

A Multi-Task BERT Model for Schema-Guided Dialogue State Tracking

Eleftherios Kapelonis¹, Efthymios Georgiou^{1,2}, Alexandros Potamianos¹

¹School of ECE, National Technical University of Athens, Athens, Greece

²Institute for Language and Speech Processing, Athena Research Center, Athens, Greece

lkapelonis@gmail.com, efthygeo@mail.ntua.gr, potam@central.ntua.gr

Abstract

Task-oriented dialogue systems often employ a Dialogue State Tracker (DST) to successfully complete conversations. Recent state-of-the-art DST implementations rely on schemata of diverse services to improve model robustness and handle zero-shot generalization to new domains [1], however such methods [2, 3] typically require multiple large scale transformer models and long input sequences to perform well. We propose a single multi-task BERT-based model that jointly solves the three DST tasks of intent prediction, requested slot prediction and slot filling. Moreover, we propose an efficient and parsimonious encoding of the dialogue history and service schemata that is shown to further improve performance. Evaluation on the SGD dataset shows that our approach outperforms the baseline SGP-DST by a large margin and performs well compared to the state-of-the-art, while being significantly more computationally efficient. Extensive ablation studies are performed to examine the contributing factors to the success of our model.

Index Terms: dialogue state tracking, schema-guided, task-oriented dialogue, zero-shot learning

1. Introduction

Task-oriented dialogue is an important and active research area that has attracted a lot of attention in both academia and industry. The aim of task-oriented dialogue systems is to assist users in accomplishing daily activities like reserving a restaurant, booking tickets etc. An important component of a task-oriented dialogue system is the Dialogue State Tracker (DST) which tracks the user goal over multiple turns of dialogue. Based on a spoken utterance and the dialogue history, the DST predicts the dialogue state which represents the user goal. The predicted dialogue state is then used by other components to retrieve elements from a database, perform the actions requested by the user and respond accordingly [4].

Both single-domain [5, 6, 7, 8] and multi-domain [9, 10, 11, 12, 13] approaches have been used for DST training. The main DST’s tasks are to predict the active user intent (intent prediction), the slots that are requested by the user (requested slot prediction) and the values for slots given by the user until the turn (slot filling) [14]. Early neural methods use slot-dependent architectures [15, 16], training different parameters for every slot. In order to improve scalability and performance, slot-independent methods were proposed [17, 18] which share parameters between all slots.

Motivated by the ever-increasing number of diverse services used by commercial task-oriented systems, the schema-guided paradigm was developed [1]. Services or dialogue domains are defined by their corresponding *schema*, a structured ontology, which is usually a set of the supported intents and slots. Schema-guided approaches [1, 10, 19] often include a natural language description of the schema elements, e.g., in

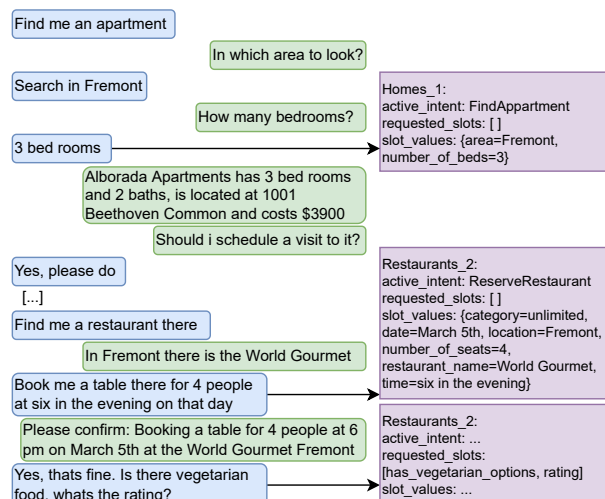


Figure 1: At every user turn the dialogue state is calculated for each involved service. The last two utterances are not always enough especially for the slot filling task. In such cases slot values can be found either in previous dialogue states or previous system actions.

the SGD dataset the schema for the service Restaurants.1 has a slot with name “party_size” and description “Party size for a reservation”. An important goal of schema-based approaches is scalability and generalization, i.e., to build systems that are capable of handling completely new domains and services.

Pre-trained transformer models (e.g. BERT [20], XLNet [21], GPT-2 [22], T5 [23] etc.) are the most popular solution for schema-based DST modeling. The importance of encoding the dialogue and schema together is highlighted in [24]. State-of-the-art methods use classification [2] or sequence-to-sequence pre-trained transformer models [25, 26, 27]. The entire dialogue is passed to the model multiple times with every possible schema element description (multi-pass approach). In [3], the authors concatenate all of the schema element descriptions with the dialogue (single-pass approach) slightly improving computational efficiency, yet still using the entire dialogue history. Other methods ([28, 29]) aim to address this issue by only encoding the last two utterances. To retrieve slot values found in earlier utterances, slot carryover mechanisms and a multi-pass approach were employed. Note that methods that encode the entire dialogue history, e.g., [2, 3], often perform better than methods that only encode the last two utterances.

In this paper, we propose a single multi-task BERT-based model that jointly performs intent prediction, requested slot prediction and slot filling. In the proposed model, we adopt slot carryover mechanisms and encode only the preceding system

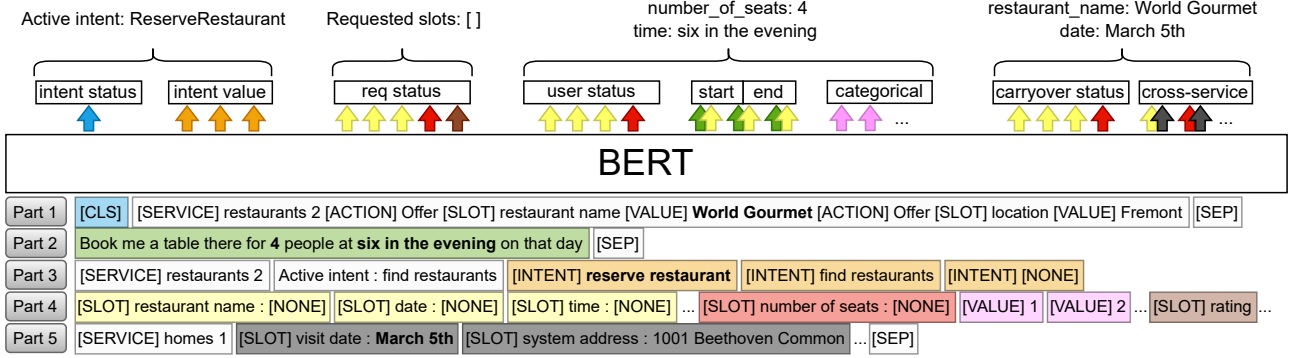


Figure 2: The inputs to the intent prediction, requested slot prediction, slot filling and slot carryover heads are shown for our proposed multi-task BERT model (top), along with an example encoding of the utterance and dialogue history that is the input to the base BERT model (bottom). Note the color coding of the input to the classification heads (top) that matches the various parts of the input sequence (bottom). For this example, the service in the system and the user utterance is *Restaurants.2*. The previous intent *FindRestaurants* changes to *ReserveRestaurant*. No slots are requested by the user. In the preceding system utterance, the system offers the value “World Gourmet” for the slot *restaurant_name* which the user accepts (slot carryover in *_sys_uttr*). The user gives the values “six in the evening” and “4” for the non-categorical slot *time* and the categorical slot *number_of_seats*. The date value is not uttered but it is implied that it has been mentioned before (slot carryover in *_cross_service_hist* from a previous service (*Homes.1*)). Part of the input is truncated for illustration purposes.

utterance and the current user utterance. Furthermore, the preceding system utterance is abstracted and represented as its underlying system actions. To achieve a more efficient and parsimonious input representation, we encode all of the schema elements together using only their names and we selectively include past dialogue states. Our proposed model significantly outperforms the baseline SGP-DST system and achieves near state-of-the-art performance. Extensive ablation studies reveal the impact of each strategy of our model on the slot filling task. Our key contributions are: 1) we propose a novel multi-task BERT-based model with slot carryover mechanisms, 2) we construct an efficient and parsimonious abstracted representation of the dialogue and schema that is shown to significantly improve performance while achieving greater computational efficiency. Our code is available as open-source ¹.

2. Method

The multi-task model architecture is shown in Fig. 2. The user utterance, previous system utterance, schema(ta) and past DST information (see Part 1 to 5) are encoded via BERT. Different pieces of the encoded sequence (see matching color coding in figure) are given as input to nine classification heads that perform the tasks of intent prediction (2 heads), requested slot prediction, slot filling (4 heads) and slot carryover (2 heads).

2.1. Notation

Let n be a dialogue service, $I(n)$ the set of intents in the service (including the special `[NONE]` intent) and $S(n)$ the set of slots in the service. Slots are divided to categorical and non-categorical slots. Let $S_{cat}(n) \subseteq S(n)$ be the set of categorical slots and $S_{noncat}(n) \subseteq S(n)$ the set of non-categorical slots. For every categorical slot, a set of possible values $V(s), s \in S_{cat}(n)$ are available. Furthermore, every slot may be informable or not depending on whether the user is allowed to give a value for it. We denote the set of the service informable slots as $S_{inf}(n) \subseteq S(n)$.

¹<https://github.com/lefteris12/multitask-schema-guided-dst>

Assume that at user turn t of a dialogue with N services we want to predict the dialogue state for service n . Essentially we have to predict the active intent $int(n)$ (intent prediction), the requested slots $req(n) \subseteq S(n)$ (requested slot prediction) and the values for the slots given by the user $usrSlotValue(s), s \in S_{inf}(n)$ (slot filling).

For every service $n', 1 \leq n' \leq N$, we denote its previous active intent as $prevInt(n')$. Also, for every slot $s \in S(n')$ we denote the last value given by the user for s as $prevUsrSlotValue(s)$. Furthermore, we use $prevSysSlotValue(s)$ and $sysUttrSlotValue(s)$ to denote the last value present in a system action, before turn $t - 1$ and at (system) turn $t - 1$ respectively. For $prevSysSlotValue(s)$ and $sysUttrSlotValue(s)$ we only use system actions that contain the slot s and exactly one value for the slot. In cases where the intent or the slot value is empty we use the `[NONE]` value.

We use S_{prev} to denote the set of slots $s \in S(n'), n' \neq n$ that $prevUsrSlotValue(s)$ or $prevSysSlotValue(s)$ is not `[NONE]` and $prevSlotValue(s)$ to denote their previous value. If $prevUsrSlotValue(s)$ is not `[NONE]` we use that value otherwise we use $prevSysSlotValue(s)$.

For every slot we employ additional binary features $x_{bin}(s)$. The binary features used are the following: 1) whether the service is new in the dialogue 2) whether the service switched (it was not present in the previous dialogue state) 3) whether exactly one value for the slot is found in the system utterance 4) whether exactly one value for the slot is found in previous system utterances 5) whether the slot is required in at least one intent 6) whether the slot is optional in all intents. Similar binary features have been used by [28].

2.2. Input representation

An example input can be seen in Fig. 2. In Part 1 we encode the preceding system utterance as a list of actions. In Part 2 we encode the current user utterance. In Part 3, the active service n , the previous active intent $prevInt(n)$ and all candidate intents belonging to service n are enumerated. Part 4 contains

the list of all slots $s \in S(n)$. If $s \in S_{inf}(n)$ we append $prevUtrSlotValue(s)$ and if $s \in S_{cat}(n) \cap S_{inf}(n)$ we also append all values in $V(s)$. Part 5 contains all other services found earlier in the dialogue. For every service we enumerate slot-value pairs from previous dialogue states or system actions, $s \in S_{prev}$ and their values $prevSlotValue(s)$. We prepend the word “system” before slots given by the system to differentiate them from slots given by the user (present in previous dialogue states).

For the schema we only use the names for the slots and intents instead of their full natural language descriptions used by other works. A number of custom tokens are introduced to the BERT vocabulary to indicate intents, slots etc.

2.3. Intent prediction task

Intent status head. We perform binary classification on the encoded [CLS] representation to predict the intent status as active or none.

Intent value head. For every intent $i \in I(n)$ we perform binary classification on its encoded [INTENT] representation to predict if the user switches to that intent.

If the intent status is active we choose the intent with the highest intent value probability. Otherwise, we keep the previous intent $prevInt(n)$.

2.4. Requested slot prediction task

Requested status head. For every slot $s \in S(n)$ we perform binary classification on its encoded [SLOT] representation in Part 4 to decide whether it is requested in the current user utterance.

2.5. Slot filling task

User status head. For every slot $s \in S_{inf}(n)$ we find the user status using its encoded [SLOT] representation in Part 4 to decide whether a value is given in the current user utterance. The user status classes are none, active and dontcare.

Categorical head. For the categorical slots $s \in S_{inf}(n) \cap S_{cat}(n)$ we perform binary classification for every possible value $v \in V(s)$ on its encoded [VALUE] representation to predict whether it is present in the user utterance.

Start and end heads. For the non-categorical slots $s \in S_{inf}(n) \cap S_{noncat}(n)$ we find the start and end span index distribution in the user utterance by performing classification on the concatenation of every user utterance token with the encoded [SLOT] representation.

If the user status is active, the value or the span with the highest probability is chosen for the slot. If the user status is dontcare, the special dontcare value is assigned to the slot.

2.6. Slot carryover

The user does not always explicitly give the value for the slot but they may instead refer to previous utterances. Therefore, we design slot carryover mechanisms to retrieve values for slots from the current or previous services.

Carryover status head. For every slot $s \in S_{inf}(n)$ we predict the carryover status using its encoded [SLOT] representation in Part 4 to find the source of the slot value. The carryover status classes are none, in_sys.uttr, in_service_hist and in_cross_service_hist.

For in_sys.uttr the slot is updated according to the value present in the preceding system utterance $sysUtrSlotValue(s)$. For in_service_hist the slot is

updated according to the value present in past system actions of service n , $prevSysSlotValue(s)$. In the above two cases, the user accepts the value given by the system and we simply carry that value over.

Cross-service head. For every slot $s' \in S_{prev}(n)$ we perform binary classification on the concatenation of its encoded [SLOT] representation in Part 5 with the encoded [SLOT] representation of s in Part 4 to decide whether we should carry the value over from slot s' to slot s . The highest probability slot s' is used as the source for the value s if the predicted carryover status is in_cross_service_hist. In this case, we assign the value $prevSlotValue(s')$ to slot s .

We first check the user status and if it is not none we update the slot value according to its output. Otherwise, we also check the carryover status. If it predicts that a carryover should take place, we update the slot value accordingly. If both user and carryover status are none the value remains the same as in the previous dialogue state, $prevUtrSlotValue(s)$.

2.7. Multi-task training

For the intent status, intent value, categorical, start, end and cross-service classification heads we derive the class probabilities with a two-layer feed-forward neural network. For the requested status, user status and carryover status classification heads we concatenate the slot binary features $x_{bin}(s)$ after the first layer.

We jointly optimize all classification heads, using the cross entropy loss for each head. For the intent prediction task the loss is $L_1 = w_1 L_{intstat} + w_2 L_{intval}$, for the requested slot prediction $L_2 = L_{reqstat}$ and for the slot filling task $L_3 = w_3 L_{usr} + w_4 L_{carry} + w_5 L_{cat} + w_6 L_{start} + w_7 L_{end} + w_8 L_{cross}$. Finally, the total loss is defined as $L = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3$.

3. Experimental Setup

Dataset. We evaluate our proposed system on the SGD dataset [1]. SGD contains a total of 21,106 dialogues over 20 domains and 45 services. We use the standard train/development/test split introduced in [1]. The test set contains 1,331 single-domain and 2,870 multi-domain dialogues and 77% of the dialogue turns contain at least one service not present in the train set. We use the following metrics: Joint Goal Accuracy (JGA), Average Goal Accuracy (Avg GA), Intent Accuracy and Requested Slot F1 as defined in [1].

Label Acquisition. In order to acquire labels for the user and carryover status, we use the user actions and search previous turns and dialogue states to find the source for the slot. We consider a slot as informable if and only if it is either required or optional in at least one intent. For every turn we run the model only for the involved services (services with at least one change in the dialogue state in the turn) according to the ground truth dialogue states during both training and evaluation for fair comparison to other works. The input to the model contains ground-truth previous dialogue states during training and during evaluation the previously predicted ones are used.

Training Setup. We use the huggingface² implementation of the BERT uncased models. For all our experiments we use a batch size of 16 and a dropout rate of 0.3 for the classification heads. We use the AdamW optimizer [30] with a linear warmup of 10% of the training steps and learning rate 2e-5. We train for a total of about 55k steps and evaluate on the development set

²https://huggingface.co/docs/transformers/model_doc/bert

Table 1: Comparison to other works

| System | Model | Params | JGA | Intent Acc | Req Slot F1 |
|----------------------|------------------------------------------------------|-------------|-------------|-------------|-------------------|
| SGD-baseline [1] | BERT _{BASE} | 110M | 25.4 | 90.6 | 96.5 |
| SGP-DST [28] | 6 × BERT _{BASE} | 660M | 72.2 | 91.9 | 99.0 |
| paDST [2] | 3 × RoBERTa _{BASE} + XLNet _{LARGE} | 715M | 86.5 | 94.8 | 98.5 |
| D3ST [3] (Base) | T5 _{BASE} | 220M | 72.9 | 97.2 | 98.9 |
| D3ST [3] (Large) | T5 _{LARGE} | 770M | 80.0 | 97.1 | 99.1 |
| D3ST [3] (XXL) | T5 _{XXL} | 11B | 86.4 | 98.8 | 99.4 |
| Ours (median result) | BERT _{BASE} | 110M | 82.7 | 94.6 | 99.4 |
| Ours (avg 3 runs) | BERT _{BASE} | 110M | 82.5 ± 1.0 | 94.7 ± 0.5 | 99.4 ± 0.1 |

Table 2: Ablation study

| System | JGA | Avg GA |
|---------------------------------|------|--------|
| Ours | 82.7 | 95.2 |
| w/o system actions | 71.9 | 91.6 |
| w. slot descriptions | 78.3 | 94.1 |
| w/o previous state | 79.8 | 94.0 |
| w/o schema augm. | 80.5 | 94.9 |
| w/o schema augm. & word dropout | 78.1 | 94.3 |
| w/o binary features | 81.0 | 94.4 |

Table 3: Effect of carryover mechanisms

| System | JGA | Avg GA |
|---------------------------------------------|------|--------|
| Ours | 82.7 | 95.2 |
| w/o in_sys_uttr | 62.8 | 87.0 |
| w/o in_service_hist | 76.4 | 92.7 |
| w/o in_cross_service_hist | 66.8 | 84.4 |
| SGD-baseline [1] | 25.4 | 56.0 |
| w/o in_service_hist & in_cross_service_hist | 61.6 | 81.9 |
| w/o all | 36.5 | 68.5 |

every 4k steps. We choose the model that performs best based on the JGA metric on the development set.

Preprocessing and augmentation. We preprocess the schema elements and the system actions by removing underscores and splitting the words when on CamelCase and snake.case style. We randomly ($p = 0.1$) replace the input tokens in the user utterance with the [UNK] token (word dropout) and shuffle the order of the schema elements in Parts 3-5 during training as proposed by [31]. We also apply random ($p = 0.1$) data augmentation through synonym replacement and random swap to the intents, slots and values in Parts 3-4 (schema augm.) via [32].

4. Results and Discussion

Comparison to other works. In Table 1 we compare our model to SGD-baseline, SGP-DST, paDST and three D3ST implementations of variable size. The SGD baseline [1] fine-tunes BERT with the last two utterances as input and uses pre-computed BERT embeddings for the schema. SGP-DST [28] uses the last two utterances and slot carryover mechanisms to retrieve values for slots which were mentioned in previous utterances. paDST [2] and D3ST [3] encode the entire dialogue history until the current turn and calculate the dialogue state from scratch. We report the metrics and the number of parameters in the pre-trained model(s) fine-tuned by each method.

Our method clearly outperforms SGP-DST in all tasks indicating that our strategies are effective. Some of the entire-dialogue models outperform our model, especially when they use much more parameters (D3ST XXL) or apply more

handcrafted features, special rules and dialogue augmentation through back-translation (paDST). Overall, the proposed approach achieves near state-of-the-art performance despite using a much smaller model size and a shorter input representation.

Ablation study. We perform an ablation study (Table 2) to show the contribution of each of the proposed strategies on the slot filling task. Replacing the system utterance with a set of system actions (w/o system actions) has the biggest effect on performance (see input sequence Part 1 in Fig. 2). The system actions contain key information including the slot names and their respective values, helping our model identify which slots are requested, offered, confirmed etc. and predict the user and carryover status most accurately. Performance drops when we additionally include the slot descriptions for the informable slots of the current service (w. slot descriptions, see Parts 3-4 of input). By removing previous intent and slot values in Parts 3-4 (w/o previous state) we observe a performance drop but also a training speedup because of the smaller input sequence. We also observe an improvement by performing schema augmentation and word dropout possibly because these strategies help to avoid overfitting (w/o schema augm. & word dropout). The hand-crafted binary features can slightly benefit the system (w/o binary features).

Effect of slot carryover mechanisms. In Table 3 we show the effect of the various slot carryover mechanisms. For these experiments the model is trained once and during evaluation we replace each carryover status class with “none”. As expected, dropping “in_sys_uttr” has the biggest impact on performance. “in_cross_service_hist” is also important because of the large number of multi-domain dialogues. By removing “in_service_hist”, performance is less affected. Without “in_service_hist” and “in_cross_service_hist” (by only considering the last two utterances) we still achieve a higher accuracy than the SGD-baseline.

5. Conclusions

We propose a multi-task model for schema-guided dialogue state tracking that reasons for all three critical DST tasks simultaneously, as well as, an efficient and parsimonious encoding of user input, schemata and dialogue history. Close to state-of-the-art performance is achieved, using a significantly smaller model and input encoding. Among the various proposed enhancements to the model we show that abstracting the preceding system utterance with system actions gives the biggest performance boost. Strategies like appending previous dialogue states, data augmentation and adding hand-crafted features further improve performance. We believe that these strategies can guide the design of accurate, efficient and ontology-independent task-oriented DST capable of scaling to large multi-domain dialogues, important in real world applications.

6. References

- [1] A. Rastogi, X. Zang, S. Sunkara, R. Gupta, and P. Khaitan, "Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset," 2020.
- [2] Y. Ma, Z. Zeng, D. Zhu, X. Li, Y. Yang, X. Yao, K. Zhou, and J. Shen, "An end-to-end dialogue state tracking system with machine reading comprehension and wide & deep classification," 2020.
- [3] J. Zhao, R. Gupta, Y. Cao, D. Yu, M. Wang, H. Lee, A. Rastogi, I. Shafran, and Y. Wu, "Description-driven task-oriented dialog modeling," 2022.
- [4] Z. Zhang, R. Takanobu, M. Huang, and X. Zhu, "Recent advances and challenges in task-oriented dialog system," *CoRR*, vol. abs/2003.07490, 2020.
- [5] J. Williams, A. Raux, D. Ramachandran, and A. Black, "The dialog state tracking challenge," in *Proceedings of the SIGDIAL 2013 Conference*. Metz, France: Association for Computational Linguistics, Aug. 2013, pp. 404–413.
- [6] M. Henderson, B. Thomson, and J. D. Williams, "The second dialog state tracking challenge," in *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. Philadelphia, PA, U.S.A.: Association for Computational Linguistics, Jun. 2014, pp. 263–272.
- [7] A. Bordes and J. Weston, "Learning end-to-end goal-oriented dialog," *CoRR*, vol. abs/1605.07683, 2016.
- [8] T. Wen, M. Gasic, N. Mrksic, L. M. Rojas-Barahona, P. Su, S. Ultes, D. Vandyke, and S. J. Young, "A network-based end-to-end trainable task-oriented dialogue system," *CoRR*, vol. abs/1604.04562, 2016.
- [9] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, and M. Gašić, "Multiwoz—a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling," *arXiv preprint arXiv:1810.00278*, 2018.
- [10] M. Eric, R. Goel, S. Paul, A. Sethi, S. Agarwal, S. Gao, and D. Hakkani-Tür, "Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines," 2019.
- [11] X. Zang, A. Rastogi, S. Sunkara, R. Gupta, J. Zhang, and J. Chen, "Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines," *arXiv preprint arXiv:2007.12720*, 2020.
- [12] T. Han, X. Liu, R. Takanabu, Y. Lian, C. Huang, D. Wan, W. Peng, and M. Huang, "Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and coreference annotation," in *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 2021, pp. 206–218.
- [13] F. Ye, J. Manotumruksa, and E. Yilmaz, "Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation," *arXiv preprint arXiv:2104.00773*, 2021.
- [14] A. Rastogi, X. Zang, S. Sunkara, R. Gupta, and P. Khaitan, "Schema-guided dialogue state tracking task at DSTC8," *CoRR*, vol. abs/2002.01359, 2020.
- [15] M. Henderson, B. Thomson, and S. J. Young, "Word-based dialog state tracking with recurrent neural networks," in *SIGDIAL Conference*, 2014.
- [16] N. Mrksic, D. Ó. Séaghdha, T. Wen, B. Thomson, and S. J. Young, "Neural belief tracker: Data-driven dialogue state tracking," *CoRR*, vol. abs/1606.03777, 2016.
- [17] A. Rastogi, D. Hakkani-Tür, and L. P. Heck, "Scalable multi-domain dialogue state tracking," *CoRR*, vol. abs/1712.10224, 2017.
- [18] L. Ren, K. Xie, L. Chen, and K. Yu, "Towards universal dialogue state tracking," *CoRR*, vol. abs/1810.09587, 2018.
- [19] J. E. M. Mosig, S. Mehri, and T. Kober, "STAR: A schema-guided dialog dataset for transfer learning," *CoRR*, vol. abs/2010.11853, 2020.
- [20] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.
- [21] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLnet: Generalized autoregressive pretraining for language understanding," *CoRR*, vol. abs/1906.08237, 2019.
- [22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.
- [23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.
- [24] J. Cao and Y. Zhang, "A comparative study on schema-guided dialogue state tracking," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 782–796.
- [25] C. Lee, H. Cheng, and M. Ostendorf, "Dialogue state tracking with a language model using schema-driven prompting," *CoRR*, vol. abs/2109.07506, 2021.
- [26] J. Zhao, M. Mahdih, Y. Zhang, Y. Cao, and Y. Wu, "Effective sequence-to-sequence dialogue state tracking," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7486–7493.
- [27] Z. Lin, B. Liu, S. Moon, P. A. Crook, Z. Zhou, Z. Wang, Z. Yu, A. Madotto, E. Cho, and R. Subba, "Leveraging slot descriptions for zero-shot cross-domain dialogue state tracking," *CoRR*, vol. abs/2105.04222, 2021.
- [28] Y.-P. Ruan, Z.-H. Ling, J.-C. Gu, and Q. Liu, "Fine-tuning bert for schema-guided zero-shot dialogue state tracking," 2020.
- [29] M. Li, H. Xiong, and Y. Cao, "The SPPD system for schema guided dialogue state tracking challenge," *CoRR*, vol. abs/2006.09035, 2020.
- [30] I. Loshchilov and F. Hutter, "Fixing weight decay regularization in adam," *CoRR*, vol. abs/1711.05101, 2017.
- [31] S. Kim, S. Yang, G. Kim, and S. Lee, "Efficient dialogue state tracking by selectively overwriting memory," *CoRR*, vol. abs/1911.03906, 2019.
- [32] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," 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)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6383–6389.