DSTEA : Improving Dialogue State Tracking via Entity Adaptive Pre-training

Yukyung Lee
Korea University
Seoul, Republic of Korea
yukyung_lee@korea.ac.kr

Takyoung Kim
Korea University
Seoul, Republic of Korea
takyoung_kim@korea.ac.kr

Hoonsang Yoon
Korea University
Seoul, Republic of Korea
hoonsang_yoon@korea.ac.kr

Pilsung Kang
Korea University
Seoul, Republic of Korea
pilsung_kang@korea.ac.kr

Junseong Bang
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
hjbang21pp@gmail.com

Misuk Kim
Sejong University
Seoul, Republic of Korea
misuk.kim@sejong.ac.kr

ABSTRACT

Dialogue State Tracking (DST) is critical for comprehensively interpreting user and system utterances, thereby forming the cornerstone of efficient dialogue systems. Despite past research efforts focused on enhancing DST performance through alterations to the model structure or integrating additional features like graph relations, they often require additional pre-training with external dialogue corpora. In this study, we propose DSTEA, improving Dialogue State Tracking via Entity Adaptive pre-training, which can enhance the encoder through by intensively learning key entities in dialogue utterances. DSTEA identifies these pivotal entities from input dialogues utilizing four different methods: ontology information, named-entity recognition, the spaCy toolkit, and the flair library. Subsequently, it employs selective knowledge masking to train the model effectively. Remarkably, DSTEA only requires pre-training without the direct infusion of extra knowledge into the DST model. This approach resulted in substantial performance improvements of four robust DST models on MultiWOZ 2.0, 2.1, and 2.2, with joint goal accuracy witnessing an increase of up to 2.69% (from 52.41% to 55.10%). Further validation of DSTEA's efficacy was provided through comparative experiments considering various entity types and different entity adaptive pre-training configurations such as masking strategy and masking rate.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence.

KEYWORDS

Task-oriented Dialogue System, Dialogue State Tracking, Entity, Adaptive Pre-training

ACM Reference Format:
Yukyung Lee, Takyoung Kim, Hoonsang Yoon, Pilsung Kang, Junseong Bang, and Misuk Kim. 2023. DSTEA : Improving Dialogue State Tracking via Entity Adaptive Pre-training. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (KDD'23).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

A task-oriented dialogue system (TOD), which aims to complete a certain task in a specific domain, such as restaurant reservation, can be modularized into four sub-tasks: natural language understanding, dialogue state tracking (DST), dialogue policy learning, and natural language generation [41]. Among them, DST, which seeks to track the structured belief state, plays an important role because the final quality of the entire dialogue system significantly depends on the accurate tracking of such belief states. However, as the dialogue system mainly focuses on the current dialogue turn, generating the correct belief state in a multi-turn dialogue is a challenging task [16].

There are two main directions of recent studies endeavoring to improve DST performance by generating the correct belief states: modifying the model structure or conducting additional pre-training using in-domain dialogue corpora. In the former, TRADE [35] and SOM-DST [16] were used in an attempt to accurately reflect the accumulated belief states by jointly training the encoder-decoder structure using a pointer network [25]. In addition, SST [5], GCDST [36], and CSFN-DST [44] were employed to provide rich information to the encoder by using the schema graph as an extra feature. In the latter, TOD-BERT [34] and DialoGLUE [20] performed additional pre-training based on masked language modeling on the encoder model (e.g., BERT [6]). In particular, DialoGLUE learned a representation suitable for the target domain by taking the same pre-training strategy using a large amount of external dialogue corpora before fine-tuning [10, 20]. In addition, models such as SimpleTOD [14] and SOLOIST [21] performed additional pre-training based on language modeling on the decoder model (e.g., GPT [22]) to capture TOD-related features.

Accepted to Second Workshop on Knowledge Augmented Methods for Natural Language Processing, in conjunction with KDD 2023.

KDD'23, August 06–10, 2018, Longbeach, CA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

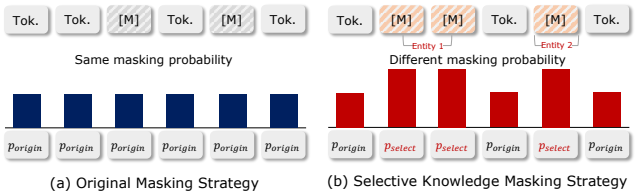

Figure 1: Comparison between original and proposed masking strategies. We use selective knowledge for effective pre-training on the DST. We give the masking probability by distinguishing between knowledge selected as important information (p_{select}) or relatively insignificant information (p_{origin}).

Although the aforementioned models have shown some positive effects on DST performance improvement, we believe that they can be further improved by considering the following points. First, language model pre-training can be enhanced by employing a customized strategy for the dialogue domain for the DST. Conversational models such as DialogGPT [40], TOD-BERT, and DialogBERT [9] were trained to capture textual and semantic features of dialogue context and showed significant performance improvement in response generation and response selection [9]. Therefore, DST performance improvement can also be achieved when training is based on a pre-trained language model specialized in the dialogue domain. Second, utilizing additional techniques to capture task-related knowledge effectively will also be helpful [10, 42]. DialoGLUE learned representations suitable for the target domain by performing additional pre-training using seven different datasets across four tasks: intent prediction, slot filling, semantic parsing, and DST. Similarly, PPTOD [26] used 11 datasets to train four tasks: natural language understanding, DST, dialogue policy learning, and natural language generation. Although this learning strategy is effective for performance improvement, computational cost issues arise because it requires a large amount of dialogue corpora. Therefore, finding an effective as well as efficient pre-training method that captures task-related knowledge from a DST dataset is necessary.

Note again that the main purpose of DST is to accurately extract the value assigned to a specific domain-slot pair from the dialogue utterance. Such an information extraction task can emphasize the semantics of important information by pre-training focusing on the entity feature of the input sequence. Accordingly, the performance of various natural language processing tasks can be improved [37]. Hence, we focused on the point that previous studies have rarely considered entities that provide rich information from dialogue utterances [30] (i.e., an entity that appears in one sentence, such as identifying people). In this paper, we introduce DSTEA (improving **D**ialogue **S**tate **T**racking via **E**ntity **A**daptive pre-training), a methodology in which the entity representations in DST models are intensively trained after extracting important information from a given utterance. Our approach is applicable to any BERT-based belief tracker and can enhance the tracking ability. We verified the effectiveness of the DSTEA using four strong DST models for various versions of MultiWOZ [4]. Experimental results showed that our training strategy had a positive effect on the performance improvement of the DST model and demonstrated the usefulness of entity-level information in multi-turn dialogue. Furthermore, additional analysis of slot-meta (domain-slot pair) information, and value, showed that the proposed DST model affords a lower error rate of predicted values than the existing models.

2 RELATED WORK

2.1 Task Adaptive Pre-training for DST

Although large-scale pre-trained language models have achieved remarkable successes in various natural language processing tasks, models specialized for target domains and tasks have been continuously studied. Adaptive pre-training refers to a process in which a language model trained in the general domain is pre-trained to learn the knowledge suitable for a specific domain or task. Beltagy et al. [2], Gururangan et al. [10], Lee et al. [18] proposed a training strategy for a specific domain or task by adding an adaptation phase between pre-training and fine-tuning. In the area of DST, some studies have attempted to conduct task adaptive pre-training. DialoGLUE employed ConvBERT, which was trained with large amounts of open-domain dialogues, and performed adaptive pre-training on the target dataset. In addition, ConvBERT-DG, which leveraged additional pre-training with seven DialoGLUE benchmarks, proved the surprising effect of self-supervised training. Further, SimpleTOD and SOLOIST performed pre-training through an auto-regressive objective with a GPT-based model and achieved high performance in DST. In particular, they achieved performance gain by training DST, action decision, and response generation together. In the case of Zhao et al. [43], Pegasus [39] pre-training objective were applied to T5 [23], and good performance was achieved without any pre/post-processing. Aug 15, 2022 10:23 PM

Inspired by the aforementioned methods, we propose a new adaptive pre-training method for DST. However, in contrast to existing methods, our strategy can enhance the performance through a modified masking strategy to further pre-train the target dataset without any external dialogue corpora.

2.2 Knowledge-enhanced Masking Strategy

Knowledge can not only be decomposed into different levels of granularity but also be divided into unstructured and structured knowledge [37]. Unstructured knowledge comprises entities and text, whereas structured knowledge refers to a predefined structure such as a knowledge graph or syntax tree. This knowledge is transferred to the pre-trained model in the form of rich information and can improve the performance of downstream tasks [31, 37]. Pre-training methods such as BERT learn general-purpose knowledge through random uniform token masking [19]. However, this masking strategy trains information about a single segment [15] and has a limitation that models cannot learn related sub-word tokens together. Therefore, several studies modified the original masking strategy of BERT and attempted to improve performance in various natural language processing downstream tasks through appropriate knowledge injection [37]. ERNIE [27] is a language model that contains entity information. It defines meaningful tokens, entities, and phrases as knowledge. Moreover, it was the first model to perform continual learning of high-level knowledge during the pre-training process. Further, SpanBERT [15] learns a continuous random span.

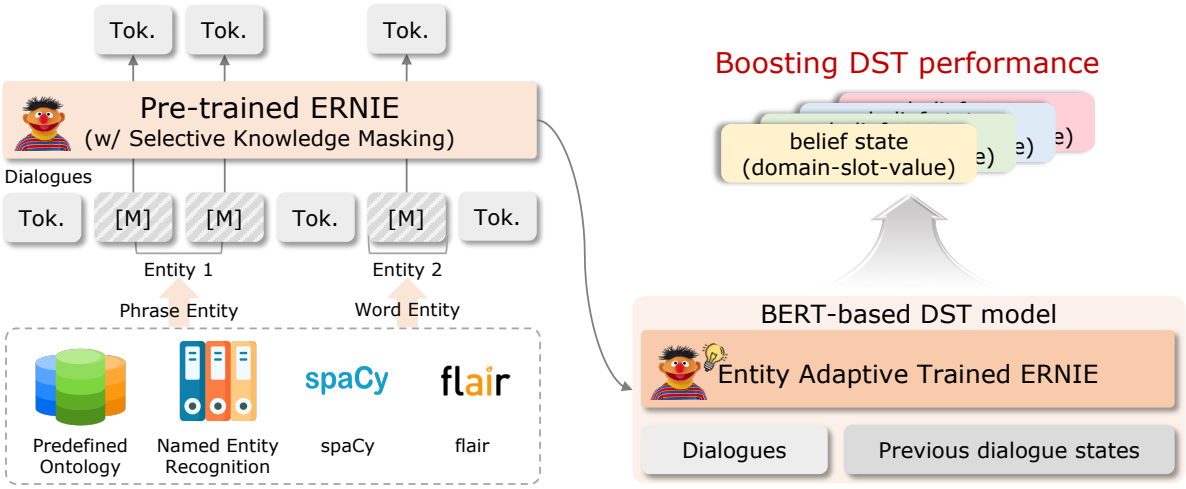

Figure 2: The architecture of DSTEA. Our method enriches word and phrase entity information to encoder during further pre-training with selective knowledge masking.

Because the span boundary representation is learned without relying on individual tokens, it is more advanced than the original BERT. Levine et al. [19] proposed pointwise mutual information masking that effectively conducts pre-training by simultaneously masking highly relevant n-gram tokens. Guu et al. [11] proposed REALM with salient-span masking to learn only named-entity and date information for question answering. Furthermore, Roberts et al. [24] considered entity information by applying salient-span masking to T5 [23]; it showed high performance in a question answering task. However, the method of properly injecting entity-level knowledge for DST has been relatively less investigated.

In this study, we found that the knowledge definition of ERNIE and the entity masking of REALM are well suited to the nature of dialogues. Inspired by this finding, we propose selective knowledge masking to focus on the important entities. In summary, we attempted to capture information suitable for DST through a new entity masking strategy after extracting fine-grained knowledge entities from dialogues using an entity-specific ERNIE-encoder.

3 PROPOSED METHOD

In this study, we propose DSTEA to intensively learn important knowledge about DST through entity adaptive pre-training. After extracting entities from the target dialogue data, DSTEA can capture DST-specific features by selectively training more focus on these entities.

The overall architecture of the DSTEA is shown in Figure 2. In particular, we assume that the entity information appearing in the utterance is significant knowledge for DST and apply entity adaptive pre-training using a selective knowledge masking strategy, as shown in Figure 1-(b). After applying this pre-training, DSTEA learns belief state tracking by utilizing previous DST models, such as SOM-DST, Trippy [12], SAVN [29] and STAR [8], which have achieved excellent performance and used BERT as an encoder.

3.1 Entity Adaptive Pre-training

The encoder architecture is an essential part of the DST model. The purpose of the encoder model in the proposed method is to learn an inductive bias suitable for DST during the pre-training process so that the representation of the pre-training model can be used to learn the dialogue information more accurately. The adaptive pre-training method proposed for DSTEA is shown in Figure 3. Pre-training comprises three steps: entity set construction, selective knowledge masking, and adaptive pre-training. After extracting entities from the utterance, a higher masking rate is assigned to them, while the original masking rate is assigned to the remaining tokens.

3.1.1 Entity Construction. One of the most important parts of entity adaptive pre-training is entity set construction. In this study, the entities were collected in four ways. First, entities were selected using ontology information. This is because an ontology is the most readily available form of information from the dataset and specifies the most important words. Second, after establishing a named-entity recognition (NER) model, the inference was conducted on the MultiWOZ dataset to collect the entities. An ERNIE-based entity tagging model was used, and the model is trained using the CONLL 2003 dataset [28]. Third, entities were extracted using the *spaCy* [13] library. The *spaCy* entity recognizer extracts entities in span units, including entity types such as location, language, person, and product. Finally, entities were extracted using the *flair* library [1]. The *flair* entity tagger is a model that is trained based on CONLL 2003 and extracts entities in span units. The entities extracted using these methods are a mixture of words and phrases. We attempted to learn these by distinguishing between word and phrase entities.

Word Entity First, each extracted entity was split into word units to compose a word entity. To prevent overfitting during this process, information about time and numbers was excluded from the entity. Because random times and numbers are used when constructing dialogue datasets [4], the appearance of unseen information during a dialogue may prevent the DST model from responding correctly

to unseen slots or values. In other words, if the entities are trained regarding time and numbers with selective knowledge masking, the biased model is highly likely to generate incorrect values. Therefore, these values were not considered. Additionally, stopwords from the *NLTK* [3] library were used to exclude words that were not helpful for training. Moreover, filtering was performed when the punctuation mark was extracted as an entity.

Phrase Entity To learn a phrase entity, cases that included an unknown token (i.e., [UNK]) in the phrase were excluded. Next, the phrase entities defined for each utterance were extracted in advance, and then pre-training was performed using randomly selected phrase entities. Even if a phrase entity included stopwords, filtering was not performed for masking in span units, and information about numbers and times was excluded, as for word entities. However, phrase entities extracted by the NER model and *flair* were of poor quality; therefore, entities were extracted using only the ontology and *spaCy*.

3.1.2 Selective Knowledge Masking. Selective knowledge masking is a method for learning important knowledge after selecting essential information from user and system utterances to inject an inductive bias suitable for DST. As shown in Figure 1, the previous DST model used an encoder such as BERT, which was trained by random masking without considering the characteristics of tokens. By contrast, we identified entities appearing in dialogues and assigned a higher masking rate to them while giving the original masking rate to non-entity tokens.

Selective knowledge masking is described in step 2 in Figure 3. The one-turn utterance token sequence is $U_t = (tok_1, tok_2, \ldots, tok_L)$, while the defined word entity is $Ent_{word} = (w_1, w_2, \ldots, w_N)$, the phrase entity is $Ent_{phrase} = (P_1, P_2, \ldots, P_M)$, the total length of the token sequence is L, the total number of word entities is N, and the total number of phrase entities is M. All tokens appearing in U_t have a masking probability, and whether each token is masked or not is determined by its masking probability before training a masked language model. $prob_{origin}$ refers to the masking probability for general tokens, whereas $prob_{select}$ refers to the masking probability for entity tokens. In the proposed DSTEA, $prob_{select}$ is always greater than $prob_{origin}$. The selective knowledge masking proposed in this study proceeds as follows. First, the masking probability of all tokens in every dialogue utterance is initialized to $prob_{origin}$. When a specific utterance includes a predefined entity, the masking probability of the entity token (Ent_{word}) or entity span (Ent_{phrase}) is changed to $prob_{select}$. After changing the masking probability of the token to $prob_{select}$, masking is performed in units of words and phrases.

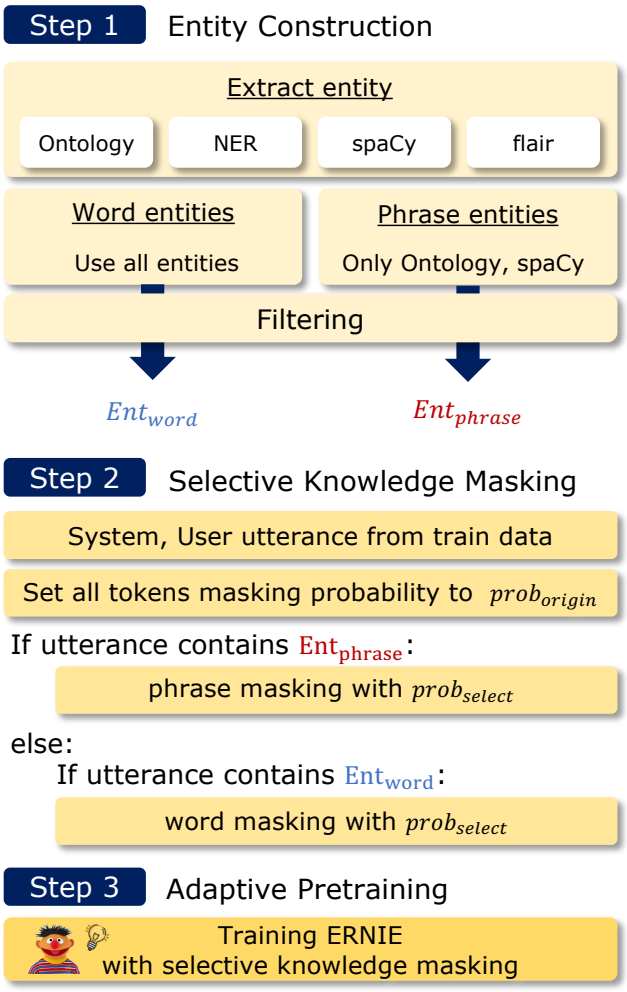

Figure 3: Selective Knowledge Masking Procedure. The proposed masking strategy comprises three steps. In step 1, word and phrase entities are extracted and filtered through ontology, NER, *spaCy*, and *flair*. In step 2, selective knowledge masking is performed using the extracted entity set. In step 3, entity adaptive pre-training is performed based on the masking probability newly defined in step 2.

4 EXPERIMENTAL SETTINGS

4.1 Datasets

Experiments were conducted using MultiWOZ 2.0, 2.1, and 2.2 [4, 7, 38], the most widely used datasets for DST. Similar to previous DST models, the experiments were conducted using five domains, namely, restaurant, train, hotel, taxi, and attraction. We followed the preprocessing procedures for each of the four baselines, most of which were provided by TRADE.

4.2 Baseline Models

Four BERT encoder structure-based baseline models were employed to evaluate the effectiveness of DSTEA: SOM-DST [16], Trippy [12], SAVN [29], and STAR [8]. Each model was trained according to publicly released implementations in the standard train/dev/test split of MultiWOZ.

SOM-DST introduced a state operation prediction that maintains the value of the previous slot instead of newly generating the value of every slot in each dialogue turn. In this model, the size of dialogue states is fixed size and some of dialogue states are selectively overwritten.

Table 1: Comparison of the proposed (+DSTEA) with four baseline models on the test sets of MultiWOZ 2.0, 2.1, and 2.2. The scores of joint goal accuracy (JGA), slot accuracy (SA), and relative slot accuracy (RSA) are computed. The underline indicates the best JGA for each dataset.

Model	MultiWOZ 2.0			MultiWOZ 2.1			MultiWOZ 2.2		
	JGA	SA	RSA	JGA	SA	RSA	JGA	SA	RSA
SOM-DST [16]	51.60	97.20	86.59	52.41	97.34	86.94	53.71	97.38	87.31
SOM-DST + DSTEA (ours)	**54.11**	**97.40**	**87.51**	**55.10**	**97.46**	**87.79**	**55.23**	**97.42**	**87.55**
Trippy [12]	52.63	97.13	86.56	52.63	97.20	87.98	53.38	97.20	88.45
Trippy + DSTEA (ours)	**52.96**	**97.18**	**86.80**	**54.87**	**97.33**	**88.50**	**54.05**	**97.31**	**88.81**
SAVN [29]	53.90	**97.43**	**86.33**	53.65	95.47	87.61	-		
SAVN + DSTEA (ours)	**54.19**	**97.43**	86.32	**54.75**	**97.52**	**87.94**			
STAR [8]	54.75	97.44	86.19	54.22	97.48	87.49	-		
STAR + DSTEA (ours)	**55.53**	**97.49**	**86.43**	**55.02**	**97.57**	**87.88**			

Trippy used three types of copy modules and classification gates, enabling the model to find values in the context of a conversation or the predictions of the previous turn.

SAVN utilized slot attention and value normalization. Slot attention improves span prediction performance by sharing information between slot and utterance, while value normalization can correct the extracted span based on ontology.

STAR utilized slot token and slot self-attention to capture a strong correlation between slots. These two self-attention operations learn the relationship between slots and values and find the value through distance-based slot value matching.

In this paper, the experimental results for MultiWOZ 2.2 are not reported in the case of SAVN and STAR because the performance fluctuated greatly owing to ontology update issues. The hyperparameters are described in detail in the supplementary material.

4.3 Adaptive Pre-training Settings

We trained the 'pre-trained ERNIE-2.0' on dialogue and used the huggingface transformers [33], with `'nghuyong/ernie-2.0-en'` as the ERNIE model. During the experiment, the masking probability $prob_{origin}$ was set to 0.2, and $prob_{select}$ was set to 0.4. The hyperparameters for adaptive and DST training are explained in detail in the supplementary material.

4.4 Evaluation Metrics

The performances of the baseline and the proposed models were evaluated according to the three following metrics: joint goal accuracy, slot accuracy, and relative slot accuracy. Joint goal accuracy [32] is an evaluation metric that verifies whether the predicted belief state exactly matches the gold label. Slot accuracy [32] is an evaluation metric that identifies the accuracy of slots among the predicted dialogue states. Relative slot accuracy [17] is a recently proposed metric that complements joint goal accuracy and slot accuracy. In contrast to slot accuracy, relative slot accuracy only focuses on the gold reference and predicted slots of the current dialogue instead of all predefined slots in slot accuracy.

Table 2: Domain-specific performance on the test sets of MultiWOZ 2.1. The score of joint goal accuracy for each domain is computed.

Model	Domain				
	Attraction	Hotel	Restaurant	Taxi	Train
SOM-DST	68.19	**49.22**	65.89	57.01	71.61
+ DSTEA	**70.65**	48.61	**69.88**	**58.57**	**73.38**
Trippy	72.17	43.45	68.48	**39.21**	70.48
+ DSTEA	**73.92**	**48.81**	**69.68**	35.53	**71.10**
SAVN	66.82	47.28	66.46	62.96	**75.51**
+ DSTEA	**66.87**	**49.25**	**69.28**	**65.91**	74.46
STAR	**68.57**	49.00	67.28	63.28	72.50
+ DSTEA	67.83	**49.09**	**68.42**	**71.03**	**74.32**

5 EXPERIMENTAL RESULTS

Since there are some differences in evaluating the baseline models in their source codes, we unified the evaluation processes using the same criteria to secure a fair comparison. All prediction results for each model and implementation codes can be found in the supplementary material.

5.1 DST Performance

The performances of DSTEA with four baseline models using MultiWOZ 2.0, 2.1, and 2.2 datasets are presented in Table 1. The joint goal accuracy improved in all datasets for all baseline models. More specifically, STAR + DSTEA in MultiWOZ 2.0 and SOM-DST + DSTEA in MultiWOZ 2.1 and 2.2 recorded the best performance. In particular, when DSTEA's entity adaptive pre-training was applied to SOM-DST, the most remarkable performance improvement

Table 3: Performance comparison table for adaptive pre-training of MultiWOZ 2.1. The strategies for three masking strategies were tested based on SOM-DST(ERNIE).

Model	MultiWOZ 2.1	
	JGA	SA
SOM-DST (BERT)	52.41	97.34
SOM-DST (ERNIE)	53.97	97.40
+ Random Masking ($P_{origin} = 0.2$)	52.70	97.27
+ Random Masking ($P_{origin} = 0.4$)	54.24	97.42
+ Selective Knowledge Masking (ours) ($P_{select} = 0.4$, $P_{origin} = 0.2$)	**55.10**	**97.46**

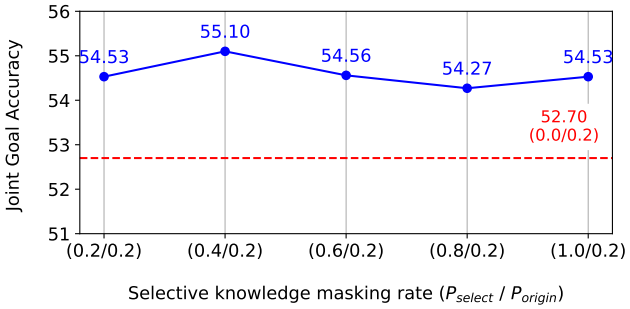

Figure 4: Performance on DSTEA of MultiWOZ 2.1 test set with different selective knowledge masking rates.

(+2.69) was recorded in MultiWOZ 2.0. These performance improvements confirm that an effective representation can be learned by training entities, essential information for the dialogue domain, more intensively than other tokens. Because joint goal accuracy accepts the prediction as correct only when the accurate belief state is predicted over all dialogue turns, it is the strictest evaluation metric. Significant performance improvement in terms of joint goal accuracy implies that the entity adaptive pre-training is sufficiently effective. In addition to joint goal accuracy, both slot accuracy and relative slot accuracy are also improved in all cases except for SAVN in MultiWOZ.

Table 2 shows the domain-level performance of DSTEA with the baseline models. Significant performance improvements were observed in most domains by applying DSTEA. The results for all slots per each model are described in detail in the supplementary material.

5.2 Effectiveness of Adaptive Pre-training

To verify the effectiveness of the proposed selective knowledge masking, its performance was compared by changing the masking probability of the original masked language modeling. This comparative experiment was conducted using the SOM-DST model, which

Table 4: Performance according to the entity type of MultiWOZ 2.1. Comparison of individual performances of word and phrase entities according to the entity module. The underline indicates the highest score for each word-level and phrase-level entity type.

Model	Type	Entities	MultiWOZ 2.1
			JGA
SOM-DST + DSTEA	**Word Level**	**Ontology Only**	54.54
		Entity (spaCy) Only	54.32
		Entity (NER) Only	54.42
		Entity (flair) Only	53.90
		Combine Words	<u>54.79</u>
	Phrase Level	**Ontology Only**	54.47
		Entity (spaCy) Only	54.31
		Combine Phrases	<u>54.83</u>
	Combine All	**Combine All**	**55.10**

showed the highest performance improvement in the abovementioned experimental results. Table 3 shows how the performance is affected by the adaptive pre-training setting. **SOM-DST (ERNIE)** indicates that the encoder of SOM-DST is changed to ERNIE, and **+ Random Masking ($P_{origin} = \alpha$)** represents the case of using random masking probability α during adaptive pre-training. P_{select} and P_{origin} of **Selective Knowledge Masking (ours)** are the masking probabilities of the entity tokens and the masking probability of the remaining tokens, respectively. The results showed that our strategy yielded the best performance. Notably, the performance of SOM-DST was improved by simply applying ERNIE, implying that the ERNIE encoder itself assuredly helps to extract the correct belief state. With respect to the masking probability, the performance was lower than that of **SOM-DST (ERNIE)** when the masking probability was set to 0.2. An interesting observation is that the DST performance can be improved by only increasing the masking probability for all tokens. However, our selective masking strategy outperformed **+ Random Masking (P_{origin} = 0.4)**, validating the greater effectiveness of the proposed selective-knowledge-masking method compared with simply increasing the masking probability.

Figure 4 shows the performance of DSTEA according to the change in the selective knowledge masking rate. The red dashed line represents the pre-training model that does not consider entities, which can be understood as the lower bound performance. The blue line indicates the performance of DSTEA according to different P_{select} ratios. Selective knowledge masking clearly enhanced the DST performance regardless of P_{select} ratio. The best joint goal accuracy was reported when $P_{select} = 0.4$, but the worst case still yielded a significantly improved joint goal accuracy compared to that without selective knowledge masking.

5.3 Effectiveness of Entity Types

In this section, we discuss how the entity extraction method affects the final DST performance. The entities were extracted using four modules, and both word and phrase units were considered. Table 4

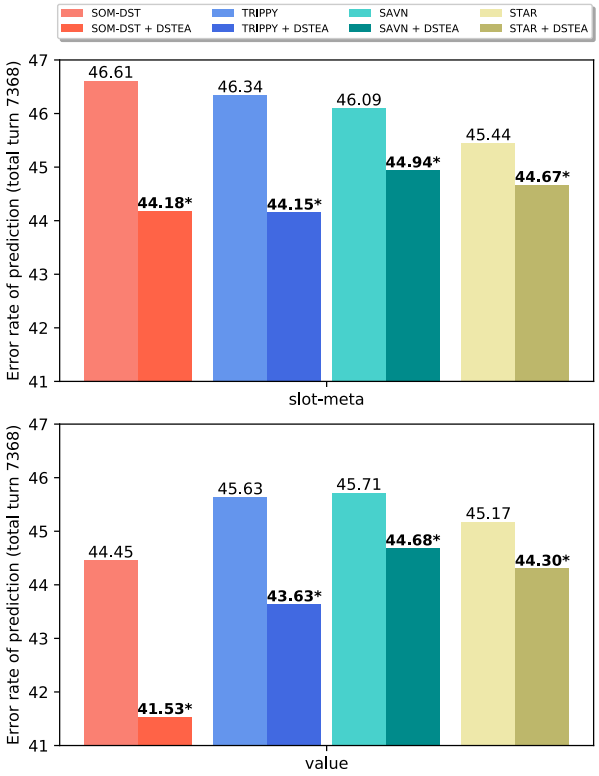

Figure 5: Error rate between the baseline models (SOM-DST, Trippy, SAVN, STAR) and the DSTEA-applied models for slot-meta and value in MultiWOZ 2.1 (Total turns = 7368). The error rate is measured when there is a mismatch between the ground truth and the predicted value; thus, a lower value means an improved model.

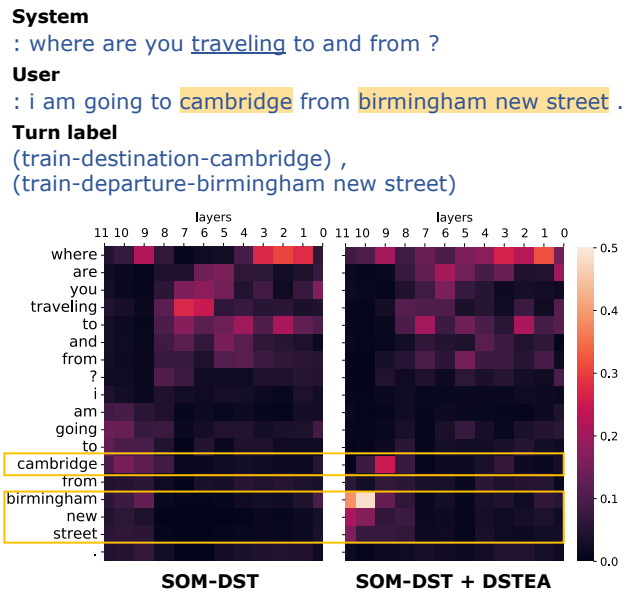

Figure 6: Visualization of attention weights between SOM-DST and SOM-DST + DSTEA. The heatmap shows the average attention weight for all layers for the word 'traveling' (MultiWOZ 2.1 MUL0671 turn 1).

shows the joint goal accuracy on MultiWOZ 2.1 with consideration of different entity types. Among the four individual extraction modules, the ontology-based entity extraction was found to be the best for both word level and phrase levels. For word-level entities, other extraction methods also showed good performances, but the *flair* library-based entity extraction afforded a slightly lower joint goal accuracy compared with the other three extraction modules. Note that when entities extracted by all four modules are combined, the joint goal accuracy was even improved than the single best extraction module for both word-level and phrase level. Moreover, when both word and phrase entities were aggregated, the best joint goal accuracy of 55.10 was achieved. Based on these results, entity set construction is also very important for DST performance improvement in addition to an appropriate learning strategy during the pre-training.

5.4 Slot-meta and Value Error Rate

The DST model generates prediction values comprising domain-slot values for each turn. We compared the degree of error rate between the baseline models and DSTEA with respect to the slot-meta (domain-slot pair) and value. Figure 5 shows that DSTEA effectively reduces the errors compared to the baseline models, which confirms that the DSTEA can accurately detect the domain and slot information of the current utterance and can appropriately predict the value for each slot. These results also support the fact that the model trained using DSTEA can learn the appropriate inductive bias to improve the final performance.

5.5 Attention Visualization

In this part, we investigate the difference in attention weights between the baseline model and DSTEA. Figure 6 is a heatmap of the average attention weights for the word 'traveling' when the first dialogue turn of MUL0671 is used as the input to both models. We can observe that much higher attention weights are assigned to the entities in the user utterance, especially to the proper nouns ('cambridge' and 'birmingham'), by the DSTEA, supporting that the proposed entity adaptive pre-training results in focused concentration on informative entities. Note also that these higher attention weights appear in the upper layers of the encoder so that this important information is certainly delivered to the decoder in the case of SOM-DST to answer the slot-meta and value correctly.

6 CONCLUSION

In this study, we propose an entity adaptive pre-training framework, named DSTEA, assuming that the essential knowledge of DST will be well captured if the pre-training of the language model focuses on informative entity tokens more intensively than others. In DSTEA, an entity-specialized language model, ERNIE, was employed for pre-training, while selective knowledge masking strategy is proposed to learn word and phrase entities more frequently than

non-entities. Experimental results on four DST models show that the proposed DSTEA framework improved the baseline models in terms of JGA, SA, and RSA for three versions of the MultiWOZ dataset. Most entity extraction methods help to improve the DST performance, and combining the four extraction methods for both word and phrase entities yielded the best performance. We also verified that selective knowledge masking is more appropriate than simply increasing the masking rate to all types of tokens. The effect of DSTEA was also confirmed by the attention heatmap in that the informative entities were given higher attention weights with the DSTEA than without the DSTEA. We expect this pre-training method to be used effectively for DST under insufficient input data and entity-related tasks.

7 FUTURE WORKS

In this paper, we have provided several contributions and discoveries in relation to our proposed model, DSTEA. However, there exist numerous prospective research trajectories that could further elevate the performance and practicability of DSTEA.

One of the key challenges to address in future works is the requirement for knowledge about all entities in various situations such as zero-shot, and few-shot. Our current model depends on comprehensive dialogue state tracking (DST) labels, and acquiring these labels can prove onerous and often impossible in real-life scenarios. Yet, our model shows a promising capability in extracting entities from unlabeled corpora, offering a feasible solution for situations where DST gold labels are unavailable. Based on this concept, we propose utilizing our entity extraction, filtering, and masking techniques as a form of weak supervision, particularly in zero-shot, few-shot settings, or domains with extremely limited data. This strategy will enable our model to operate efficiently in scenarios involving unseen data and incomplete DST details.

Moreover, although our DSTEA model is constructed upon BERT, we would like to underscore its adaptability with other transformer-based architectures, such as BART and T5. We propose to extend our masking method to include text-infilling, thereby making our model compatible with GPT-based structures. This adaptability opens avenues for researchers and practitioners to employ our model across a diverse array of dialogue state tracking applications, encouraging wider acceptance and comparability in this research area.

Additionally, our present study's primary focus lies in utilizing DST as the fundamental component within a modular approach (NLU-DST-DP-NLG). This selection allows us to delve deeply into the complexities of dialogue state tracking. Future studies could investigate the advantages of integrating multi-task learning or other tasks within the proposed framework. By harnessing external corpora and implementing an end-to-end approach, we can potentially enhance the performance and robustness of the overall dialogue system. Therefore, exploring the potential of multitask learning and considering additional tasks within the modular design represents a promising pathway for future investigations.

By addressing these prospective research directions, we anticipate significant advancements in the field of dialogue state tracking. This, in turn, will enhance the functionality, adaptability, and overall performance of DSTEA and other similar models, rendering them even more useful in real-world applications.

REFERENCES

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. 54–59.

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. https://doi.org/10.18653/v1/D19-1371

[3] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[4] 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*. Association for Computational Linguistics, Brussels, Belgium, 5016–5026. https://doi.org/10.18653/v1/D18-1547

[5] 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. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 7521–7528. https://doi.org/10.1609/aaai.v34i05.6250

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. 2019. MultiWOZ 2.1: Multi-Domain Dialogue State Corrections and State Tracking Baselines. *arXiv preprint arXiv:1907.01669* (2019).

[8] Ye Fanghua, Manotumruksa Jarana, Zhang Qiang, Li Shenghui, and Yilmaz Emine. 2021. Slot Self-Attentive Dialogue State Tracking. In *The Web Conference (WWW)*.

[9] Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. DialogBERT: Discourse-Aware Response Generation via Learning to Recover and Rank Utterances. *In Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)* (2021).

[10] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8342–8360. https://doi.org/10.18653/v1/2020.acl-main.740

[11] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. *ArXiv* abs/2002.08909 (2020).

[12] Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gašić. 2020. TripPy: A Triple Copy Strategy for Value Independent Neural Dialog State Tracking. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, 35–44.

[13] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[14] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796* (2020).

[15] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics* 8 (2020), 64–77. https://doi.org/10.1162/tacl_a_00300

[16] 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*. Association for Computational Linguistics, Online, 567–582. https://doi.org/10.18653/v1/2020.acl-main.53

[17] Takyoung Kim, Hoonsang Yoon, Yukyung Lee, Pilsung Kang, and Misuk Kim. 2022. Mismatch between Multi-turn Dialogue and its Evaluation Metric in Dialogue State Tracking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Dublin, Ireland, 297–309. https://doi.org/10.18653/v1/2022.acl-short.33

[18] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* (sep 2019). https://doi.org/10.1093/bioinformatics/btz682

[19] Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2021. {PMI}-Masking: Principled masking of correlated spans. In *International Conference on Learning Representations*. https://openreview.net/forum?id=3Aoft6NWFej

[20] S. Mehri, M. Eric, and D. Hakkani-Tur. 2020. DialoGLUE: A Natural Language Understanding Benchmark for Task-Oriented Dialogue. *ArXiv* abs/2009.13570 (2020).

[21] Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *Transactions of the Association for Computational Linguistics* 9 (2021), 807–824. https://doi.org/10.1162/tacl_a_00399

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[24] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model?. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 5418–5426. https://doi.org/10.18653/v1/2020.emnlp-main.437

[25] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1073–1083. https://doi.org/10.18653/v1/P17-1099

[26] Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-Task Pre-Training for Plug-and-Play Task-Oriented Dialogue System. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 4661–4676. https://doi.org/10.18653/v1/2022.acl-long.319

[27] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 8968–8975. https://doi.org/10.1609/aaai.v34i05.6428

[28] Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 142–147. https://aclanthology.org/W03-0419

[29] 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 (EMNLP)*. Association for Computational Linguistics, Online, 3019–3028. https://www.aclweb.org/anthology/2020.emnlp-main.243

[30] Yijun Wang, Changzhi Sun, Yuanbin Wu, Junchi Yan, Peng Gao, and Guotong Xie. 2020. Pre-training Entity Relation Encoder with Intra-span and Inter-span Information. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 1692–1705. https://doi.org/10.18653/v1/2020.emnlp-main.132

[31] Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew Arnold. 2021. Knowledge Enhanced Pretrained Language Models: A Compreshensive Survey. https://doi.org/10.48550/ARXIV.2110.08455

[32] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference*. Association for Computational Linguistics, Metz, France, 404–413. https://aclanthology.org/W13-4065

[33] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[34] Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained Natural Language Understanding for Task-Oriented Dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 917–929. https://doi.org/10.18653/v1/2020.emnlp-main.66

[35] 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*. Association for Computational Linguistics, Florence, Italy, 808–819. https://doi.org/10.18653/v1/P19-1078

[36] Peng Wu, Bowei Zou, Ridong Jiang, and AiTi Aw. 2020. GCDST: A Graph-based and Copy-augmented Multi-domain Dialogue State Tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 1063–1073. https://doi.org/10.18653/v1/2020.findings-emnlp.95

[37] Jian Yang, Gang Xiao, Yulong Shen, Wei Jiang, Xinyu Hu, Ying Zhang, and Jinghui Peng. 2021. A survey of knowledge enhanced pre-trained models. *arXiv preprint arXiv:2110.00269* (2021).

[38] Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. MultiWOZ 2.2: A Dialogue Dataset with Additional Annotation Corrections and State Tracking Baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*. 109–117.

[39] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 11328–11339. https://proceedings.mlr.press/v119/zhang20ae.html

[40] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 270–278. https://doi.org/10.18653/v1/2020.acl-demos.30

[41] Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences* 63, 10 (2020), 2011–2027.

[42] Zhuosheng Zhang and Hai Zhao. 2021. Structural Pre-training for Dialogue Comprehension. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 5134–5145. https://doi.org/10.18653/v1/2021.acl-long.399

[43] Jeffrey Zhao, Mahdis Mahdieh, Ye Zhang, Yuan Cao, and Yonghui Wu. 2021. Effective Sequence-to-Sequence Dialogue State Tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 7486–7493. https://doi.org/10.18653/v1/2021.emnlp-main.593

[44] Su Zhu, Jieyu Li, Lu Chen, and Kai Yu. 2020. Efficient Context and Schema Fusion Networks for Multi-Domain Dialogue State Tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 766–781. https://doi.org/10.18653/v1/2020.findings-emnlp.68