

Dual Slot Selector via Local Reliability Verification for Dialogue State Tracking

Jinyu Guo¹, Kai Shuang^{1*}, Jijie Li¹ and Zihan Wang²

¹State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications

²Graduate School of Information Science and Technology, The University of Tokyo

{guojinyu, shuangk, lijijie}@bupt.edu.cn

zwang@tkl.iis.u-tokyo.ac.jp

Abstract

The goal of dialogue state tracking (DST) is to predict the current dialogue state given all previous dialogue contexts. Existing approaches generally predict the dialogue state at every turn from scratch. However, the overwhelming majority of the slots in each turn should simply inherit the slot values from the previous turn. Therefore, the mechanism of treating slots equally in each turn not only is inefficient but also may lead to additional errors because of the redundant slot value generation. To address this problem, we devise the two-stage DSS-DST which consists of the Dual Slot Selector based on the current turn dialogue, and the Slot Value Generator based on the dialogue history. The Dual Slot Selector determines each slot whether to update slot value or to inherit the slot value from the previous turn from two aspects: (1) if there is a strong relationship between it and the current turn dialogue utterances; (2) if a slot value with high reliability can be obtained for it through the current turn dialogue. The slots selected to be updated are permitted to enter the Slot Value Generator to update values by a hybrid method, while the other slots directly inherit the values from the previous turn. Empirical results show that our method achieves 56.93%, 60.73%, and 58.04% joint accuracy on MultiWOZ 2.0, MultiWOZ 2.1, and MultiWOZ 2.2 datasets respectively and achieves a new state-of-the-art performance with significant improvements.¹

1 Introduction

Task-oriented dialogue has attracted increasing attention in both the research and industry communities. As a key component in task-oriented dialogue systems, Dialogue State Tracking (DST) aims to

extract user goals or intents and represent them as a compact dialogue state in the form of slot-value pairs of each turn dialogue. DST is an essential part of dialogue management in task-oriented dialogue systems, where the next dialogue system action is selected based on the current dialogue state.

Early dialogue state tracking approaches extract value for each slot predefined in a single domain (Williams et al., 2014; Henderson et al., 2014a,b). These methods can be directly adapted to multi-domain conversations by replacing slots in a single domain with domain-slot pairs predefined. In multi-domain DST, some of the previous works study the scalability of the model (Wu et al., 2019), some aim to fully utilizing the dialogue history and context (Shan et al., 2020; Chen et al., 2020a; Quan and Xiong, 2020), and some attempt to explore the relationship between different slots (Hu et al., 2020; Chen et al., 2020b). Nevertheless, existing approaches generally predict the dialogue state at every turn from scratch. The overwhelming majority of the slots in each turn should simply inherit the slot values from the previous turn. Therefore, the mechanism of treating slots equally in each turn not only is inefficient but also may lead to additional errors because of the redundant slot value generation.

To address this problem, we propose a DSS-DST which consists of the Dual Slot Selector based on the current turn dialogue, and the Slot Value Generator based on the dialogue history. At each turn, all slots are judged by the Dual Slot Selector first, and only the selected slots are permitted to enter the Slot Value Generator to update their slot value, while the other slots directly inherit the slot value from the previous turn. The Dual Slot Selector is a two-stage judging process. It consists of a Preliminary Selector and an Ultimate Selector, which jointly make a judgment for each slot according to the current turn dialogue. The intuition behind

*Corresponding author.

¹Code is available at
<https://github.com/guojinyu88/DSSDST>

this design is that the Preliminary Selector makes a coarse judgment to exclude most of the irrelevant slots, and then the Ultimate Selector makes an intensive judgment for the slots selected by the Preliminary Selector and combines its confidence with the confidence of the Preliminary Selector to yield the final decision. Specifically, the Preliminary Selector briefly touches on the relationship of current turn dialogue utterances and each slot. Then the Ultimate Selector obtains a temporary slot value for each slot and calculates its reliability. The rationale for the Ultimate Selector is that if a slot value with high reliability can be obtained through the current turn dialogue, then the slot ought to be updated. Eventually, the selected slots enter the Slot Value Generator and a hybrid way of the extractive method and the classification-based method is utilized to generate a value according to the current dialogue utterances and dialogue history.

Our proposed DSS-DST achieves state-of-the-art joint accuracy on three of the most actively studied datasets: MultiWOZ 2.0 (Budzianowski et al., 2018), MultiWOZ 2.1 (Eric et al., 2019), and MultiWOZ 2.2 (Zang et al., 2020) with joint accuracy of 56.93%, 60.73%, and 58.04%. The results outperform the previous state-of-the-art by +2.54%, +5.43%, and +6.34%, respectively. Furthermore, a series of subsequent ablation studies and analysis are conducted to demonstrate the effectiveness of the proposed method.

Our contributions in this paper are three folds:

- We devise an effective DSS-DST which consists of the Dual Slot Selector based on the current turn dialogue and the Slot Value Generator based on the dialogue history to alleviate the redundant slot value generation.
- We propose two complementary conditions as the base of the judgment, which significantly improves the performance of the slot selection.
- Empirical results show that our model achieves state-of-the-art performance with significant improvements.

2 Related Work

Traditional statistical dialogue state tracking models combine semantics extracted by spoken language understanding modules to predict the current dialogue state (Williams and Young, 2007; Thomson and Young, 2010; Wang and Lemon, 2013;

Williams, 2014) or to jointly learn speech understanding (Henderson et al., 2014c; Zilka and Jurcicek, 2015; Wen et al., 2017). With the recent development of deep learning and representation learning, most works about DST focus on encoding dialogue context with deep neural networks and predicting a value for each possible slot (Xu and Hu, 2018; Zhong et al., 2018; Ren et al., 2018; Xie et al., 2018). For multi-domain DST, slot-value pairs are extended to domain-slot-value pairs for the target (Ramadan et al., 2018; Gao et al., 2019; Wu et al., 2019; Chen et al., 2020b; Hu et al., 2020; Heck et al., 2020; Zhang et al., 2020a). These models greatly improve the performance of DST, but the mechanism of treating slots equally is inefficient and may lead to additional errors. SOM-DST (Kim et al., 2020) considered the dialogue state as an explicit fixed-size memory and proposed a selectively overwriting mechanism. Nevertheless, it arguably has limitations because it lacks the explicit exploration of the relationship between slot selection and local dialogue information.

On the other hand, dialogue state tracking and machine reading comprehension (MRC) have similarities in many aspects (Gao et al., 2020). In MRC task, unanswerable questions are involved, some studies pay attention to this topic with straightforward solutions. (Liu et al., 2018) appended an empty word token to the context and added a simple classification layer to the reader. (Hu et al., 2019) used two types of auxiliary loss to predict plausible answers and the answerability of the question. (Zhang et al., 2020c) proposed a retrospective reader that integrates both sketchy and intensive reading. (Zhang et al., 2020b) proposed a verifier layer to context embedding weighted by start and end distribution over the context words representations concatenated to [CLS] token representation for BERT. The slot selection and the mechanism of local reliability verification in our work are inspired by the answerability prediction in machine reading comprehension.

3 The Proposed Method

Figure 1 illustrates the architecture of DSS-DST. DSS-DST consists of Embedding, Dual Slot Selector, and Slot Value Generator. In the task-oriented dialogue system, given a dialogue $Dial = \{(U_1, R_1); (U_2, R_2) \dots; (U_T, R_T)\}$ of T turns where U_t represents user utterance and R_t represents system response of turn t . We define

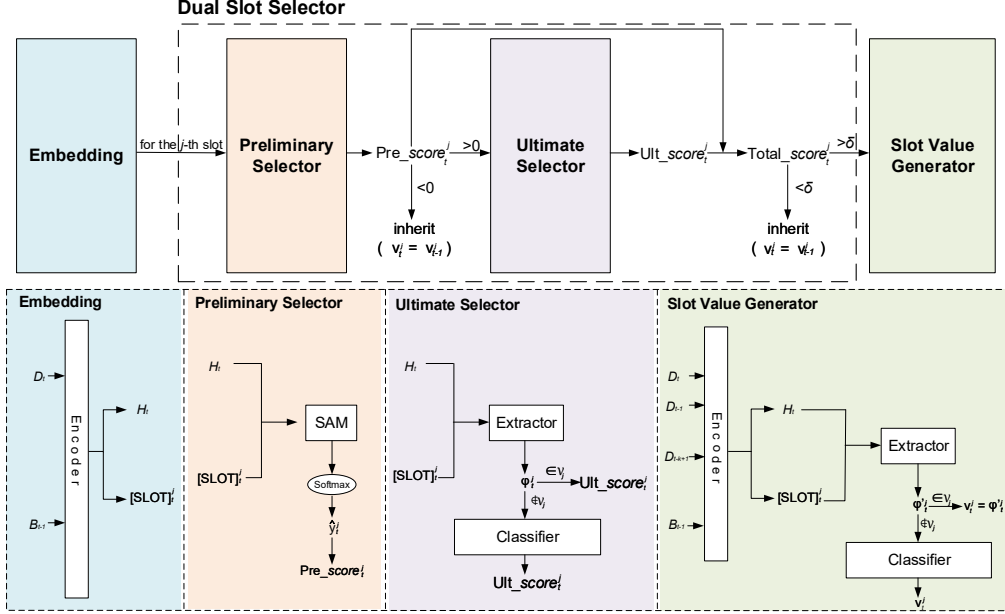


Figure 1: The architecture of the proposed DSS-DST model. The upper part of the figure is the process between each module. The four blocks in the lower part of the figure are the internal structures of the modules with the same color above. At each turn, all slots are judged first, and the slots selected to be updated are permitted to enter the Slot Value Generator to update slot values, while the other slots directly inherit the slot values from the previous turn. The input utterances of the Slot Value Generator are the dialogues of the previous $k - 1$ turns and the current turn, while the Dual Slot Selector only utilizes the current turn dialogue as the input utterances.

the dialogue state at turn t as $\mathcal{B}_t = \{(S^j, V_t^j) | 1 \leq j \leq J\}$, where S^j are the slots, V_t^j are the corresponding slot values, and J is the total number of such slots. Following (Lee et al., 2019), we use the term ‘‘slot’’ to refer to the concatenation of a domain name and a slot name (e.g., ‘‘restaurant – food’’).

3.1 Embedding

We employ the representation of the previous turn dialog state B_{t-1} concatenated to the representation of the current turn dialogue D_t as input:

$$X_t = [\text{CLS}] \oplus D_t \oplus B_{t-1} \quad (1)$$

where [CLS] is a special token added in front of every turn input. Following SOM-DST (Kim et al., 2020), we denote the representation of the dialogue at turn t as $D_t = R_t \oplus U_t \oplus [\text{SEP}]$, where R_t is the system response and U_t is the user utterance. \oplus 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. The representation of the dialogue state at turn t is $B_t = B_t^1 \oplus \dots \oplus B_t^J$, where $B_t^j = [\text{SLOT}]^j \oplus S_j \oplus - \oplus V_t^j$ is the representation of the j -th slot-value pair. $-$ is a special token

used to mark the boundary between a slot and a value. $[\text{SLOT}]^j$ is a special token that represents the aggregation information of the j -th slot-value pair. We feed a pre-trained ALBERT (Lan et al., 2019) encoder with the input X_t . Specifically, the input text is first tokenized into subword tokens. For each token, the input is the sum of the input tokens X_t and the segment id embeddings. For the segment id, we use 0 for the tokens that belong to B_{t-1} and 1 for the tokens that belong to D_t .

The output representation of the encoder is $O_t \in \mathbb{R}^{|X_t| \times d}$, and $h_t^{[\text{CLS}]}, h_t^{[\text{SLOT}]^j} \in \mathbb{R}^d$ are the outputs that correspond to [CLS] and $[\text{SLOT}]^j$, respectively. To obtain the representation of each dialogue and state, we split the O_t into H_t and H_{t-1}^B as the output representations of the dialogue at turn t and the dialogue state at turn $t - 1$.

3.2 Dual Slot Selector

The Dual Slot Selector consists of a Preliminary Selector and an Ultimate Selector, which jointly make a judgment for each slot according to the current turn dialogue.

Slot-Aware Matching Here we first describe the Slot-Aware Matching (SAM) layer, which will be

used as the subsequent components. The slot can be regarded as a special category of questions, so inspired by the previous success of explicit attention matching between passage and question in MRC (Kadlec et al., 2016; Dhingra et al., 2017; Wang et al., 2017; Seo et al., 2016), we feed a representation H and the output representation $h_t^{[\text{SLOT}]^j}$ at turn t to the Slot-Aware Matching layer by taking the slot presentation as the attention to the representation H :

$$\text{SAM}(H, j, t) = \text{softmax}(H(h_t^{[\text{SLOT}]^j})^\top) \quad (2)$$

The output represents the correlation between each position of H and the j -th slot at turn t .

Preliminary Selector The Preliminary Selector briefly touches on the relationship of current turn dialogue utterances and each slot to make an initial judgment. For the j -th slot ($1 \leq j \leq J$) at turn t , we feed its output representation $h_t^{[\text{SLOT}]^j}$ and the dialogue representation H_t to the SAM as follows:

$$\alpha_t^j = \text{SAM}(H_t, j, t) \quad (3)$$

where $\alpha_t^j \in \mathbb{R}^{N \times 1}$ denotes the correlation between each position of the dialogue and the j -th slot at turn t . Then we get the aggregated dialogue representation $H_t^j \in \mathbb{R}^{N \times d}$ and passed it to a fully connected layer to get classification the j -th slot’s logits \hat{y}_t^j composed of selected (logit_sel_t^j) and fail (logit_fai_t^j) elements as follows:

$$H_t^j, m = \alpha_{t,m}^j H_{t,m}, 0 \leq m < N \quad (4)$$

$$\hat{y}_t^j = \text{softmax}(\text{FC}(H_t^j)) \quad (5)$$

We calculate the difference as the Preliminary Selector score for the j -th slot at turn t : $\text{Pre_score}_t^j = \text{logit_sel}_t^j - \text{logit_fai}_t^j$, and define the set of the slot indices as $U_{1,t} = \{j | \text{Pre_score}_t^j > 0\}$, and its size as $J_{1,t} = |U_{1,t}|$. In the next paragraph, the slot in $U_{1,t}$ will be processed as the target object of the Ultimate Selector.

Ultimate Selector The Ultimate Selector will make the judgment on the slots in $U_{1,t}$. The mechanism of the Ultimate Selector is to obtain a temporary slot value for the slot and calculate its reliability through the dialogue at turn t as its confidence for each slot. Specifically, for the j -th slot in $U_{1,t}$ ($1 \leq j \leq J_{1,t}$), we first attempt to obtain the temporary slot value φ_t^j using the extractive method: We employ two different linear layers and feed H_t

as the input to obtain the representation $H_{\text{-}s_t}$ and $H_{\text{-}e_t}$ for predicting the start and end, respectively. Then we feed them to the SAM with the j -th slot to obtain the correlation representation $\alpha_{\text{-}s_t}^j$ and $\alpha_{\text{-}e_t}^j$ as follows:

$$H_{\text{-}s_t} = W_t^s H_t \quad (6)$$

$$H_{\text{-}e_t} = W_t^e H_t \quad (7)$$

$$\alpha_{\text{-}s_t}^j = \text{SAM}(H_{\text{-}s_t}, j, t) \quad (8)$$

$$\alpha_{\text{-}e_t}^j = \text{SAM}(H_{\text{-}e_t}, j, t) \quad (9)$$

The position of the maximum value in $\alpha_{\text{-}s_t}^j$ and $\alpha_{\text{-}e_t}^j$ will be the start and end predictions of φ_t^j :

$$\text{ps}_t^j = \underset{m}{\text{argmax}}(\alpha_{\text{-}s_t}^j, m) \quad (10)$$

$$\text{pe}_t^j = \underset{m}{\text{argmax}}(\alpha_{\text{-}e_t}^j, m) \quad (11)$$

$$\varphi_t^j = \text{Dial}_t[\text{ps}_t^j : \text{pe}_t^j] \quad (12)$$

Here we define \mathcal{V}_j , the candidate value set of the j -th slot. If φ_t^j belongs to \mathcal{V}_j , we calculate its proportion of all possible extracted temporary slot values and calculate the Ult_score_t^j as the score of the j -th slot:

$$\text{logit_span}_t^j = \frac{\exp(\alpha_{\text{-}s_t}^j[\text{ps}_t^j] + \alpha_{\text{-}e_t}^j[\text{pe}_t^j])}{\sum_{p_1=0}^{N-1} \sum_{p_2=p_1+1}^{N-1} \exp(\alpha_{\text{-}s_t}^j[p_1] + \alpha_{\text{-}e_t}^j[p_2])} \quad (13)$$

$$\text{logit_null}_t^j = \frac{\exp(\alpha_{\text{-}s_t}^j[0] + \alpha_{\text{-}e_t}^j[0])}{\sum_{p_1=0}^{N-1} \sum_{p_2=p_1+1}^{N-1} \exp(\alpha_{\text{-}s_t}^j[p_1] + \alpha_{\text{-}e_t}^j[p_2])} \quad (14)$$

$$\text{Ult_score}_t^j = \text{logit_span}_t^j - \text{logit_null}_t^j \quad (15)$$

If φ_t^j does not belong to \mathcal{V}_j , we employ the classification-based method instead to select a temporary slot value from \mathcal{V}_j . Specifically, the dialogue representation H_t^j is passed to a fully connected layer to get the distribution of \mathcal{V}_j . We choose the candidate slot value corresponding to the maximum value as the new temporary slot value φ_t^j , and calculate the distribution probability difference between φ_t^j and “None” as the Ult_score_t^j :

$$\alpha_{\text{-}c_t}^j = \text{softmax}(\text{FC}(H_t^j)) \quad (16)$$

$$\text{max}_c = \underset{m}{\text{argmax}}(\alpha_{\text{-}c_t}^j, m) \quad (17)$$

$$\text{Ult_score}_t^j = \alpha_{\text{-}c_t}^j[\text{max}_c] - \alpha_{\text{-}c_t}^j[0] \quad (18)$$

We choose 0 as index because $\mathcal{V}_j[0] = \text{“None”}$.

Threshold-based decision Following previous studies (Devlin et al., 2019; Yang et al., 2019; Liu et al., 2019; Lan et al., 2019), we adopt the threshold-based decision to make the final judgment for each slot in $U_{1,t}$. The slot-selected threshold δ is set and determined in our model. The total score of the j -th slot is the combination of the predicted Preliminary Selector’s score and the predicted Ultimate Selector’s score:

$$\text{Total_score}_t^j = \beta \text{Pre_score}_t^j + (1 - \beta) \text{Ult_score}_t^j \quad (19)$$

where β is the weight. We define the set of the slot indices as $U_{2,t} = \{j | \text{Total_score}_t^j > \delta\}$, and its size as $J_{2,t} = |U_{2,t}|$. The slot in $U_{2,t}$ will enter the Slot Value Generator to update the slot value.

3.3 Slot Value Generator

After the judgment of the Dual Slot Selector, the slots in $U_{2,t}$ are the final selected slots. For each j -th slot in $U_{2,t}$, the Slot Value Generator generates a value for it. Conversely, the slots that are not in $U_{2,t}$ will inherit the slot value of the previous turn (i.e., $V_t^i = V_{t-1}^i, 1 \leq i \leq J - J_{2,t}$). For the sake of simplicity, we sketch the process as follows because this module utilizes the same hybrid way of the extractive method and the classification-based method as in the Ultimate Selector:

$$X_{-g_t} = [\text{CLS}] \oplus D_t \oplus \dots \oplus D_{t-k+1} \oplus B_{t-1} \quad (20)$$

$$H_{-g_t} = \text{Embedding}(X_{-g_t}) \quad (21)$$

$$\varphi_{-g_t}^j = \text{Ext_method}(H_{-g_t}), 1 \leq j \leq J_{2,t} \quad (22)$$

$$V_t^j = \varphi_{-g_t}^j, \varphi_{-g_t}^j \in \mathcal{V}_j \quad (23)$$

$$V_t^j = \text{Cls_method}(H_{-g_t}), \varphi_{-g_t}^j \notin \mathcal{V}_j \quad (24)$$

Significantly, the biggest difference between the Slot Value Generator and the Ultimate Selector is that the input utterances of the Slot Value Generator are the dialogues of the previous $k - 1$ turns and the current turn, while the Ultimate Selector only utilizes the current turn dialogue as the input utterances.

3.4 Optimization

During training, we optimize both Dual Slot Selector and Slot Value Generator.

Preliminary Selector We use cross-entropy as a training objective:

$$L_{\text{pre},t} = -\frac{1}{J} \sum_{j=1}^J [y_t^j \log \hat{y}_t^j + (1 - y_t^j) \log(1 - \hat{y}_t^j)] \quad (25)$$

where \hat{y}_t^j denotes the prediction and y_t^j is the target indicating whether the slot is selected.

Ultimate Selector The training objectives of both extractive method and classification-based method are defined as cross-entropy loss:

$$L_{\text{ext},t} = -\frac{1}{J_{1,t}} \sum_j^{J_{1,t}} \log(\text{logit}_{-p_t}^j) \quad (26)$$

$$L_{\text{cls},t} = -\frac{1}{J_{1,t}} \sum_j^{J_{1,t}} \sum_i^{|\mathcal{V}_j|} y_{-c_{t,i}}^j \log \alpha_{-c_{t,i}}^j \quad (27)$$

where $\text{logit}_{-p_t}^j$ is the target indicating the proportion of all possible extracted temporary slot values which is calculated according to the form of Equation 13, and $y_{-c_{t,i}}^j$ is the target indicating the probability of candidate values.

Slot Value Generator The training objective $L_{\text{gen},t}$ of this module has the same form of training objective as in the Ultimate Selector.

4 Experimental Setup

4.1 Datasets and Metrics

We choose MultiWOZ 2.0 (Budzianowski et al., 2018), MultiWOZ 2.1 (Eric et al., 2019), and the latest MultiWOZ 2.2 (Zang et al., 2020) as our training and evaluation datasets. These are the three largest publicly available multi-domain task-oriented dialogue datasets, including over 10,000 dialogues, 7 domains, and 35 domain-slot pairs. MultiWOZ 2.1 fixes the previously existing annotation errors. MultiWOZ 2.2 is the latest version of this dataset. It identifies and fixes the annotation errors of dialogue states on MultiWOZ2.1, solves the inconsistency of state updates and the problems of ontology, and redefines the dataset by dividing all slots into two types: non-categorical and categorical. In conclusion, it helps make a fair comparison between different models and will be crucial in the future research of this field.

Following TRADE (Wu et al., 2019), we use five domains for training, validation, and testing, including *restaurant, train, hotel, taxi, attraction*.

These domains contain 30 slots (i.e., $J = 30$). We use joint accuracy and slot accuracy as evaluation metrics. Joint accuracy refers to the accuracy of the dialogue state in each turn. Slot accuracy only considers individual slot-level accuracy.

4.2 Baseline Models

We compare the performance of DSS-DST with the following competitive baselines:

DSTreader formulates the problem of DST as an extractive QA task and extracts the value of the slots from the input as a span (Gao et al., 2019). **TRADE** encodes the whole dialogue context and decodes the value for every slot using a copy-augmented decoder (Wu et al., 2019). **NADST** uses a Transformer-based non-autoregressive decoder to generate the current turn dialogue state (Le et al., 2019). **PIN** integrates an interactive encoder to jointly model the in-turn dependencies and cross-turn dependencies (Chen et al., 2020a). **DS-DST** uses two BERT-base encoders and takes a hybrid approach (Zhang et al., 2020a). **SAS** proposes a Dialogue State Tracker with Slot Attention and Slot Information Sharing to reduce redundant information’s interference (Hu et al., 2020). **SOM-DST** considers the dialogue state as an explicit fixed-size memory and proposes a selectively overwriting mechanism (Kim et al., 2020). **DST-Picklist** performs matchings between candidate values and slot-context encoding by considering all slots as picklist-based slots (Zhang et al., 2020a). **SST** proposes a schema-guided multi-domain dialogue state tracker with graph attention networks (Chen et al., 2020b). **TripPy** extracts all values from the dialog context by three copy mechanisms (Heck et al., 2020).

4.3 Training

We employ a pre-trained ALBERT-large-uncased model (Lan et al., 2019) for the encoder of each part. 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 to 0.03 for the Preliminary Selector and 0.0001 for the Ultimate Selector and the Slot Value Generator, respectively. The max-gradient normalization is utilized and the threshold of gradient clipping is set to 0.1. We use a batch size of 8 and set the dropout (Srivastava et al., 2014) rate to 0.1. In addition, we utilize word dropout (Bowman et al., 2016) by randomly replacing the input tokens with

the special [UNK] token with the probability of 0.1. The max sequence length for all inputs is fixed to 256.

We train the Preliminary Selector for 10 epochs and train the Ultimate Selector and the Slot Value Generator for 30 epochs. During training the Slot Value Generator, we use the ground truth selected slots instead of the predicted ones. We set k to 2, β to 0.55, and δ to 0. For all experiments, we report the mean joint accuracy over 10 different random seeds to reduce statistical errors.

5 Experimental Results

5.1 Main Results

Table 1 shows the joint accuracy and the slot accuracy of our model and other baselines on the test sets of MultiWOZ 2.0, 2.1, and 2.2. As shown in the table, our model achieves state-of-the-art performance on three datasets with joint accuracy of 56.93%, 60.73%, and 58.04%, which has a significant improvement over the previous best joint accuracy. Particularly, the joint accuracy on MultiWOZ 2.1 beyond 60%. Despite the sparsity of experimental result on MultiWOZ 2.2, our model still leads by a large margin in the existing public models. Similar to (Kim et al., 2020), our model achieves higher joint accuracy on MultiWOZ 2.1 than that on MultiWOZ 2.0. For MultiWOZ 2.2, the joint accuracy of categorical slots is higher than that of non-categorical slots. This is because we utilize the hybrid way of the extractive method and the classification-based method to treat categorical slots. However, we can only utilize the extractive method for non-categorical slots since they have no ontology (i.e., candidate value set).

5.2 Ablation Study

Pre-trained Language Model For a fair comparison, we employ different pre-trained language models with different scales as encoders for training and testing on MultiWOZ 2.1 dataset. As shown in Table 2, the joint accuracy of other implemented ALBERT and BERT encoders decreases in varying degrees. In particular, the joint accuracy of BERT-base-uncased decreased by 1.38%, but still outperformed the previous state-of-the-art performance on MultiWOZ 2.1. The result demonstrates the effectiveness of DSS-DST.

Separate Slot Selector To explore the effectiveness of the Preliminary Selector and Ultimate Selector respectively, we conduct an ablation study

Model	MultiWOZ 2.0		MultiWOZ 2.1		MultiWOZ 2.2			
	Joint	Slot	Joint	Slot	Joint	Slot	Cat-joint	Noncat-joint
DSTreader	39.41	-	36.40	-	-	-	-	-
TRADE	48.60	96.92	45.60	-	45.40	-	62.80	66.60
NADST	50.52	-	49.04	-	-	-	-	-
PIN	52.44	97.28	48.40	97.02	-	-	-	-
DS-DST	-	-	51.21	97.35	51.70	-	70.60	70.10
SAS	51.03	97.20	-	-	-	-	-	-
SOM-DST	52.32	-	53.68	-	-	-	-	-
DST-Picklist	54.39	-	53.30	97.40	-	-	-	-
SST	51.17	-	55.23	-	-	-	-	-
TripPy	-	-	55.30	-	-	-	-	-
DSS-DST	56.93 (± 0.43)	97.55 (± 0.05)	60.73 (± 0.51)	98.05 (± 0.06)	58.04 (± 0.49)	97.66 (± 0.06)	76.32 (± 0.27)	73.39 (± 0.32)

Table 1: Joint accuracy (%) and slot accuracy (%) on the test sets of MultiWOZ 2.0, 2.1, and 2.2 vs. various approaches as reported in the literature. Cat-joint and noncat-joint denote joint accuracy on categorical and non-categorical slots, respectively.

Pre-Trained Language Model	MultiWOZ 2.1
Our Model	60.73
BERT (large)	60.11 (-0.62)
ALBERT (base)	59.98 (-0.75)
BERT (base)	59.35 (-1.38)

Table 2: The ablation study of the DSS-DST on the MultiWOZ 2.1 dataset with joint accuracy (%).

Model	MultiWOZ 2.1
Our Model	60.73
-Ultimate Selector	58.82 (-1.91)
-Preliminary Selector	52.22 (-8.51)
-above two	40.69 (-20.04)

Table 3: The ablation study of the DSS-DST on the MultiWOZ 2.1 dataset with joint accuracy (%).

of the two slot selectors on MultiWOZ 2.1. As shown in Table 3, we observe that the performance of the separate Preliminary Selector is better than that of the separate Ultimate Selector. This is presumably because the Preliminary Selector is the head of the Dual Slot Selector, it is stable when it handles all slots. Nevertheless, the input of the Ultimate Selector is the slots selected by the Preliminary Selector, and its function is to make a refined judgment. Therefore, it will be more vulnerable when handling all the slots independently. In addition, when the two selectors are removed, the performance drops drastically. This demonstrates

Model	MultiWOZ 2.1
Our Model	60.73
Dialogue History [†]	58.36 (-2.37)

Table 4: The ablation study of the DSS-DST on the MultiWOZ 2.1 dataset with joint accuracy (%). [†] means attaching the dialogue of the previous turn to the current turn dialogue as the input of the Dual Slot Selector.

k	MultiWOZ 2.1
1	53.96
2 (Our Model)	60.73
3	59.34

Table 5: The joint accuracy (%) of different k on MultiWOZ 2.1 dataset. The k represents the dialogue history of the previous $k - 1$ turns.

that the slot selection is integral before slot value generation.

Dialogue History for the Dual Slot Selector As aforementioned, we consider that the slot selection only depends on the current turn dialogue. In order to verify it, we attach the dialogue of the previous turn to the current turn dialogue as the input of the Dual Slot Selector. We observe in Table 4 that the joint accuracy decreases by 2.37%, which implies the redundant information of dialogue history confuse the slot selection in the current turn.

Our Model		SOM-DST	
Operation	F1	Operation	F1
<i>inherit</i>	99.71	CARRYOVER	98.66
<i>update</i>	90.65	UPDATE	80.10
		DELETE	32.51
		DONTCARE	2.86

Table 6: Statistics of the state operations and the corresponding F1 scores of our model and SOM-DST in the test set of MultiWOZ 2.1.

MultiWOZ 2.2	
Domain	Joint Accuracy (%)
Attraction	79.88
Hotel	62.47
Restaurant	75.79
Taxi	54.84
Train	76.25

Table 7: Domain-specific results on the test set of MultiWOZ 2.2. We are the first to list Domain-specific results on the test set of MultiWOZ 2.2 to the best of our knowledge.

Dialogue History for the Slot Value Generator

We try the number from one to three for the k to observe the influence of the selected dialogue history on the Slot Value Generator. As shown in Table 5, the model achieves better performance on MultiWOZ 2.1 when $k = 2, 3$ than that of $k = 1$. Furtherly, the performance of $k = 2$ is better than that of $k = 3$. We conjecture that the dialogue history far away from the current turn is little helpful because the relevance between two sentences in dialogue is strongly related to their positions.

The above ablation studies show that dialogue history confuses the Dual Slot Selector, but it plays a crucial role in the Slot Value Generator. This demonstrates that there are fundamental differences between the two processes, and confirms the necessity of dividing DST into these two sub-tasks.

6 Analysis

6.1 Comparative Analysis of Slot Selector

We analyze the performance of the Dual Slot Selector and compare it with other previous work in MultiWOZ 2.1. Here we choose the SOM-DST and list the state operations and the corresponding F1 scores as a comparison. The SOM-DST sets four state operations (i.e., CARRYOVER, DELETE, DONTCARE, UPDATE), while our model classifies the slots into two classes (i.e., *inherit* and

Model	MultiWOZ 2.2	
	Joint	Cat-joint
Our Model	58.04	76.32
-Extractive Method	50.01	66.15

Table 8: The ablation study of the DSS-DST on the MultiWOZ 2.2 dataset with joint accuracy (%) and joint accuracy on categorical slots.

update). It means that DELETE, DONTCARE, and UPDATE in SOM-DST all correspond to *update* in our model. As shown in Table 6, our model still achieves superior performance when dealing with *update* slots, which contain DONTCARE, DELETE, and other difficult cases.

6.2 Domains and Ontology

Table 7 shows the domain-specific results of our model on the latest MultiWOZ 2.2 dataset. We can observe that the performance of our model in *taxi* domain is lower than that of the other four domains. We investigate the dataset and find that all the slots in *taxi* domain are non-categorical slots. This indicates the reason that we can only utilize the extractive method for non-categorical slots since they have no ontology. Furthermore, we test the performance of using the separate classification-based method for categorical slots. As illustrated in Table 8, the joint accuracy of our model and categorical slots decreased by 8.03% and 10.17%, respectively.

7 Conclusion

We introduce an effective two-stage DSS-DST which consists of the Dual Slot Selector based on the current turn dialogue, and the Slot Value Generator based on the dialogue history. The Dual Slot Selector determines each slot whether to update or to inherit based on the two conditions. The Slot Value Generator employs a hybrid method to generate new values for the slots selected to be updated according to the dialogue history. Our model achieves state-of-the-art performance of 56.93%, 60.73%, and 58.04% joint accuracy with significant improvements (+2.54%, +5.43%, and +6.34%) over previous best results on MultiWOZ 2.0, MultiWOZ 2.1, and MultiWOZ 2.2 datasets, respectively. The mechanism of a hybrid method is a promising research direction and we will exploit a more comprehensive and efficient hybrid method for slot value generation in the future.

Acknowledgements

This work was supported by the National key research and development project (2017YFB1400603) and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (Grant No. 61921003). We thank the anonymous reviewers for their insightful comments.

Ethical Considerations

The claims in this paper match the experimental results. The model utilizes the hybrid method for slot value generation, so it is universal and scalable to unseen domains, slots, and values. The experimental results can be expected to generalize.

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.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz—a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.
- Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2020a. Parallel interactive networks for multi-domain dialogue state generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1921–1931.
- Lu Chen, Boer Lv, Chi Wang, Su Zhu, Bowen Tan, and Kai Yu. 2020b. 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.
- Bhuvan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1832–1846.
- 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*.
- 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.
- Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialogue state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGDial Meeting on Discourse and Dialogue*, pages 264–273.
- Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geischauser, 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. 2014a. 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.
- Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014b. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329. IEEE.
- Matthew Henderson, Blaise Thomson, and Steve Young. 2014c. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299.
- Jiaying Hu, Yan Yang, Chencai Chen, Zhou Yu, et al. 2020. Sas: Dialogue state tracking via slot attention and slot information sharing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6366–6375.
- Minghao Hu, Furu Wei, Yuxing Peng, Zhen Huang, Nan Yang, and Dongsheng Li. 2019. Read+ verify: Machine reading comprehension with unanswerable questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6529–6537.
- Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the*

- 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 908–918.
- 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*.
- Hung Le, Richard Socher, and Steven CH Hoi. 2019. Non-autoregressive dialog state tracking. In *International Conference on Learning Representations*.
- Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. Sumbt: Slot-utterance matching for universal and scalable belief tracking. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5478–5483.
- Xiaodong Liu, Wei Li, Yuwei Fang, Aerin Kim, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for squad 2.0. *arXiv preprint arXiv:1809.09194*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.
- Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.
- Jun Quan and Deyi Xiong. 2020. Modeling long context for task-oriented dialogue state generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7119–7124.
- Osman Ramadan, Paweł Budzianowski, and Milica Gasic. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 432–437.
- Liliang Ren, Kaige Xie, Lu Chen, and Kai Yu. 2018. Towards universal dialogue state tracking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2780–2786.
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension.
- 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.
- Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198.
- 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.
- Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve J Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *EACL (1)*.
- Jason D Williams. 2014. Web-style ranking and slu combination for dialog state tracking. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 282–291.
- Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The dialog state tracking challenge series. *AI Magazine*, 35(4):121–124.
- Jason D Williams and Steve Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.
- 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*.
- Kaige Xie, Cheng Chang, Liliang Ren, Lu Chen, and Kai Yu. 2018. Cost-sensitive active learning for dialogue state tracking. In *Proceedings of the 19th Annual SIGDial Meeting on Discourse and Dialogue*, pages 209–213.
- 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.

- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763.
- 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. 2020a. 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.
- Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, and Rui Wang. 2020b. Sg-net: Syntax-guided machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9636–9643.
- Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2020c. Retrospective reader for machine reading comprehension. *arXiv preprint arXiv:2001.09694*.
- Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1458–1467.
- Lukas Zilka and Filip Jurcicek. 2015. Incremental lstm-based dialog state tracker. In *2015 Ieee Workshop on Automatic Speech Recognition and Understanding (Asru)*, pages 757–762. IEEE.

Appendices

A Accuracy per Slot on MultiWOZ 2.2 Testset

Domain-Slot	Our Model
attraction-area	97.95
attraction-name	93.38
attraction-type	97.37
hotel-area	97.29
hotel-book day	100
hotel-book people	100
hotel-book stay	100
hotel-internet	94.94
hotel-name	95.29
hotel-parking	95.26
hotel-price range	97.67
hotel-stars	97.98
hotel-type	93.24
restaurant-area	97.34
restaurant-book day	100
restaurant-book people	100
restaurant-book time	100
restaurant-food	96.76
restaurant-name	94.26
restaurant-price range	97.88
taxi-arrive by	98.68
taxi-departure	97.24
taxi-destination	97.05
taxi-leave at	99.25
train-arrive by	96.63
train-book people	100
train-day	99.59
train-departure	98.32
train-destination	98.48
train-leave at	94.14

Table 9: The detailed results of accuracy (%) per slot on MultiWOZ 2.2 test set. We sort them according to their domains.

B Data Statistics

Domain	Slots	Dialogues			Turns		
		Train	Valid	Test	Train	Valid	Test
Hotel	price range, type, parking, book stay, book day, book people, area, stars, internet, name	3,381	416	394	14,793	1,781	1,756
Attraction	area, name, type	2,717	401	395	8,073	1,220	1,256
Restaurant	food, price range, area, name, book time, book day, book people	3,813	438	437	15,367	1,708	1,726
Taxi	leave at, destination, departure, arrive by	1,654	207	195	4,618	690	654
Train	destination, day, departure, arrive by, book people, leave at	3,103	484	494	12,133	1,972	1,976

Table 10: Data statistics of MultiWOZ 2.1.