# Dialogue State Tracking with Sparse Local Slot Attention

**Longfei Yang**[1], **Jiyi Li**[2], **Sheng Li**[3], **Takahiro Shinozaki**[1]
[1]Tokyo Institute of Technology
[2]University of Yamanashi
[3]National Institute of Information and Communications Technology
longfei.yang.cs@gmail.com, jyli@yamanashi.ac.jp, sheng.li@nict.go.jp,
shinot@ict.e.titech.ac.jp

## Abstract

Dialogue state tracking (DST) is designed to track the dialogue state during the conversations between users and systems, which is the core of task-oriented dialogue systems. Mainstream models predict the values for each slot with fully token-wise slot attention from dialogue history. However, such operations may result in overlooking the neighboring relationship. Moreover, it may lead the model to assign probability mass to irrelevant parts, while these parts contribute little. It becomes severe with the increase in dialogue length. Therefore, we investigate sparse local slot attention for DST in this work. Slot-specific local semantic information is obtained at a sub-sampled temporal resolution capturing local dependencies for each slot. Then these local representations are attended with sparse attention weights to guide the model to pay attention to relevant parts of local information for subsequent state value prediction. The experimental results on MultiWOZ 2.0 and 2.4 datasets show that the proposed approach effectively improves the performance of ontology-based dialogue state tracking, and performs better than token-wise attention for long dialogues.

## 1 Introduction

Task-oriented dialogue systems aim to assist users to complete certain tasks and have drawn great attention in both academia and industry (Young et al., 2010, 2013; Chen et al., 2017). As the core of task-oriented dialogue systems, dialogue state tracking (DST) is designed to track the dialogue states during the conversation between users and systems, which is generally expressed as a list of {*(domain, slot, value)*} representing user's goal (Rastogi et al., 2017, 2018). The estimated dialogue states are used for subsequent actions.

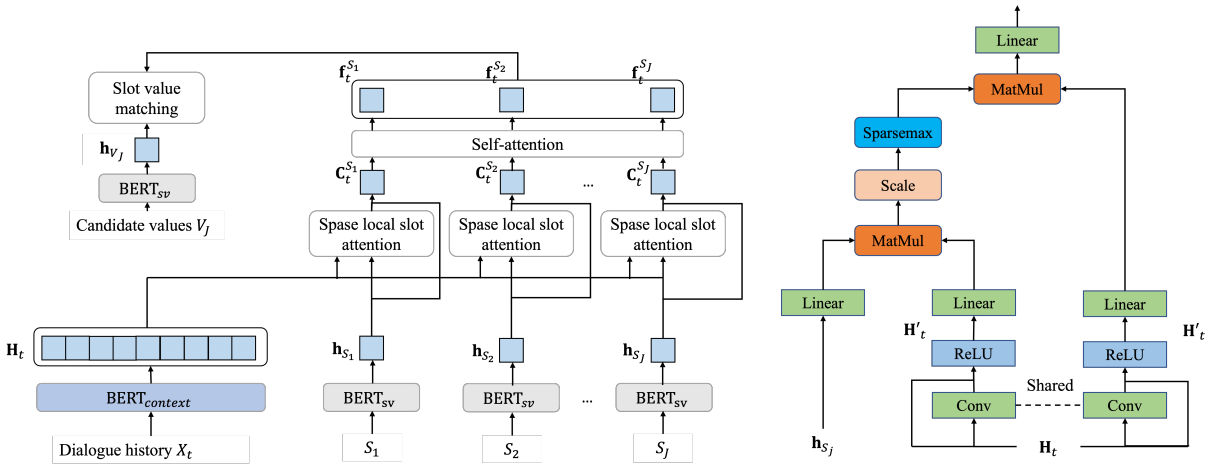
To achieve the dialogue state, value prediction is made for each slot given the dialogue history. At each turn, the model inquires of the dialogue history and predicts the state values accordingly (Xu and Hu, 2018; Ren et al., 2018; Wu et al., 2019; Zhang et al., 2019; Heck et al., 2020). With it, how to extract appropriate context information in the noisy dialogue history is crucial and challenging (Hu et al., 2020). Yang et al. (2021) make an empirical study about the effect of different contexts on the performance of DST with several manually designed rules. It indicates that the performance of DST models benefits from selecting appropriate context granularity.

In recent mainstream models, a fully token-wise slot attention mechanism is widely used to capture slot-specific information with dialogue history. The attention assigns an attention weight to each token, measuring the relationship of each token in dialogue history for the specified slot, and then attends them with these weights. Although encouraging results have been achieved, it also brings some risks. First, such operations disperse the distribution of attention, which results in overlooking the neighboring relation (Yang et al., 2018). Some entities (e.g., restaurant and attraction names) in spoken dialogue are generally informal, diverse, and local-compact, where the non-semantic tokens may be included. Moreover, a limitation of the used softmax computation is that the probability distribution in the outputs always has full support (Martins and Astudillo, 2016), i.e., $\text{softmax}(\mathbf{z}) > 0$ for every vector $\mathbf{z}$. It may lead a model to assign probability mass to implausible parts of dialogue history. Involving noise may make the model difficult to focus on the essential parts, and it may be more severe with the increase in dialogue length (Peters et al., 2019).

To tackle this problem, we propose a sparse local slot attention mechanism for this task. In our approach, local semantic information is firstly achieved at a sub-sampled temporal resolution capturing local dependencies for each slot. Then, these local information is attended with sparse attention weights generated by sparsemax function (Martins



*Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 39–46




(a) The entire framework of our model.

(b) Sparse local slot attention.

Figure 1: A demonstration of our model: (a) the entire framework, (b) the proposed sparse local slot attention.

and Astudillo, 2016), which outputs sparse posterior distributions by assigning zero probability to irrelevant contents in the dialogue history.

We conduct experiments to verify our approach on MultiWOZ 2.0 and 2.4 datasets. The contributions can be addressed as follows: 1) We propose a sparse local slot attention mechanism to lead the model to focus on relevant local parts to the specific slot for the DST task; 2) We demonstrate that the performance of DST benefits from introducing local information with our proposed approach, and make an empirical study that shows that our model performs better in state prediction for name-related slots and long dialogues than the models based on fully token-wise attention.

## 2 Related Works

Dialogue state tracking (DST) is the core of task-oriented dialogue systems. In the early years, DST highly relies on hand-crafted semantic features to predict the dialogue states (Williams and Young, 2007; Thomson and Young, 2010; Wang and Lemon, 2013), which is hard to handle lexical and morphological variations in spoken language (Lee et al., 2019). Benefiting from the rapid development of deep learning methods and their successful application in natural language processing, neural method-based DST models have been proposed. (Mrkšić et al., 2017) proposes a novel neural belief tracking (NBT) framework with learning n-gram representation of the utterance. Inspired by it, sorts of neural network-based models have been investigated for DST task (Nouri and Hosseini-Asl, 2018; Ren et al., 2018; Zhong et al., 2018; Hu et al., 2020; Ouyang et al., 2020; Wu et al., 2019) and achieves encouraging results.

Pre-trained models have brought natural language processing to a new era in recent years. Many substantial works have shown that the pre-trained models can learn universal language representations, which are beneficial for downstream tasks (Mikolov et al., 2013; Pennington et al., 2014; McCann et al., 2017; Sarzynska-Wawer et al., 2021; Devlin et al., 2019). More recently, very deep pre-trained language models, such as bidirectional encoder representation from the transformer (BERT) (Devlin et al., 2019) and generative pre-training (GPT) (Radford et al., 2018), trained with an increasing number of self-supervised tasks have been proposed to make the models capturing more knowledge from a large scale of corpora, which have shown their abilities to produce promising results in downstream tasks. In view of it, many pieces of research of DST have explored to establish the models on the basis of pre-trained language models (Hosseini-Asl et al., 2020; Kim et al., 2020; Lee et al., 2019; Zhang et al., 2019; Chen et al., 2020; Chao and Lane, 2019; Ye et al., 2021b; Heck et al., 2020; Lin et al., 2020).

Related to extracting slot-specific information, most of the previous studies rely on dense token-wise attention (Lee et al., 2019; Wang et al., 2020; Ye et al., 2021b). However, several pieces of research have indicated that local information may be missing with it (Yang et al., 2018; Shaw et al., 2018; Sperber et al., 2018; Luong et al., 2015; Yang et al., 2022). Motivated by it, we investigate introducing local modeling in this task. The most rele-

vant research is (Yang et al., 2021), which makes a comprehensive study of how different granularities affect DST. However, this research employs simple hand-crafted rules to neglect several utterances in a dialogue history. Our proposed approach in this work is data-driven.

## 3 Dialogue State Tracking with Sparse Local Slot Attention

### 3.1 Encoding

As shown in Figure 1(a), BERT$_{context}$ is used for encoding the dialogue context, whose parameters are fine-tuned during training. Let's define the dialogue history $D_T = \{R_1, U_1, ..., R_T, U_T\}$ as a set of system responses $R$ and user utterances $U$ in $T$ turns of dialogue, where $R = \{R_t\}_{t=1}^{T}$ and $U = \{U_t\}_{t=1}^{T}$. We define $E_T = \{B_1, ..., B_T\}$ as the dialogue states of $T$ turns, and each $E_t$ is a set of slot value pairs $\{(S_1, V_1), ..., (S_J, V_J)\}$ of $J$ slots. The context encoder accepts the dialogue history till turn $t$, which can be denoted as $X_t = \{D_t, E'_{t-1}\}$, as the input and generates context vector representations $\mathbf{H}_t = \text{BERT}_{context}(X_t)$.

Another pre-trained BERT$_{sv}$ is employed to encode the slots and candidate values. Its parameters remain frozen during training. For those slots and values containing multiple tokens, the vector corresponding to the [CLS] token is employed to represent them. For each slot $S_j$ and value $V_j$, $\mathbf{h}_{S_j} = \text{BERT}_{sv}(S_j)$, $\mathbf{h}_{V_j} = \text{BERT}_{sv}(V_j)$.

### 3.2 Sparse Local Slot Attention

To extract slot-specific information, we propose sparse local slot attention (SLSA). As shown in Figure 1(b), sparse local slot attention accepts the dialogue history $\mathbf{H}_t$ and the representation $\mathbf{h}_{S_j}$ of the specific slot $S_j$. To obtain local information, we employ a convolutional layer whose kernel has size $l$ and stride $m$ over the context vector representation of dialogue history. The convolutional kernel accepts the local area in the dialogue history representation and multiplies it with the learnable parameters to obtain the local semantic representations.

$$\mathbf{H}'_t = ReLU(Conv(\mathbf{H}_t) + \mathbf{H}_t) \tag{1}$$

After that, multi-head attention with the sparsemax function is employed to retrieve relevant information for each slot. It generates sparse distribution to each local area. The sparsemax function returns the Euclidean projection of the input vector $\mathbf{z}$ onto the probability simplex $\Delta^{K-1} := \{\mathbf{p} \in \mathbb{R}^K | \mathbf{1}^T\mathbf{p} = 1, \mathbf{p} \geq 0\}$. The projection is likely to hit the boundary of the simplex, in which case $sparsemax(\mathbf{z})$ becomes sparse (Martins and Astudillo, 2016).

$$Sparsemax(\mathbf{z}) := \arg\min_{\mathbf{p}\in\Delta^{K-1}} ||\mathbf{p} - \mathbf{z}||^2 \tag{2}$$

Then the output is concatenated with each slot to generate slot-specific representations through a feed-forward layer.

$$\mathbf{Q}_t^{S_j} = \mathbf{h}_{S_j}\mathbf{W}_Q + \mathbf{b}_Q \tag{3}$$

$$\mathbf{K}_t^{S_j} = \mathbf{H}'_t\mathbf{W}_K + \mathbf{b}_K \tag{4}$$

$$\mathbf{V}_t^{S_j} = \mathbf{H}'_t\mathbf{W}_V + \mathbf{b}_V \tag{5}$$

$$\boldsymbol{\alpha}_t^{S_j} = Sparsemax(\frac{\mathbf{Q}_t^{S_j}{\mathbf{K}_t^{S_j}}^T}{\sqrt{d_k}})\mathbf{V}_t^{S_j} \tag{6}$$

$$\mathbf{C}_t^{S_j} = \mathbf{W}_2 ReLU(\mathbf{W}_1[\mathbf{h}_{S_j}, \boldsymbol{\alpha}_t^{S_j}] + \mathbf{b}_1) + \mathbf{b}_2 \tag{7}$$

Where $\mathbf{W}_Q, \mathbf{b}_Q, \mathbf{W}_K, \mathbf{b}_K$, $\mathbf{W}_V$, and $\mathbf{b}_V$ are the parameters of the linear layers for projecting query, key, and value respectively. $d_k = d_h/N$ in which $d_h$ is the hidden size of the model, and $N$ is the number of heads.

### 3.3 Slot Self-Attention

Slot self-attention is introduced to communicate information across different slots. Each sub-layer in the self-attention layer consists of the self-attention block and two fully connected layers of ReLU activation with layer normalization and residual connection. Let $\mathbf{C}_t = [\mathbf{C}_t^{S_1}, ..., \mathbf{C}_t^{S_J}]$ and $\mathbf{F}_t^1 = \mathbf{C}_t$ at the first sub layer, then for the $l$-th sub-layer,

$$\widetilde{\mathbf{F}}_t^l = LayerNorm(\mathbf{F}_t^l), \tag{8}$$

$$\mathbf{G}_t^l = MultiHead(\widetilde{\mathbf{F}}_t^l, \widetilde{\mathbf{F}}_t^l, \widetilde{\mathbf{F}}_t^l) + \widetilde{\mathbf{F}}_t^l. \tag{9}$$

For the $l$-th feed forward sub-layer,

$$\widetilde{\mathbf{G}}_t^l = LayerNorm(\mathbf{G}_t^l), \tag{10}$$

$$\mathbf{F}_t^{l+1} = FFN(\widetilde{\mathbf{G}}_t^l) + \widetilde{\mathbf{G}}_t^l. \tag{11}$$

The output of the final layer is regarded as the final slot specific vector $\mathbf{F}_t^{L+1} = [\mathbf{f}_t^{S_1}, ..., \mathbf{f}_t^{S_J}]$.

### 3.4 Slot Value Matching

A Euclidean distance-based value prediction is performed for each slot, the nearest value is chosen to predict the state value.

$$p(V_t^j|X_t, S_j) = \frac{\exp(-d(\mathbf{h}^{V_j}, \mathbf{f}_t^{S_j}))}{\sum_{V_j' \in \nu_j} \exp(-d(\mathbf{h}^{V_j'}, \mathbf{f}_t^{S_j}))} \tag{12}$$

where $d(\cdot)$ is Euclidean distance function, and $\nu_j$ denotes the value space of the slot $S_j$. The model is trained to maximize the joint probability of all slots. The loss function at each turn $t$ is denoted as the sum of the negative log-likelihood, $\mathcal{L}_t = \sum_{j=1}^{J} -\log(p(V_t^j|X_t, S_j))$.

## 4 Experiments

### 4.1 Datasets

We conduct experiments using a couple of mainstream datasets of task-oriented dialogue: MultiWOZ 2.0 and 2.4 datasets. MultiWOZ2.0 (Budzianowski et al., 2018) is currently the largest open-source human-human conversational dataset of multiple domains. MultiWOZ 2.4 is the latest version and fixes the incorrect and inconsistent annotations (Ye et al., 2021a).

### 4.2 Implementation Details

The $\text{BERT}_{context}$ is a pre-trained BERT-base-uncased model, which has 12 layers with 768 hidden units and 12 self-attention heads. Another BERT-base-uncased model is used as the $\text{BERT}_{sv}$. For the sparse local slot attention, window size and stride are investigated in the experiment. Padding is added on both sides of the input if needed. The number of attention heads is 4. Adam optimizer is adopted with a batch size of 16, which trains the model with a learning rate of 4e-5 for the encoder and 1e-4 for other parts. The hyper-parameters are selected from the best-performing model over the validation set. We use a dropout with a probability of 0.1 on the dialogue history during training.

### 4.3 Main Results

The main results are shown in Table 1. As we can see, our model achieves the best performance on all the datasets. We utilize the Wilcoxon signed-rank test, the proposed method is statistically significantly better ($p < 0.05$) than baselines. For the MultiWOZ 2.0 dataset, our proposed SLSA model (window size is 3 and stride is 1) achieves a JGA of 54.83% performing better than STAR with a JGA of 54.53%, which is the previous SOTA. Moreover, on the latest refined version MultiWOZ 2.4 fixing many annotations in the test set, our model obtains a JGA of 77.92%. To sum up, our proposed model achieves a slight improvement on the original MultiWOZ 2.0 dataset, and a significant improvement on the latest refined MultiWOZ 2.4 dataset with a clean test set. We also make an investigation about the effects of local granularities, as shown in Appendix A.1.

Table 1: The joint goal accuracy (JGA) of different models. SLSA denotes our proposed sparse local slot attention.

| Model | MW2.0 | MW2.4 |
|---|---|---|
| TRADE (Wu et al., 2019) | 48.93 | 54.97 |
| SOM (Kim et al., 2020) | 51.72 | 66.78 |
| TripPy (Heck et al., 2020) | - | 59.62 |
| SimpleTOD (Hosseini-Asl et al., 2020) | - | 66.78 |
| SUMBT (Lee et al., 2019) | 46.65 | 61.86 |
| DS-DST (Zhang et al., 2019) | 52.24 | - |
| DS-Picklist (Zhang et al., 2019) | 54.39 | - |
| SAVN (Wang et al., 2020) | 54.52 | 60.55 |
| SST (Chen et al., 2020) | 51.17 | - |
| STAR (Ye et al., 2021b) | 54.53 | 73.62 |
| SLSA | **54.83** | **77.92** |

Table 2: The results on the MultiWOZ 2.4 dataset using our model with different settings.

| | JGA (%) | SA (%) |
|---|---|---|
| SLSA | 77.92 | 99.06 |
| w/o Sparse | 75.79 | 98.96 |
| w/o Local | 74.65 | 98.89 |
| w/o Both | 73.88 | 98.84 |

### 4.4 Ablation Study

To further verify the proposed approach, we present some results that show the effectiveness of the components in the proposed approaches. Table 2 presents the joint goal accuracy and slot accuracy obtained when we progressively remove the components in our proposed model on MultiWOZ 2.4 dataset. On one hand, comparing SLSA and "w/o Local" (or "w/o Sparse" and "w/o both"), when the local pattern component is removed, the performance of corresponding model decreases. On the other hand, comparing SLSA and "w/o Sparse" (or "w/o Local" and "w/o both" when the sparse component is removed, the performance of the corresponding model decreases. It shows that the sparse and the local components are effective and important to the proposed model.

### 4.5 Error Analysis

An error analysis of each slot for the previous SOTA model STAR and our models on MultiWOZ 2.4 is shown in Figure 2, in which the lower the better. The four slots with the highest error rates

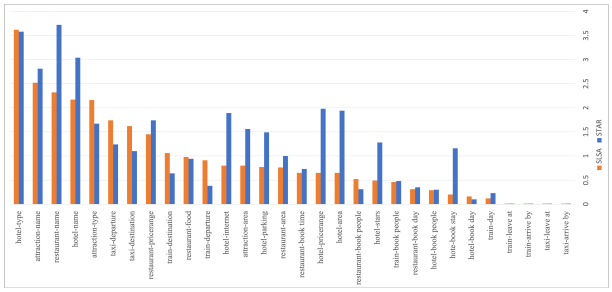



Figure 2: The error rate per slot of STAR and our models on MultiWOZ 2.4 dataset.

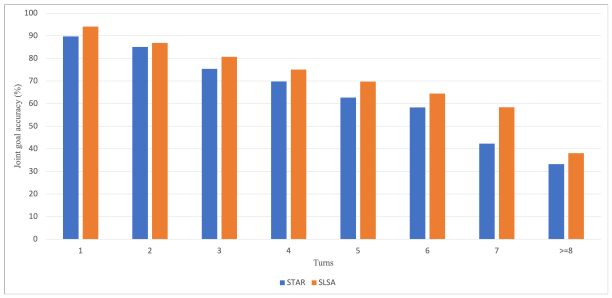



Figure 3: Joint goal accuracy per turn of STAR and our models on MultiWOZ 2.4 dataset.

are *hotel-type* with 3.62%, *attraction-name* with 2.52%, *restaurant-name* with 2.32% and *hotel-name* with 2.17%. It can be noticed that the later three are *name*-related whose values are diverse, local-compact, and may includes several non-semantic tokens. Our proposed models perform better than STAR on these three slots, evidenced by that the error rates are lower. In addition, our model performs better in several categorical slots such as *hotel-internet*, *hotel-parking*, *hotel-stars* and *book stay*. We make a case study shown in Appendix A.2 to have a straightforward understanding of our proposed approach.

### 4.6 Performance for Long Dialogues

Figure 3 depicts the joint goal accuracy per turn of our models and STAR on MultiWOZ 2.4 dataset. Joint goal accuracy per turn is to measure the performance for long dialogues. It is considered correct if and only all of the values are correctly predicted for each slot until the $n$-th turn. In the beginning, the performance of these two models for short turns is comparable. Then it decreases as the dialogue length becomes longer since the previous states are employed as part of the input where some mistakes may be included. The trend of our model is a little milder. For very long dialogues whose length is larger than 7, our model performs better than STAR. It shows our model performs better for the long dialogues DST.

## 5 Conclusion

In his work, we propose a sparse local slot attention for dialogue state tracking to alleviate allocating attention weights to content unrelated to the specific slot of interest. In our approach, local semantic information is firstly achieved at a sub-sampled temporal resolution capturing local dependencies for each slot. Then, these local information is attended with sparse attention weights generated by sparsemax function. The experimental results show that, comparing to several existing models based on dense token-wise attention, our approach effectively improves the performance of ontology-based dialogue state tracking in the state prediction for name-related slots and long dialogues.

## Acknowledgement

This work was supported by JSPS KAKENHI Grand Number JP22K12069 and partially supported by JSPS KAKENHI Grant Number 23K11227 and 23H03402.

## Limitations

In this work, we propose a sparse local slot attention (SLSA) mechanism to make the model pay attention to slot-specified local areas in dialogue history, and then attend them with sparse distribution generated by sparsemax to neglect some redundant parts. This paper shows the effectiveness of our proposed approaches in state prediction for some specified slots and long dialogues. While we show that the model with SLSA is competitive in dialogue state tracking, there are limitation of that provide avenues for future works. First, it is not as easy to apply SLSA to generation-based dialogue state tracking. Different from ontology-based manners, the condition may be different in the case of generative DST since entire successive information involved in language modeling may be important for language generation. Therefore, how to handle the local and sparse properties for the generative model need to further consider. Second, convolution operation considers a fixed bounded local context. It is a challenge to handle local properties of various lengths.

## References


Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

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*.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

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.

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*, pages 4171–4186.

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.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A Simple Language Model for Task-Oriented Dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191.

Jiaying Hu, Yan Yang, Chencai Chen, Liang He, and Zhou Yu. 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.

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.

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.

Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. 2020. Mintl: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3391–3405.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Andre Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International conference on machine learning*, pages 1614–1623.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *arXiv preprint arXiv:1708.00107*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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*, pages 1777–1788.

Elnaz Nouri and Ehsan Hosseini-Asl. 2018. Toward scalable neural dialogue state tracking. In *NeurIPS 2018, 2nd Conversational AI workshop*.

Yawen Ouyang, Moxin Chen, Xinyu Dai, Yinggong Zhao, Shujian Huang, and Jiajun Chen. 2020. Dialogue state tracking with explicit slot connection modeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 34–40.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Ben Peters, Vlad Niculae, and André FT Martins. 2019. Sparse sequence-to-sequence models. *arXiv preprint arXiv:1905.05702*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Abhinav Rastogi, Raghav Gupta, and Dilek Hakkani-Tur. 2018. Multi-task learning for joint language understanding and dialogue state tracking. *arXiv preprint arXiv:1811.05408*.

Abhinav Rastogi, Dilek Hakkani-Tür, and Larry Heck. 2017. Scalable multi-domain dialogue state tracking. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 561–568.

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.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 464–468.

Matthias Sperber, Jan Niehues, Graham Neubig, Sebastian Stüker, and Alex Waibel. 2018. Self-attentional acoustic models. *arXiv preprint arXiv:1803.09519*.

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.

Yexiang Wang, Yi Guo, and Siqi Zhu. 2020. Slot attention with value normalization for multi-domain dialogue state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3019–3028.

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.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819.

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*, pages 1448–1457.

Baosong Yang, Zhaopeng Tu, Derek F Wong, Fandong Meng, Lidia S Chao, and Tong Zhang. 2018. Modeling localness for self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4449–4458.

Longfei Yang, Jiyi Li, Sheng Li, and Takahiro Shinozaki. 2022. Multi-domain dialogue state tracking with top-k slot self attention. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 231–236.

Puhai Yang, Heyan Huang, and Xian-Ling Mao. 2021. Comprehensive study: How the context information of different granularity affects dialogue state tracking? *arXiv preprint arXiv:2105.03571*.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021a. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv:2104.00773*.

Fanghua Ye, Jarana Manotumruksa, Qiang Zhang, Shenghui Li, and Emine Yilmaz. 2021b. Slot self-attentive dialogue state tracking. In *Proceedings of the Web Conference 2021*, pages 1598–1608.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Jian-Guo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wan, Philip S Yu, Richard Socher, and Caiming Xiong. 2019. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. *arXiv preprint arXiv:1910.03544*.

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*, pages 1458–1467.

Table 3: The results with different sizes $l$s and strides $m$s of the local window in our model.

| Setup | SLSA$_{conv}$ |
|---|---|
| $l = 1, m = 1$ | 74.82 |
| $l = 3, m = 1$ | 77.92 |
| $l = 3, m = 2$ | 76.45 |
| $l = 3, m = 3$ | 76.87 |
| $l = 5, m = 1$ | 74.89 |
| $l = 5, m = 3$ | 74.46 |
| $l = 5, m = 5$ | 73.44 |

# A Appendix

## A.1 Effects of Different Locality Granularities

We compare our model with different sizes and strides of the window of the local pattern to see how different granularities affect the performance on MultiWOZ 2.4 dataset, as shown in Table 3.

It shows that the best result is achieved when the size of 3 and the stride of 1, while the performance is not improved by enlarging the size of the local window or decreasing it. Note that, as mentioned in the experimental settings, in the main results, the hyperparameters of window size and stride are selected by tuning on the validation set.

## A.2 Case Study

Figure 4 and 5 demonstrate the predicted states of STAR and our model on two pieces of dialogues from the MultiWOZ 2.4 dataset. We color the input with the weights generated by sparse local slot attention in our model and the dense token-wise attention used in STAR. Note that in our model, one position with a dark background means the local area around this position is focused. It is different from STAR, in which one position denotes a token.

As shown in Figure 4, although STAR captures the relevant information for *attraction-name* but not the best. Our models are able to focus on the local area covering the entity. As shown in Figure 5, the user says "nothing in particular" indicating he/him does not prefer "a certain area". STAR fails to capture this information, and its attention is scattered. Our model realizes this and successfully gets the user's point. Although the values "none" and "do not care" indicate the *attraction-area* does not need concrete values, they denote the user's different intentions.

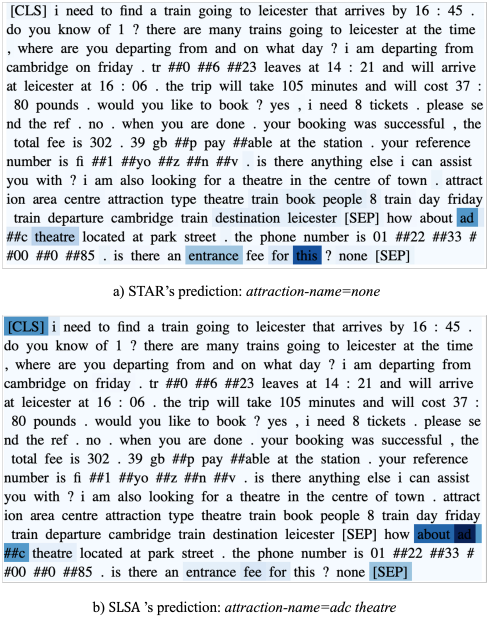



Figure 4: The predicted dialogue states for slot $attraction - name$ with STAR and our model on dialogue PMUL1424.

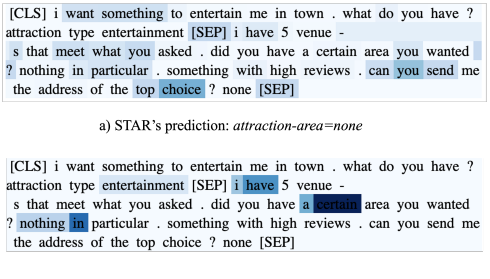



Figure 5: The predicted dialogue states for slot $attraction - area$ with STAR and our model on dialogue PMUL2415.