 4 star hotel in northern Cambridge.

U4: How long can I book it from Monday for 7 people?

S5: You can book it for 3 nights.

U5: That's fine. Please book it for me.

··· *hotel-name: ['worth house'] hotel-bookday: ['monday'] hotel-bookpeople: ['7'] hotel-bookstay: ['3']*



B Statistics of datasets in experiments

Characteristics	MultiWOZ 2.1	MultiWOZ 2.2	Sim-M & Sim-R	DSTC2
No. of domains	7	8	2	1
No. of dialogues	8,438	8,438	1,500	1612
Total no. of turns	113,556	113,556	14,796	23354
Avg. turns per dialogue	13.46	13.46	9.86	14.49
Avg. tokens per turn	13.38	13.13	8.24	8.54
No. of categorical slots	37	21	0	3
No. of non-categorical slots	0	40	14	0
Have schema description	Yes	Yes	No	No