
Proceedings Paper:
Wang, D., Lin, C. orcid.org/0000-0003-3454-2468, Liu, Q. et al. (1 more author) (2021) Fast and scalable dialogue state tracking with explicit modular decomposition. In: Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T. and Zhou, Y., (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 06-11 Jun 2021, Virtual conference. Association for Computational Linguistics (ACL) , pp. 289-295. ISBN 9781954085466

10.18653/v1/2021.naacl-main.27

Fast and Scalable Dialogue State Tracking with Explicit Modular Decomposition

Dingmin Wang[♣], **Chenghua Lin**[♠], **Qi Liu**[♣], **Kam-Fai Wong**[♢]
[♣]Department of Computer Science, University of Oxford, UK
[♠] Department of Computer Science, The University of Sheffield, UK
[♢] The Chinese University of Hong Kong, Hong Kong SAR
{dingmin.wang, qi.liu}@cs.ox.ac.uk c.lin@shef.ac.uk
kfwong@se.cuhk.edu.hk

Abstract

We present a fast and scalable architecture called Explicit Modular Decomposition (EMD), in which we incorporate both classification-based and extraction-based methods and design four modules (for classification and sequence labelling) to jointly extract dialogue states. Experimental results based on the MultiWoz 2.0 dataset validates the superiority of our proposed model in terms of both complexity and scalability when compared to the state-of-the-art methods, especially in the scenario of multi-domain dialogues entangled with many turns of utterances.

1 Introduction

Dialogue state tracking (DST), responsible for extracting user goals/intentions from dialogues, is a core component in task-oriented dialogue systems (Young et al., 2013). A dialogue state is commonly represented as a (DOMAIN, SLOT TYPE, SLOT VALUE) triplet, e.g., (hotel, people, 3). We show an illustrated example of a multi-domain dialogue in Figure 1, which involves two domains, i.e., TRAIN and HOTEL.

Previous approaches for DST usually fall into the following four categories: (1) adopt encoder-decoder models to generates states (Kim et al., 2020; Ren et al., 2019; Li et al., 2019; Lee et al., 2019; Wu et al., 2019) ; (2) cast DST as a multi-label classification task when a full candidate-value list is available (Shan et al., 2020; Ramadan et al., 2018; Zhong et al., 2018; Ren et al., 2018); (3) employ span-based methods to directly extract the states (Chao and Lane, 2019; Gao et al., 2019); and (4) combine both classification-based and span-based methods to jointly complete the dialogue state extraction (Zhang et al., 2019).

The most related work to ours is DS-DST (Zhang et al., 2019), a joint model which highlights the problem that using classification-based or span-based approach alone is insufficient to cover all cases of DST in the task-oriented dialogue. While DS-DST has achieved some promising result on dialogue state tracking and demonstrated the utility of combining these two types of methods, some problems still remain unaddressed. On one hand, since the model is conditioned on domain-slot pairs, the computational complexity is not constant and will grow as the number of domains and slots involved in dialogues increases. To be more specific, if there are 1000 domain-slot pairs, the model needs to run 1000 times to obtain the expected dialogue states for the current turn at each time, which is a huge computational overhead. On the other hand, previous works usually directly concatenate the history content and the current utterance as input, which is difficult to scale in the multi-turn scenarios, especially when the number of turns of a dialogue is large. Furthermore, we observe that generative approaches may generate some *domain outlier*[1] triplets due to lack of domain constraints.

To tackle these issues, we propose a fast and

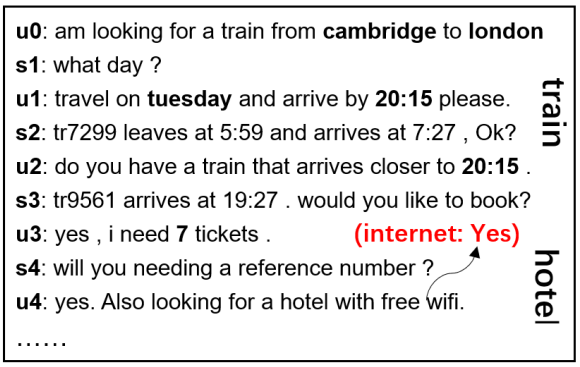

Figure 1: A multi-domain dialogue example extracted from MultiWoz 2.0. The **S-type** slot values are marked in bold and the arrow points to a pair of **C-type** slots and its corresponding value. The domain discussed changes from "train" to "hotel" at the fourth turn. Refer to Section 2 for the definitions of **C-type** and **S-type**.

[1]We refer a predicted result as "domain outlier" when slot types are out of the domain pertaining to current utterances.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 289–295
June 6–11, 2021.

scalable method called EMD, where we decompose DST into three classification modules and one sequence labeling module to jointly extract the dialogue states. The benefits of our approach are summarised below:

- *Efficient*: Different to the previous work, we employ a sequence labeling approach to directly annotate the domain-slot values in the utterance instead of iterating over all domain-slot pairs one by one, and thus greatly reduce the model complexity.
- *Constrained output*: To effectively model the relationship between the predicted domain and its associated slots, as well as to reduce the occurrence of *domain outlier* results, we propose a list-wise global ranking approach which uses Kullback-Leibler divergence to formulate the training objective.
- *Scalable*: Based on turn-level utterances rather than the whole history dialogue content, our proposed model offers better scalability, especially in tackling dialogues with multiple turns. Additionally, we employ a correction module to handle the changes of the states as the dialogue proceeds.

2 Our Proposed Model

Formally, a multi-turn dialogue is represented as $T = \{(s_1, u_1, d_1), (s_2, u_2, d_2), \cdots, (s_n, u_n, d_n)\}$, $d_i \in D$, where s_i, u_i and d_i refer to the system utterance, the user utterance, and the domain at turn i, respectively[2], and D represents the set of all domains in the training dataset. The overall architecture of our model is shown in Figure 2.

In our proposed model, we choose MT-DNN (Liu et al., 2019), pretrained model which has the same architecture as BERT but trained on multiple GLUE tasks (Wang et al., 2019). MT-DNN has been shown to be a better contextual feature extractor for downstream NLP tasks. Given dialogue utterances as input, we represent the output of MT-DNN as $\{H_{[CLS]}, H_1, H_2, \cdots, H_n\}$, where n is the length of the concatenation of the system and user utterances. As a sentence-level representation, $H_{[CLS]}$ is expected to encode the information of the whole input sequence (Devlin et al., 2019; Liu et al., 2019). Based on these contextual representations, we predict the domain (see §2.1) and belief states (see §2.2 and §2.3).

Figure 1 shows a typical multi-domain dialogue example, from which we can observe that some slot values can be directly found from utterances (e.g. `cambridge` and `london`), while other slot values are implicit which are more challenging to discover, e.g., requiring classification to infer the values (e.g. `internet:Yes`). We divide slots into two categories that are handled by two two separate modules: **S-type** slots whose values could be extracted from dialogue utterances, and **C-type** slots whose values do not appear in utterances and are chosen from one of the three values {yes, no, don't care}.

2.1 Domain Prediction Module (DPM)

In a multi-domain dialogue, the target domain may change as the dialogue proceeds. Different from some previous works (Chen et al., 2019; Castellucci et al., 2019), which directly use the first hidden state ($H_{[CLS]}$), in our model, apart from $H_{[CLS]}$, we additionally incorporate D_l, the domain result of the last turn into the our domain prediction module. The rationale behind is that when the domain of current utterances is not explicit, D_l can provide useful reference information for domain identification. Formally, the domain is predicted as:

$$y^d = \text{softmax}(W^d[H_{[CLS]}; E(D_l)]) \quad (1)$$

$$D_c = \arg\max(y^d), D_c \in D \quad (2)$$

where ; denotes the concatenation operation and $E(\cdot)$ embeds a word into a distributed representation using fixed MT-DNN (Liu et al., 2019). D_c is the predicted domain result.

2.2 S-type Slots Tagging Module (SSTM)

Domain-slot-matching constraints R To prevent our model from predicting some slots not belonging to the current domain, we generate a domain constrained contextual record $R \in \mathbb{R}^{1\times(s+1)}$, where s is number of **S-type** slots of all domains[3]. Concretely speaking, R is a distribution over all **S-type** slots and *[EMPTY]* using

$$R = \text{softmax}(W^R[H_{[CLS]}; E(D_l]) \quad (3)$$

[2]We assume that the turn-level utterances only contain one domain, and the Multiwoz 2.0 dataset we use in this paper also conforms to this assumption.

[3]We add a *[EMPTY]*, the value of which is expected to be 1 when there is no slot needed to be predicted. In particular, we consider the "don't care" as a special case in which the corresponding slot is considered not to be predicted.

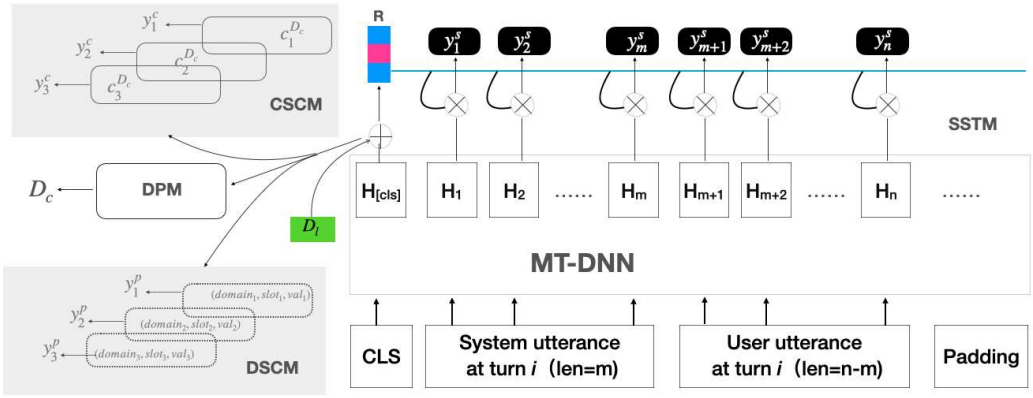

Figure 2: Our neural model architecture, which includes DPM for the domain prediction, whose output is the predicted domain, D_c. D_l denotes the domain at the previous turn. CSCM for the three classification of the domain-associated **C-type** slots, in which $c_i^{D_c}$ denotes one of **C-type** slots in D_c, and SSTM for tagging **S-type** slots in the given input, where tagging results are in IOB format; DSCM is for deciding whether to remove outdated states from the history state set. $y_i^p \in \{\text{yes, no}\}$, $y_i^c \in \{\text{yes, no, don't care}\}$ and $y_i^s \in \{O\} \bigcup \{\text{all } \textbf{S-type} \text{ slots}\}$.

In particular, L_R, the loss for R is defined as the Kullback-Leibler (KL) divergence between $Div(R^{real}||R)$, where distribution R^{real} from the ground truth is computed as follows:

- If there is no slot required to be predicted, $R^{real}_{[EMPTY]}$ receives a probability mass of 1 for the special slot *[EMPTY]*.
- If the number of slots needed to be predicted is $k(\geq 1)$, then corresponding k slot positions receive an equal probability mass of $1/k$.

Next, we employ a sequence labeling approach to directly annotate the domain-slot values in the utterance instead of iterating over all domain-slot pairs one by one. Specifically, to tag **S-type** slots of the given input, we feed the final hidden states of $H_1, H_2, \cdots, H_n$ into a softmax layer to classify all the **S-type** slots,

$$y_i^s = \text{softmax}(W^s H_i), i \in [1, 2, \cdots, N] \quad (4)$$

Instead of directly predicting **S-type** slot results based on y_i^s, we introduce a domain-slot-matching constraint R, which helps avoid generating **S-type** slots that do not belong to the predicted domain. The multiplication operation is given below,

$$\hat{y_i^s} = R \odot y_i^s \quad (5)$$

where $\odot$ is the element-wise multiplication.

2.3 C-type Slots Classification Module (CSCM)

Given the currently predicted domain result D_c, we build a set C_{D_c} which contains all **C-type** slots from all domains D. If C_{D_c} is empty, it indicates that there is no **C-type** slot needed to be predicted in the current domain. Otherwise, we classify each slot $c_i^{D_c}$ in C_D into one of the following following categories, i.e., {yes, no, don't care}, with the classification function below.

$$y^c = \text{softmax}(W^c[E(c_i^{D_c}); E(D_l); H_{[CLS]}]) \quad (6)$$

2.4 Dialogue State Correction Module (DSCM)

Previous models such as TRADE (Wu et al., 2019) and COMER (Ren et al., 2019) requires that all dialogue states need to be predicted from scratch at each turn, including those dialogue states that have already been predicted at previous turns. This poses a big challenge to the model in terms of scalability, especially when the number of dialogue turns increases. Conversely, the input of our model consists of the system utterance and the user utterance at the current turn, so our model only outputs the estimates of the dialogue states for the current turn, and the previous dialogues are directly included where no re-prediction is needed.

However, there is an issue with direct inclusion of previously predicted results in that some states may need to be updated or removed as the dialogue proceeds. For example, a user firstly looks for a hotel located in the center area, then a state (hotel, area, center) is estimated. Subsequently, the user utters a specified hotel name, e.g. "*I wanna the King House*", then the previous state (hotel, area, center) is outdated and should be removed. To this end, we design the dialogue state correction module to update previously predicted results in

order to improve the precision of the outputted dialogues states at each turn. Similar to the C-type classification module, we cast this situation as a classification task, and for each triple tuple p from the previous dialogue states, the classifier is formulated as

$$y^p = \text{sigmoid}(W^p[\hat{p}; E(D_l); H_{[CLS]}]) \quad (7)$$

Here each item in p is embedded using $E(\cdot)$ and $\hat{p}$ is the embedding sum of the three items in p.

During training, we use cross entropy loss for y^d, y^c, y^s and y^p, which are represented as L_{y^d}, L_{y^c}, L_{y^s} and L_{y^p}, respectively. The loss for R (denoted as L_R) is defined as Kullback-Leibler (KL) divergence between R^{real} and R (i.e, KL($R^{real}||R$)). All parameters are jointly trained by minimizing the weighted-sum of five losses (α, β, γ, θ, ϵ are hyper-parameters),

$$\text{Loss} = \alpha L_{y^d} + \beta L_{y^c} + \gamma L_{y^s} + \theta L_{y^p} + \epsilon L_R \quad (8)$$

2.5 Analysis of model complexity

Table 1 reports the Inference Time Complexity (ITC) proposed by (Ren et al., 2019), which is used to measure the model complexity. ITC calculates how many times inference must be performed to complete a prediction of the belief state in a dialogue turn. By comparison, we can observe that our model achieves the lowest complexity, $\mathcal{O}(1)$, attributed to the modular decomposition and the usage of the sequence label based model.

Model	ITC
DS-DST (Zhang et al., 2019)	$\mathcal{O}(n)$
SOM-DST (Kim et al., 2020)	$\mathcal{O}(n)$
SUMBT (Lee et al., 2019)	$\mathcal{O}(mn)$
GLAD (Zhong et al., 2018)	$\mathcal{O}(mn)$
COMER (Ren et al., 2019)n	$\mathcal{O}(n)$
TRADE (Wu et al., 2019)	$\mathcal{O}(n)$
EMD	$\mathcal{O}(1)$

Table 1: Inference Time Complexity (ITC) proposed in (Ren et al., 2019), m is the number of values in a pre-defined ontology list and n is the number of slots. Note that the ITC reported refers to the worst scenarios.

3 Experimental Setup

3.1 Setup

Dataset We evaluate our model performance based on the MultiWoZ 2.0 dataset (Budzianowski et al., 2018), which contains $10,000$ dialogues of 7 domains and 35 domain-slot pairs. Detailed dataset statistics is summarised in Table 2.

Evaluation metrics We utilize joint goal accuracy (JGA) (Henderson et al., 2014) to evaluate the model performance. Joint goal accuracy is the accuracy of the dialogue state of each turn and a dialogue state is regarded as correct only if all the values of slots are correctly predicted.

Implementation details The hyper-parameters of our model go as follows: both the embedding and the hidden size is 1024; we used a learning rate of 0.0001 with a gradient clip of 2.0, mini-batch SGD with a batch size of 32, and Adam optimizer (Kingma and Ba, 2014) for 50 epoch training. We set a value of 1 to the five weighted hyper-parameters: α, β, γ, θ, ϵ.

Metric	Train	Dev	Test
# of multi-domain dialogs	5,459	796	777
# of single-domain dialogs	2,979	204	223
# of total dialogs	8,438	1,000	1,000
Avg. # turns by dialog	6.7	7.4	7.3

Table 2: The statistics of the MultiWoZ2.0.

3.2 Results

Overall comparison We compare our models against six strong baselines on the multi-domain dataset MultiWoz. Results are reported in Table 3 based on joint goal accuracy (JGA). Our model achieves the best performance of 50.18% in the multi-domain testset, while the accuracy achieved in the single-domain is on par with the state-of-the-art results, which demonstrates the superiority of our model.

Model	JGA^s	JGA^m	JGA
SOM-DST (Kim et al., 2020)	-	-	**51.72**
COMER (Ren et al., 2019)	48.62	41.21	45.72
SUMBT (Lee et al., 2019)	46.99	39.68	42.40
DS-DST (Zhang et al., 2019)	**51.99**	48.69	51.01
GLAD (Zhong et al., 2018)	37.19	33.76	35.58
TRADE (Wu et al., 2019)	49.57	47.01	48.62
EMD	51.92	**50.18**	51.03

Table 3: Experimental results. JGA^s represents the accuracy calculated in all single domain dialogues and JGA^m refers to all multi-domain dialogues.

Analysis of model scalability We select 200 samples from the testing dataset, in which each dialogue has more than 8 turns of utterances between the system and the user. Then, taking the turn number 6 as a threshold, we divide the dialogue content into two categories, i.e., COLD and

Turn	Previous States	Domain	Target states	Predicted states for the current turn		
				COMMER	TRADER	EMD
1	{ }	Hotel	(hotel, internet, yes)	(hotel, internet, yes)	(hotel, internet, yes)	(hotel, internet, yes)
...	...	...	...	...	...	...
3	(hotel, internet, yes) (hotel, name, holiday inn)	Taxi	(hotel, internet, yes) (hotel, name, holiday inn) (taxi, destination, holiday inn)	(hotel, internet, yes) (hotel, name, holiday inn) (train, destination, holiday inn)	(hotel, internet, yes) (hotel, name, holiday inn) (taxi, destination, holiday inn)	(hotel, internet, yes) (hotel, name, holiday inn) (taxi, destination, holiday inn)
...	...	...	...	...	...	...
8	(hotel, internet, yes) (hotel, name, holiday inn) (taxi, destination, holiday inn)	Taxi	(hotel, internet, yes), (hotel, name, holiday inn), (taxi, destination, holiday inn)	(hotel, internet, yes) (hotel, name, holiday inn) (train, destination, holiday inn)	(hotel, internet, no) (hotel, name, holiday inn) (taxi, destination, holiday inn)	(hotel, internet, yes) (hotel, name, holiday inn) (taxi, destination, holiday inn)

Figure 3: Case study of predicated states by our model and two baselines. Erroneous states are highlighted in red.

HOT. Utterances with turn numbers lower than 6 are assigned to the COLD category and those above 6 to the HOT category.

Model	JGA	
	COLD	HOT
SOM-DST (Kim et al., 2020)	**52.21**	48.92
COMER (Ren et al., 2019)	46.01	40.72
SUMBT (Lee et al., 2019)	42.51	33.99
TRADE (Wu et al., 2019)	47.98	46.12
EMD	51.89	**51.01**

Table 4: Experimental results for the analysis of model scalabitiy. The sample size is 200.

From Table 4, we observe that the model performance has a big drop for the four baseline models, but our model achieves a relatively stable performance, achieving 51.01% in HOT and 51.89% in COLD, respectively. This demonstrates that our model is not only fast in terms of inference speed (cf. §2.5), but also has a good scalability which can maintain a high accuracy even when the dialogue proceeds into more turns and the input length becomes larger.

Ablation study We conduct two ablation experiments to investigate the impacts of D_l and R. We introduce a metric, called outlierslot ratio (OSR), denoting the proportion of slots predicted by our model that do not belong to the current domain. From Table 5, we notice that adding D_l improves the domain accuracy, where one possible reason is that some utterances may not have a clear domain attribute, and thus the incorporated previous domain is believed to provide useful guiding information in domain prediction. Besides, by comparing OSR with and without using R, we can observe that using R reduces the proportion of generating slots that do not align to the predicted domain, which further improves the model performance.

Case study To evaluate our proposed model qualitatively, we show an exemplary dialogue and illustrate some generated results by EMD and two baseline models in Figure 3. At turn 3 when the dialogue domain change from *hotel* to *taxi*, COMMER fails to capture the domain information and generates a domain outlier, "*train*", which does not conform to the current context. Conversely, dialogue generated by our model always conforms to the domain at the current turn, which may benefit from the incorporation of the domain constrained contextual record R. Besides, another observation is that as the dialogue proceeds to the turn 8 when the history dialogue content accumulates, TRADER makes an incorrect prediction in the hotel-internet slot, which is correctly identified at the turn 1. One possible reason is that it becomes more challenging for the model to correctly predict all dialogue state from scratch when both the history dialogue content and states involved increase. Instead of repeatedly generating those previously predicted states at each turn, our model only outputs the states for the current turn, and updates previous dialogue states with a separate module.

Model	Domain Acc.	OSR	JGA
EMD	95.23	44.62	51.03
- D_l	91.83	45.62	48.62
- R	93.19	54.83	47.23

Table 5: Ablation study results.

4 Conclusion

In this paper, we propose to decompose DST into multiple submodules to jointly estimate dialogue states. Experimental results based on the MultiWoz 2.0 dataset show that our model not only reduces the model complexity, but also gives high scalability in coping with multi-domain and long task-oriented dialogue scenarios.

References

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026.

Giuseppe Castellucci, Valentina Bellomaria, Andrea Favalli, and Raniero Romagnoli. 2019. Multi-lingual intent detection and slot filling in a joint bert-based model. *CoRR*, abs/1907.02884.

Guan-Lin Chao and Ian Lane. 2019. BERT-DST: scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1468–1472.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. BERT for joint intent classification and slot filling. *CoRR*, abs/1902.10909.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tür. 2019. Dialog state tracking: A neural reading comprehension approach. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 264–273. Association for Computational Linguistics.

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, ACL 2020, Online, July 5-10, 2020*, pages 567–582. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Hwaran Lee, Jinsik Lee, and Tae-Yoon Kim. 2019. SUMBT: slot-utterance matching for universal and scalable belief tracking. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5478–5483.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2019. A dual-attention hierarchical recurrent neural network for dialogue act classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 383–392, Hong Kong, China. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4487–4496.

Osman Ramadan, Pawel 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 432–437.

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.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 2780–2786.

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, ACL 2020, Online, July 5-10, 2020*, pages 6322–6333. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 808–819.

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.

Jianguo 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. *CoRR*, abs/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 (Volume 1: Long Papers)*, pages 1458–1467.