INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Coreference Augmentation for Multi-Domain Task-Oriented Dialogue State Tracking

Ting Han[1*], *Chongxuan Huang*[2*], *Wei Peng*[2]

[1]University of Illinois at Chicago, Chicago, USA
[2]Artificial Intelligence Application Research Center, Huawei Technologies Shenzhen, PRC

than24@uic.edu, {huang.chongxuan, peng.wei1}@huawei.com

Abstract

Dialogue State Tracking (DST), which is the process of inferring user goals by estimating belief states given the dialogue history, plays a critical role in task-oriented dialogue systems. A coreference phenomenon observed in multi-turn conversations is not addressed by existing DST models, leading to sub-optimal performances. In this paper, we propose Coreference Dialogue State Tracker (CDST) that explicitly models the coreference feature. In particular, at each turn, the proposed model jointly predicts the coreferred domain-slot pair and extracts the coreference values from the dialogue context. Experimental results on MultiWOZ 2.1 dataset show that the proposed model achieves the state-of-the-art joint goal accuracy of 56.47%.

Index Terms: task-oriented dialogue, dialogue state tracking, coreference

1. Introduction

Developing task-oriented dialogue systems has attracted interest from academia and industry due to its value in real-world applications. Dialogue state tracking (DST), one of the core functions of a task-oriented dialogue system, infers users' requirements by estimating the most probable belief states in the multi-run dialogue process. Given a user utterance, a DST model records user goals by filling a predefined slot set, as shown in Figure 1. The recorded state is updated at each turn and passed to downstream modules in the pipeline to generate a proper system response.

Traditional DST approaches rely on predefined ontology to produce belief states in terms of sets of slot-value pairs with the highest probability [1, 2, 3, 4]. It remains a challenge for these DST methods to address out-of-domain (OOD) dialogue states not accessible from the predefined ontology. Recent DST models [5, 6, 7] resort to dialogue context to deal with the OOD problem mentioned above. Although progress has been made, they lack a mechanism to model the coreference features prevalent in human conversations, leading to sub-optimal performances. As illustrated in Figure 1, the slot value *"Saturday"* (highlighted in red) mentioned in the previous turn does not explicitly appear in the current turn, in which only the coreferred term *"the day of my hotel booking"* is present. It is a critical step to model the coreference features across dialogue turns in DST.

Several research works claim to improve the performance remarkably by accommodating coreference in the model. [8] show a significant performance enhancement through including coreference into training datasets. [9] boost the response quality of the dialogue system by restoring incomplete utterances with coreference labels. [10] achieve a notable improvement to the baseline by recovering coreference information in the utterances. Recently, TripPy [11] implicitly models coreference relations among slots by a copy mechanism via which the coreference value for a particular slot is copied from other slots. However, the predefined schema adopted by TripPy for training coreference slot relationships is affected by annotation noise, thus preventing coreference feature learning.

Instead of leveraging coreference-enriched datasets or predefined ontology mentioned above, we tackle the coreference problem in DST from another angle. Inspired by [6], we propose Coreference Dialogue State Tracker (CDST) to model the coreference feature explicitly. In particular, at each turn, the proposed model jointly predicts the coreferred domain-slot pair and extracts the coreference values from the dialogue context. Experimental results on MultiWOZ 2.1 [12] dataset demonstrate the effectiveness of CDST by achieving the state-of-the-

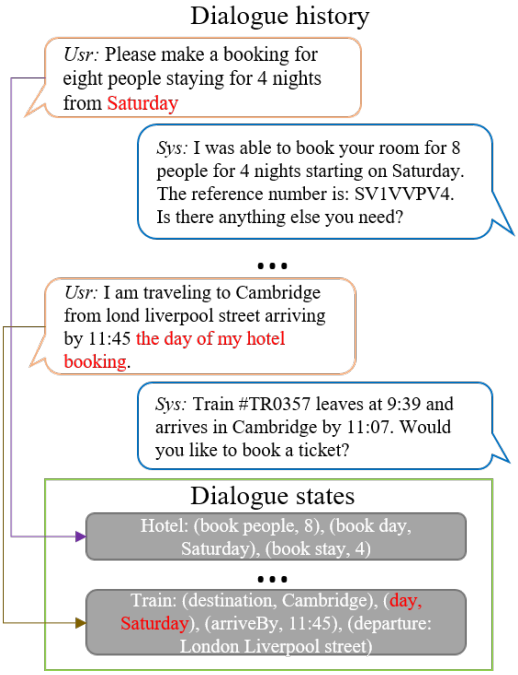

Figure 1: *An example of dialogue state tracking in a conversation. The arrows connect user utterance and its paired dialogue state. The state tracker needs to track slot values mentioned by the user for all slots during the conversation.*

[*]equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-1463

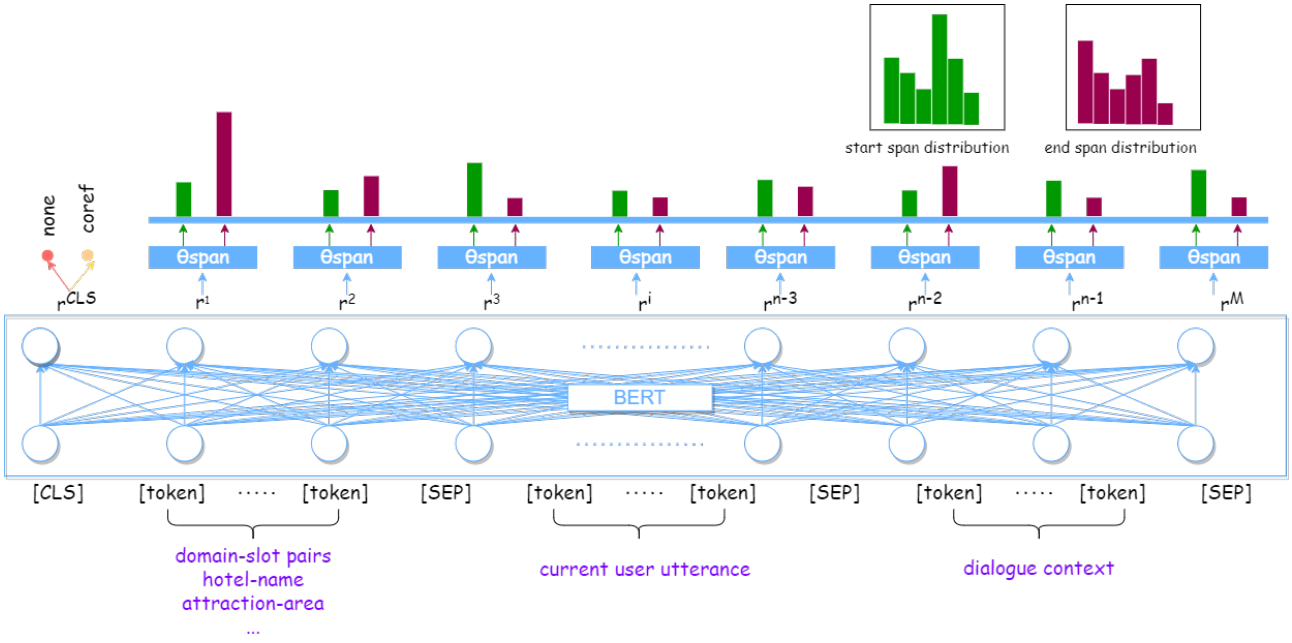

Figure 2: *Network architecture of our proposed CDST. The input consists of domain-slot pairs, current user utterance and dialogue context, and the CDST outputs values for the domain-slot pairs if coreferences exist.*

art joint goal accuracy of 56.47%. Additionally, we perform an empirical analysis to verify the necessity of the coreference solution.

The rest of the paper is organized as follows. In the next section, the proposed approach is presented. Section 3 provides experimental results and discussions followed by the conclusion of this paper.

2. The proposed approach

The network architecture of CDST is presented in Figure 2. Denoting $X = \{(U_1, S_1), \ldots, (U_T, U_T)\}$ as the sequence of pairs which represents a dialogue of length T. U_t and S_t are user utterance and system utterance at turn t respectively, and $DS = \{DS_1, DS_2, ..., DS_N\}$ is N domain-slot pairs describing different subjects in a task-oriented dialogue system, for instance, *Restaurant-name*. A restaurant name requested by a user during a conversation should be paired with the slot. CDST performs coreference dialogue state tracking by predicting coreference or "none" value for each given slot DS_n at each turn. As shown in Figure 2, the contextual representations are extracted from sequential combinations of domain-slot pairs, the current user utterance and dialogue context via BERT [13]. The aggregated sentence representation, r^{CLS} is used to differentiate the coreference slots from the none-coreferred ones. The token-level representations, $[r^1, ..., r^M]$ are used to pinpoint coreference entities in the dialogue context of the input.

2.1. Dialogue context encoder

We adopt a pre-trained BERT as the contextual encoder. At each turn t, the n_{th} domain-slot pairs DS_n, the current user utterance U_t and dialogue history C_t are concatenated as inputs to the encoder:

$$R_{tn} = BERT([CLS] \oplus DS_n \oplus [SEP] \oplus U_t \oplus [SEP] \oplus C_t \oplus [SEP]) \quad (1)$$

where [CLS] is the special token mandatory to be the first token of each input, and [SEP] is the special separator token. $C_t = (U_1, S_1), ..., (U_{t-1}, S_{t-1})$ denotes the dialogue context up to the previous turn $t-1$. The outputs of Equation (1) is the last hidden layer of BERT, i.e., $R_{tn} = [r_{tn}^{CLS}, r_{tn}^1, ..., r_{tn}^M]$, where M denotes the total number of the input tokens, r_{tn}^{CLS} is the aggregated representation of the entire input, which is later used by slot type classification. The remaining vectors $[r_{tn}^1, ..., r_{tn}^M]$ are the token-level representations for the input sequence and delivered to the coreference value predictions. The BERT is initialized from a pre-trained checkpoint and the parameters are further finetuned.

2.2. Slot type classification

One task of CDST is to determine, for each domain-slot pair $DS_n \in DS$, if the domain-slot is coreferred. It can be treated as a binary classification on a variable with values $\{coref, none\}$, indicating whether coreference exists for the given domain-slot pair. The classification takes r_{tn}^{CLS} as the input and outputs the probability of being coreferred for the domain-slot pair DS_n at the turn t.

$$P_{tn} = softmax(W_n^c \cdot r_{tn}^{CLS} + b_n^c) \quad (2)$$

where W_n^c and b_n^c are trainable parameters for the domain-slot pair DS_n.

2.3. Coreference value retrieval

Simultaneously, CDST retrieves coreference value from the given dialogue context through span prediction. For each

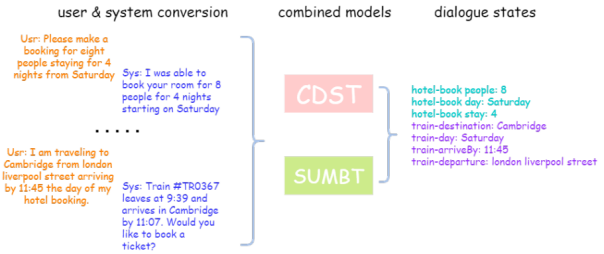

Figure 3: *An illustration of the combination of CDST and SUMBT. The two models work jointly as a dialogue state tracker.*

domain-slot pair $DS_n \in DS$, a span prediction layer takes each token of $[r^1_{tn}, ..., r^M_{tn}]$ as the input and outputs the start and end positions for the token:

$$(start^i_{tn}, end^i_{tn}) = W^{span}_n \cdot r^i_{tn} + b^{span}_n \tag{3}$$

$$P_{tn,start} = softmax(start_{tn}) \tag{4}$$

$$P_{tn,end} = softmax(end_{tn}) \tag{5}$$

where W^{span}_n and b^{span}_n are learnable parameters. The final span for the domain-slot pair DS_n is obtained through an argmax function over the start and end positions of all tokens:

$$SPAN_{tn,start} = argmax(P_{tn,start}) \tag{6}$$

$$SPAN_{tn,end} = argmax(P_{tn,end}) \tag{7}$$

The slot type classification and coreference value retrieval coordinate each other to complete coreference dialogue state tracking. The retrieved coreference value is filled into the given domain-slot pair if it is classified as coreferred. Otherwise, the retrieved values are discarded.

3. Experiments and results

We conduct three different experiments: (1) evaluating CDST solely on coreference; (2) jointly examining CDST along with SUMBT[1] [3] on the purpose of a DST task, and a combination of two models is provided in Figure 3; (3) adopting T5 [14] for comparisons and the empirical analysis with CDST. We use the joint goal accuracy (JGA) as the evaluation metric.

3.1. Dataset

All three experiments are conducted on the MultiWOZ 2.1 dataset, which is a multi-domain task-oriented dialogue dataset with more than 10k dialogues spanning over seven different domains. There are 30 slots and five domains (train, restaurant, hotel, taxi, attraction) commonly used for DST tasks. The number of dialogues for the training, development, and test sets are 8,348, 1,000, and 1,000, respectively. The coreference data is from MultiWOZ 2.3 [15], an improved dataset based on MultiWOZ 2.1 with extra coreference annotations. In total, 20.16% of the dialogues contain coreference.

[1]We skip the details of SUMBT, which is accessible in the references.

3.2. Implementation details

We adopt BERT-medium, uncased[2] [16] for CDST, and use a T5-base framework from Simple Transformers[3]. The training hyperparameters of the two models are listed in the Table 1.

Table 1: *Hyperparameters for BERT-medium and T5-base*

Model Hyperparameters	BERT-Medium	T5-base
learning rate	1×10^{-4}	1×10^{-3}
max seq length	512	196
warmup ratio	0.1	0.06
train epoch	10	20
optimizer	ADAM	Adafactor
train batchsize	2	16

The loss function of the T5-base remains intact during training. Comparatively, the loss of CDST is computed through a summation of two loss components as follow:

$$\mathcal{L}_{total} = \beta \cdot \mathcal{L}_{slot\_type} + (1 - \beta) \cdot \mathcal{L}_{SPAN} \tag{8}$$

where $\mathcal{L}_{slot\_type}$ and $\mathcal{L}_{SPAN}$ are cross-entropy loss for the corresponding predictions. $\mathcal{L}_{SPAN}$ is the average loss of the span start and end losses. β is a parameter that is empirically set to 0.8.

We finetune T5's encoder-decoder architecture to handle coreference modeling, which is treated as a question-answering (QA) task. The question includes the domain-slot pair concatenated with the current user utterance. We treat dialogue context as the passage in the QA task. T5 is trained using coreferred data samples to maximize the performance of the QA task. The predicted coreference value for a particular slot is used to update the slot. A rule-based mechanism is applied to merge results of T5 into SUMBT before evaluating the joint goal accuracy.

3.3. Results

As shown in Table 2, SUMBT+CDST outperforms other dialogue state trackers by achieving the state-of-the-art joint goal accuracy (JGA). CDST adds 3.9% of JGA to the performance of SUMBT, demonstrating its effectiveness in the DST task. Although the result of SUMBT+T5 is comparable to our proposed model, it contains twice as many parameters (220 million) as that of the CDST, putting T5 a less desirable option.

3.4. Empirical analysis and discussions

Empirical analyses on the coreference results of T5 and CDST are summarized in Table 3 and Figure 4.

We perform an ablation study on coreference performance under various combinations for inputs. Table 3 presents the performances of different input combinations among domain-slot pairs, current user utterance, and dialogue context. It can be observed that current user utterance (uttr.) and domain-slot pairs (aka slot.) contribute approximately 10% of the joint goal accuracy (JGA) to the CDST, boosting the performance from 45.52% to 55.84%.

In the ablation test, we evaluate T5 using accuracy as the performance metric instead of JGA because T5 alone can only handle single-turn dialogue data. It is confirmed that current

[2]https://github.com/google-research/bert
[3]https://github.com/ThilinaRajapakse/simpletransformers

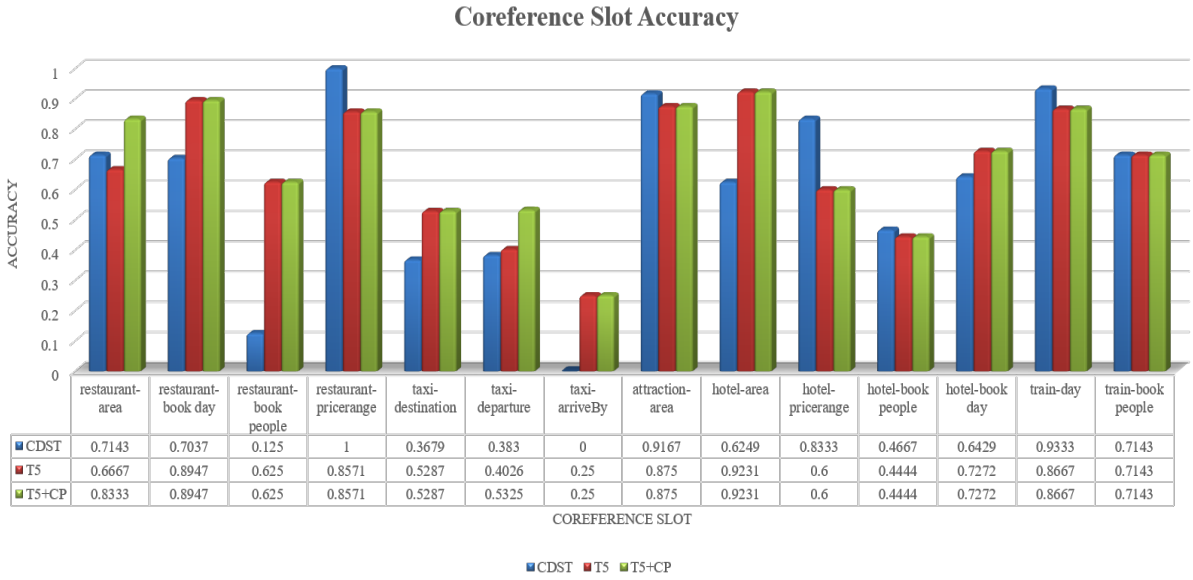

Figure 4: *A comparison of the accuracy of coreference slots of CDST, T5 and T5+CP.*

Table 2: *DST results on MultiWOZ 2.1. The result of the original SUMBT is trained and evaluated on MultiWOZ 2.0, depicted with a notation '*'. We further implement SUMBT with MultiWOZ 2.1 and the result is labelled by '+'.*

Models	MultiWOZ 2.1
SUMBT [3]	42.40%*
TRADE [5]	45.60%
DSTQA [17]	51.17%
DS-DST [4]	51.21%
SOM-DST [7]	52.57%
DST-picklist [4]	53.50%
SST [18]	55.23%
TripPy [11]	55.29%
SUMBT	52.57%+
SUMBT+T5	56.41%
SUMBT+CDST	**56.47%**

Table 3: *Coreference results of CDST and T5 in an ablation study. "-uttr." means experimenting without including the current user utterance in the input. "-slot." refers to an experiment without inputting domain-slot pairs. We do not record the score for "T5-uttr.,-slot." because "slot." has been treated as a question (a mandatory input) by T5 in this study.*

Models	Accuracy
T5+CP	70.72%
T5	65.13%
T5-uttr.	58.55%
T5-uttr.,-slot.	-
Models	**JGA**
CDST	55.84%
CDST-uttr.	48.31%
CDST-uttr., -slot.	45.52%

user utterance (uttr.) and domain-slot pairs (slot.) also contribute significantly to T5. We further include "coreference expression phrase" ("CP"), i.e., "the day" in Figure 1, into the experiment to investigate their effects on the DST task when using T5. As shown in Table 3, the performance of T5 is significantly enhanced to 70.72% as a consequence.

In the Figure 4, we record the coreference accuracy per slot for the top 3 performances in Table 3. In total, there are 14 domain-slot pairs with coreference. It can be observed that:

- Both T5 and CDST are insensitive to numbers leading to relatively lower accuracies for slots related to "book people". The worst performance is for "taxi-arriveBy" which is a time-related slot.
- The poor accuracy of "taxi-destination" and "taxi-departure" indicates a lack of discriminating power for the model to handle departure and destination information concurrently. The slight improvement of the performance of T5+CP relating to "taxi-departure" may be attributed to the information introduced by the "coreference context phrase" (CP).
- The inclusion of the "coreference expression phrase" (CP) has only benefited two slots ("restaurant-area" and "taxi-departure"). Further works are required to pinpoint the potential cause of this limitation.

4. Conclusion

In this paper, we propose Coreference Dialogue State Tracker (CDST) to explicitly model coreference in the DST task. CDST directly retrieves coreferred slot values from the given dialogue context without a predefined ontology. The experimental results demonstrate the effectiveness of CDST by achieving the state-of-the-art joint goal accuracy. Furthermore, we conduct an empirical analysis to disclose the insights associated with coreference modeling in DST.

5. References

[1] N. Mrkšić, D. Ó Séaghdha, T.-H. Wen, B. Thomson, and S. Young, "Neural belief tracker: Data-driven dialogue state tracking," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, Jul. 2017, pp. 1777–1788.

[2] V. Zhong, C. Xiong, and R. Socher, "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)*, Melbourne, Australia, Jul. 2018, pp. 1458–1467.

[3] H. Lee, J. Lee, and T.-Y. Kim, "SUMBT: Slot-utterance matching for universal and scalable belief tracking," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul. 2019, pp. 5478–5483.

[4] J. Zhang, K. Hashimoto, C.-S. Wu, Y. Wang, P. Yu, R. Socher, and C. Xiong, "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*, Barcelona, Spain (Online), Dec. 2020, pp. 154–167.

[5] C.-S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, and P. Fung, "Transferable multi-domain state generator for task-oriented dialogue systems," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul. 2019, pp. 808–819.

[6] G.-L. Chao and I. Lane, "BERT-DST: Scalable End-to-End Dialogue State Tracking with Bidirectional Encoder Representations from Transformer," in *Proc. Interspeech 2019*, 2019, pp. 1468–1472.

[7] S. Kim, S. Yang, G. Kim, and S.-W. Lee, "Efficient dialogue state tracking by selectively overwriting memory," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, Jul. 2020, pp. 567–582.

[8] J. Quan, D. Xiong, B. Webber, and C. Hu, "GECOR: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue," 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, Nov. 2019, pp. 4547–4557.

[9] Z. Pan, K. Bai, Y. Wang, L. Zhou, and X. Liu, "Improving open-domain dialogue systems via multi-turn incomplete utterance restoration," 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, Nov. 2019, pp. 1824–1833.

[10] H. Su, X. Shen, R. Zhang, F. Sun, P. Hu, C. Niu, and J. Zhou, "Improving multi-turn dialogue modelling with utterance ReWriter," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, Jul. 2019, pp. 22–31.

[11] M. Heck, C. van Niekerk, N. Lubis, C. Geishauser, H.-C. Lin, M. Moresi, and M. Gasic, "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*, 1st virtual meeting, Jul. 2020, pp. 35–44.

[12] M. Eric, R. Goel, S. Paul, A. Sethi, S. Agarwal, S. Gao, A. Kumar, A. Goyal, P. Ku, and D. Hakkani-Tur, "MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines," in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 422–428.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, Minneapolis, Minnesota, Jun. 2019.

[14] 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.

[15] T. Han, X. Liu, R. Takanobu, Y. Lian, C. Huang, W. Peng, and M. Huang, "Multiwoz 2.3: A multi-domain task-oriented dataset enhanced with annotation corrections and co-reference annotation," *arXiv preprint arXiv:2010.05594*, 2020.

[16] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: On the importance of pre-training compact models," *arXiv preprint arXiv:1908.08962*, 2019.

[17] L. Zhou and K. Small, "Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering," *arXiv preprint arXiv:1911.06192*, 2019.

[18] L. Chen, B. Lv, C. Wang, S. Zhu, B. Tan, and K. Yu, "Schema-guided multi-domain dialogue state tracking with graph attention neural networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 7521–7528, Apr. 2020.