Pre-training and Data Augmentation for Dialogue State Tracking

Stanford CS224N Custom Project

Paridhi Maheshwari, Sharmila Reddy Nangi
Department of Computer Science
Stanford University
{paridhi, srnangi}@cs.stanford.edu

Abstract

Dialogue State Tracking (DST) is an important component of task-oriented dialogue systems, which involves tracking of user's goal and slot values based on dialogue context. While large-scale pre-trained language models like BERT and T5 have proven to achieve state-of-the-art performance for DST, these models are pre-trained on generic text datasets from the web and require huge amounts of training data to adapt to the task of dialogue understanding. In this work, we address these challenges by exploring dialogue-focused pre-training strategies for different transformer architectures. Additionally, we augment the training data with several rule-based as well as deep learning networks to enable the robustness of these models. Our detailed experiments show that baseline models benefit immensely when pre-trained with our span-level objectives in multiple phases. Our experiments also suggest that the use of data augmentation techniques such as paraphrasing also improve the performance on DST.

1 Key Information to include

- Mentor: Ethan A. Chi
- External Collaborators: No
- Sharing project: No

2 Introduction

Task-Oriented Dialogue Systems (TODS) form the backbone of various conversational AI assistants such as Alexa, Siri and Google Assistant. It is a challenging NLP task that focuses on solving specific user goals such as travel planning or restaurant reservation. This requires an understanding of the internally complex human language and the ability to maintain an engaging dialogue to achieve a certain task. A typical dialogue system consists of 4 major steps: Natural Language Understanding (NLU) [1], Dialogue State Tracking (DST) [2], Dialogue Policy Learning [3] and Natural Language Generation (NLG) [4]. The aim of DST is to calibrate the dialogue states based on the current utterance and dialogue history. It is imperative for the DST component to make accurate predictions as the rest of the pipeline is heavily reliant on its output. Therefore, DST is a crucial intermediate task to enhance the overall performance of TODS, and will be the focus of our work.

The dialogue state is a structured mechanism to track the users' intentions at every step of the dialogue in the form of slot-value pairs. We illustrate an example in Table 1, which shows the annotated state after every turn in a conversation. The set of all possible slots and corresponding entities are pre-defined using an ontology. Although it might appear to be a simple task of finding mentions of ontology entities in user utterances, this is hardly the case. It is a complex task owing to the lexical and linguistic variations in language (for example, rephrases such as 'affordable', 'budget', 'low-cost', 'economic' etc for the slot *restaurant-price=cheap*), and multi-turn dynamics where model needs to infer slot-value pairs from previous context (as illustrated in the last turn of Table 1).

Dialogue State Tracking is an active research direction in both academia and industry [5]. There have been several attempts to solve it [6, 7] and the state-of-the-art methods achieve good overall performance. However, all the proposed approaches are highly data-intensive and require large amounts of conversational data to build better systems. Publicly available datasets for dialogue tasks usually contain only a few thousand dialogues, and literature also suggests that the conversations tend to follow a set pattern without much variations [8]. Lack of sufficient training data poses a serious limitation in learning these models to their full potential. While pre-trained Language Models (LM) offer a promising alternative, there exists a significant domain gap between free-flowing text on the web and multi-turn, goal-driven conversations [9], and this makes the adaptation of LMs to dialogue tasks non-trivial. In this work, we try to address this data scarcity problem using a two-pronged approach as follows:

Sys:	Hi, what can I do for you?
User:	Please find me a Chinese restaurant.
State:	*restaurant-food=chinese*
Sys:	Inchin fits your criterion, can I book it?
User:	Yes, I need a table on Monday at 12:15
State:	*restaurant-food=chinese; restaurant-name=inchin; restaurant-day=monday; restaurant-time=12:15*
Sys:	Booking is successful. Anything else I can assist with?
User:	I need a taxi to get to the restaurant on time.
State:	*restaurant-food=chinese; restaurant-name=inchin; restaurant-day=monday; restaurant-time=12:15; taxi-destination=inchin; taxi-arrive=12:15*

Table 1: Example dialogue spanning multiple domains. The slots for *taxi* domain in the last turn need to be inferred from context.

- First, we propose to utilize large-scale dialogue datasets for pre-training of language models. By leveraging unsupervised objectives, we enable the use of abundant conversation data without requiring labels and adapt the language models for dialogues.
- Second, we explore several data augmentation techniques in NLP to increase the labeled data for downstream DST task. We use both rule-based augmentations as well as deep learning techniques to generate dialogue variations. Such automatic and inexpensive ways to rephrase utterances also provide diversity and variability in language, improving generalization capability.

We present a thorough quantitative evaluation of our methods on the MultiWOZ datasets [10, 11] and corroborate with qualitative experiments where possible. The results demonstrate improved performance on DST task, indicating the viability of our approach. We also perform a detailed analysis of our model predictions and highlight the type of dialogues which are prone to error.

3 Related Work

3.1 Dialogue State Tracking

Traditional methods in Dialogue State Tracking used heuristic feature extractors and pre-defined ontology [12] to identify slots-value pairs in the dialogue, but in the past few years, neural models have taken over. Recurrent neural networks with attention-based copy mechanisms [13] and the very recent transformer architectures are used to encode the user utterance and predict the start and end positions for slot values. TripPy [6] uses a BERT-based architecture to encode the current dialogue context and employs a triple copy strategy, which allows it to copy values from the context, previous turns' predictions and system informs. This model provides state of the art performances for DST on multiple datasets. MinTL [7] proposed a framework to efficiently utilise the knowledge from pre-trained generative models like T5 and BART for the task of DST and response generation. Unlike other models that predict the dialogue state, MinTL generates the change in the dialogue state as a Levenshtein belief state. This unique approach showed more robust results in low resource domains. As an alternative to span predictions, some works have also modelled DST as a generative task, where Hosseini et al. [14] directly fine-tuned GPT-2 for dialogue state and response generation. While these prior work demonstrated the usability of large LMs, they do not venture further into the ideas of data augmentation and adaptive pre-training, which forms the core of our work.

3.2 Adaptive Pre-training for Dialogues

Though large-scale pre-training results in strong performance when transferring to downstream tasks, performing self-supervised training on a target dataset allows the model to better adapt to the dataset prior to fine-tuning [15]. In dialogue domain, when trained with large amounts of open-domain dialogues, DialoGPT [16] showed better context consistency in responses and ConveRT demonstrated significant performance improvements over BERT on both intent prediction [17] and slot filling [18]. Inspired by the success of these methods, we experimented with multi phase adaptive pre-training and explored the usage of span level corruption strategies for the setting of multi-domain dialogues.

3.3 Data Augmentation

These techniques aims to generate new training data by conducting transformations on existing data. It has been widely used in computer vision (e.g. rotating or flipping images) [19], but relatively under explored in NLP, perhaps due to the challenges posed by the discrete nature of language [20]. Recently, data augmentation has gained importance for NLP tasks which deal with low-resource domains and large-scale neural networks that require huge amounts of training data. The most common methods involve word-level perturbations or sentence-level transformations with neural models. In the dialogue domain, Louvan et al. [21] propose a lightweight augmentation method based on word/token substitutions for slot filling and Hou et al. [22] present a seq2seq framework to augment dialogue utterances for dialogue language understanding. Most augmentation methods are not widely applied for the challenging MultiWOZ dataset, and this is one of our major contributions.

4 Approach

4.1 Multi-phase Adaptive Pre-training

Training large-scale models on abundant text data crawled from the web have proven to be a very powerful tool in NLP, improving the performance on numerous downstream tasks such as question-answering and sentiment classification [23]. These pre-training methods aim to learn contextual word embeddings using unsupervised learning objectives. Common strategies include Masked Language Modelling (MLM) where inputs tokens are randomly replaced by `<mask>` and model learns to predict the masked tokens, and Next Sentence Prediction (NSP) that takes in two sentences to determine if the latter is the actual subsequent sentence to the former sentence.

Dialogue State Tracking (DST) is an intermediate step of dialogue systems, which involves tracking the user's goal and slot values as the conversation progresses. Similar to the other NLP tasks, DST has also benefited from the advances of pre-trained language models [15]. But because these models are trained on free-flowing text, they are not particularly good at representing goal-driven, multi-turn dependencies. Recent works have shown that they are limited in their ability to model structural context of dialogues [9], especially when it comes to language understanding tasks. To overcome this, we propose to incorporate multi-phase adaptive pre-training on span-level objectives in existing frameworks for dialogue state tracking.

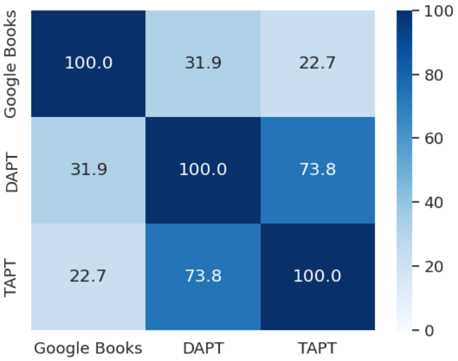

Figure 1: Vocabulary overlap (%) between Google Books corpus and the datasets used for domain and task adaptive pre-training, DAPT and TAPT, respectively. Vocabulary is defined as the top 10K most frequently occurring words.

Inspired by Gururangan et al. [24], we adopt a multi-phase adpative pre-training. First, we curate a large open-domain dialogue dataset by collecting task-oriented dialogues from multiple sources. We believe this *Domain Adaptive Pre-training (DAPT)* will learn more meaningful representations for user and system utterances. Next, we perform *Task Adaptive Pre-training (TAPT)* on the target dataset before fine-tuning on the downstream task. This continued pre-training mitigates the domain mismatch between training corpora and task domain [25] and also adapts language models across tasks. To illustrate the extent of difference between dialogues and free-flowing text, we analyse the vocabulary overlap in Figure 1. We use Google Books dataset as a proxy for conventional language modelling dataset. We observe that there is a strong overlap between DAPT and TAPT datasets (both consist of dialogues but from varied domains), but they are far more dissimilar to vocabulary used in books. This is a simple indication of the necessity and potential of adaptive pre-training.

The multi-turn setting of DST naturally involves reasoning and inferring relationships across two or more utterances. For example, in the user dialogues "Find me a movie theatre" and "I am staying at the Lensfield Hotel", it is crucial for the model to understand that the user is looking for theatres near the specified hotel. To facilitate reasoning across spans of text, we leverage span-level pre-training objectives [26]. Unlike traditional MLM, we randomly mask contiguous tokens and predict the entire span. This strategy has also proven to be beneficial for span-selection tasks, i.e, where entities are sequence of contiguous words (such as `Cambridge Arts Theatre`) rather than single tokens [27]

We evaluate our method on two popular transformer architectures, encoder-based BERT [28] and generative model T5 [29]. For BERT, we work with the TripPy [6] model for DST and compare masked language modelling and span-prediction objectives. For T5, we employ the work of Lin et al. [7] as our baseline and experiment with the span corruption objective.

4.2 Data Augmentation for Dialogue Understanding

For dialogue tasks, acquiring labeled data through human annotations is an expensive and time-consuming task. It is even more challenging due to human errors and inconsistencies in convention and normalization [30]. Even the MutiWOZ dataset has significant discrepancies despite being curated through multiple rounds of annotation as well as several iterations of the dataset release [10, 11]. This poses serious problems while training dialogue systems, which are especially data hungry.

We try to alleviate this data shortage using various data augmentation techniques. This enables us to increase the size and variance (diversity) of the training data, without the acquisition cost. We consider rule-based approaches such as entity replacement, crop and rotate using dependency parse trees, and sequential augmentation to increase complexity. We also experiment with deep learning techniques such as sequence-to-sequence models for paraphrasing and Neural Machine Translation (NMT). These strategies enhance the language structure and linguistic variability of utterances so that the model avoids memorizing templates, which also provides diversity for better generalization.

Original Utterance	Can you find an Indian restaurant for me that is also in the town centre ?
Entity Replacement	Can you find an Mexican restaurant for me that is also in the town east ?
Crop	find an Indian restaurant that is in the town centre
Rotate	an Indian restaurant for me that is also in the town centre find you
Sequential	Can you find an Indian restaurant for me that is also in the town centre ? I want to make a reservation for two people.
Paraphrase	I want to go to an Indian restaurant in the centre of the town.
Translate	I am looking for an Indian restaurant that is also in the city center ?

Table 2: Example of data augmentations for a given sample utterance. Entity replacement is a value-based augmentation where the sentence structure remains same. Other augmentations alter the context but the dialogue state remains unchanged as the users' end-goal (slot-value pairs) is the same.

We illustrate the augmentation process in Table 2, where we show the original utterance and its modifications using different methods, and delineate the augmentation techniques as follows:

- **Entity Replacement**: We take advantage of the ontology (all possible entities for a given slot) along with the slot-value pairs for every utterance. For a given slot label, we randomly sample a different entity from the ontology and edit the text in user utterance and the ground truth dialogue state. Consider Table 2, where we replaced "Indian → Mexican" for *restaurant-food* and "centre → east" for *restaurant-area*. Note that editing multi-turn dialogues is a challenging task as we have to maintain consistency in the slot substitutions across all the turns in the dialogue.
- **Crop and Rotate**: Following [31], we use the dependency parse tree to play around with the language structure of dialogues. Crop focuses on particular fragments of a sentence (e.g., subject

and predicate, or object and predicate), and removes the rest of the fragments, including its sub-tree, to create a smaller sentence. Rotate aims to rotate the target fragment of a sentence around the root of the dependency parse structure, producing a new variation of the utterance.

- **Sequential**: For each turn, we concatenate the future user and system utterances of upto η consecutive turns. The hyperparameter η controls the level of complexity of the augmented dialogue. This has been shown to improve the generalizability of dialogue models [32] due to deeper intent understanding and more ground truth labels for effective training.
- **Paraphrase**: We use Pegasus [33], a sequence-to-sequence transformer model, for generating paraphrases of all user utterances in the dialogue. The model was originally trained on abstractive summarization using a novel self-supervised objective gap-sentences generation, and we employ a pre-trained network which was fine-tuned for paraphrasing. We use a simple heuristic of generating multiple sentences using beam search and picking the longest sentence.
- **Translate**: We use Google's machine translation model [1] to convert English utterances to another language, and then run back-translation to generate context paraphrases of the input sentence. We select Spanish as our target language due to its popularity and ease of reproducibility.

5 Experiments

Data: We work with the Multi-Domain Wizard-of-Oz (MultiWOZ) dataset, a prominent choice for task-oriented dialogue systems. It contains 8438 / 1000 / 1000 dialogues for train / validation / test sets respectively. The dialogues involve multi-turn utterances along with the dialogue state (a dictionary of slot-value pairs such as `Location:San Francisco`) at every turn. The dataset spans across several domains such as `restaurants`, `hotels`, `trains` etc, and over 30 domain-slot pairs. Each domain is defined by its own ontology (set of slots and entities). The task is to correctly predict this dialogue state based on the user utterance and dialogue history. There are two versions for this dataset - MultiWOZ 2.0 [10] and MultiWOZ 2.1 [11] - and we test our approach on both.

For the first phase of pre-training (DAPT), we combine MultiWOZ with 6 other datasets from the DialoGLUE benchmark [34], namely BANKING77 [35], CLINC150 [36], HWU64 [37], RESTAURANT8K [18], DSTC8 [38], TOP [39]. The datasets comprise of 4 tasks - DST, intent prediction, slot filling, semantic parsing - all sharing the common goal of understanding language in dialogues.

Evaluation Metrics: The standard metric for this task is *Joint Goal Accuracy* which is the percentage of dialogue turns for which all slots were filled correctly. Additionally, we also compute *Slot Accuracy*, which is the ratio of successful slot value predictions among all the ground-truth slots of a dialogue turn. We also report the *Slot F1* score for the slot values across all turns of all dialogues in the dataset.

Experimental Details: We use the PyTorch implementations for the two baselines and the hyper-parameter settings mentioned by the authors. We implement our pre-training methods using the transformers library of Hugging Face. We initialize the weights from large-scale pretrained models and tokenizers (`bert-base-uncased` and `t5-small`). We set the masking probability to 0.15 and for span-based objectives, we randomly sample the span length between 1 to 5 tokens. We optimize using the default values in Hugging Face - initial learning rate of $5e^{-5}$ with AdamW [40]. Finally, we use a block size of 128, batch size of 32 and train for 3 epochs ($\sim$ 20 minutes per epoch). For sequential data augmentation, we set the hyperparameter η to be 3. We provide the final statistics of the augmented data in Table 8. The main contribution of our work - pre-training pipelines with span-level objectives; various rule-based and deep learning methods for data augmentation - were implemented by us from scratch. Our code is available at https://github.com/paridhimaheshwari2708/DialogueSystems.

6 Results

Does span-level pre-training improve performance? (refer Table 3) For BERT, span prediction objective consistently outperforms masked language modelling for pre-training. Therefore, masking span of tokens forces the model to learn better representations from the entire context.

Does multi-phase pre-training on dialogue datasets help? (refer Table 3) We notice that the DST performance for BERT, in terms of JGA, improves when we pre-train with MLM objective on

[1] https://translate.google.com

Model	Objective	Pre-training	MultiWOZ 2.0			MultiWOZ 2.1		
			JGA	S F1	SA	JGA	S F1	SA
BERT	MLM	Base	**51.2**	75.2	**92.6**	53.5	**75.8**	**92.2**
		+ DAPT	50.9	74.9	92.5	**54.0**	75.7	**92.2**
		+ TAPT	49.8	**75.3**	**92.6**	51.6	75.4	92.1
		+ DAPT + TAPT	49.1	74.7	92.5	52.0	75.5	92.1
BERT	Span Prediction	Base	51.2	**75.2**	**92.6**	53.5	75.8	92.2
		+ DAPT	**52.1**	75.1	**92.6**	54.6	**76.3**	**92.3**
		+ TAPT	50.7	**75.2**	92.5	**54.9**	75.9	92.2
		+ DAPT + TAPT	51.7	75.1	**92.6**	53.6	75.9	92.2
T5	Span Corruption	Base	50.2	89.6	96.5	50.3	90.2	96.3
		+ DAPT	50.2	89.4	96.4	50.8	90.2	96.5
		+ TAPT	**51.6**	**89.9**	**96.6**	**51.3**	**90.4**	**96.7**
		+ DAPT + TAPT	51.5	89.6	96.5	50.2	90.2	96.4

Table 3: Performance on DST on Joint Goal Accuracy (JGA in %), Slot F1 (S F1) and Slot Accuracy (SA in %). We experiment with multi-phase pre-training on different learning objectives. Note that slot-level metrics for BERT and T5 are not comparable as they use different preprocessing steps.

extended dialogue datasets (DAPT). This follows our hypothesis DAPT will help in learning better representations for dialogue tasks when compared to the traditional pre-training on Wikipedia or books. However, we observe that the performance does not improve when we continue pre-training on the specific MultiWOZ dataset (TAPT). We attribute this to the long dialogue history in MultiWOZ along with multi-token slots that might prevent the model from learning long range dependencies. On the other hand, span prediction strategy significantly improves the performance in both DAPT (1.1% increase in JGA) and TAPT (1.4% increase in JGA).

Span corruption also helped the generative model T5, where we notice an improvement of $\sim 1.3\%$ in both TAPT and DAPT + TAPT. Note that we can not compare BERT and T5 models directly because they have different set of entities in the ontology, owing to a difference in their slot F1 accuracies. But, this does not impact the relative comparison between different pre-training techniques in the same model, which is what we want to study. We note that the overall slot F1 and accuracies do not vary profoundly, but JGA is a stricter metric to evaluate on and we believe that improvement on this metric alone is substantial to show the impact of our pre-training methods.

Model	Augmentation	MultiWOZ 2.0			MultiWOZ 2.1		
		JGA	S F1	SA	JGA	S F1	SA
BERT	Base	51.2	75.2	92.6	53.5	**75.8**	92.2
	+ Entity Replacement	51.1	75.5	92.1	53.3	75.2	**92.6**
	+ Crop	**52.0**	**75.5**	**92.6**	53.7	75.6	92.2
	+ Rotate	51.1	75.4	92.6	**53.9**	75.7	92.2
	+ Sequential	44.1	74.5	92.4	51.7	75.3	92.1
	+ Paraphrase	50.4	75.4	92.6	**53.9**	75.6	92.2
	+ Translate	51.8	75.3	92.6	53.1	75.6	92.2
T5	Base	50.2	89.6	96.5	50.3	90.2	96.3
	+ Entity Replacement	50.5	89.4	96.3	50.6	89.9	96.3
	+ Crop	**51.0**	**89.8**	**96.6**	50.7	**90.3**	96.4
	+ Rotate	50.3	89.5	96.5	50.8	90.0	96.4
	+ Sequential	50.9	89.6	96.5	50.2	90.2	96.4
	+ Paraphrase	49.7	89.6	96.5	**50.9**	**90.3**	**96.5**
	+ Translate	49.2	89.1	96.3	49.5	89.8	96.3

Table 4: Performance on DST with different data augmentation strategies.

How do different data augmentation methods fare? (refer Table 4) For MultiWOZ 2.1 dataset, data augmentation through paraphrase and rotate show significant boost in the performance for both BERT and T5 models. For MultiWOZ 2.0, Crop showed substantial improvement in JGA. Since paraphrase and rotate are different ways to modify the language structure of a sentence, we can interpret that a diversity in the language structure of the training data enhances the performance of the models. While we hypothesised that entity replacements would provide more training data and improvement, they did not fare very well. In other augmentations, translate does well for BERT providing sentence diversity and sequential does better for T5 as it helps the generative model learn better from the contextual complexity.

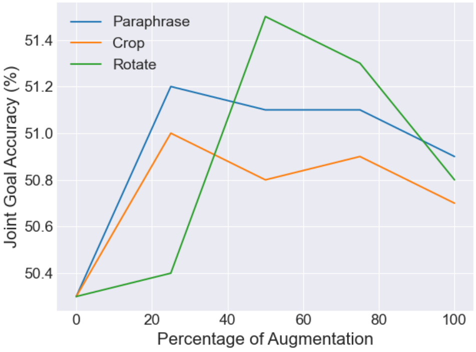

Figure 2: Plot of performance versus level of augmentation for top-3 augmentation strategies of T5. We vary the percentage of augmented data from 0 to 100 at intervals of 25.

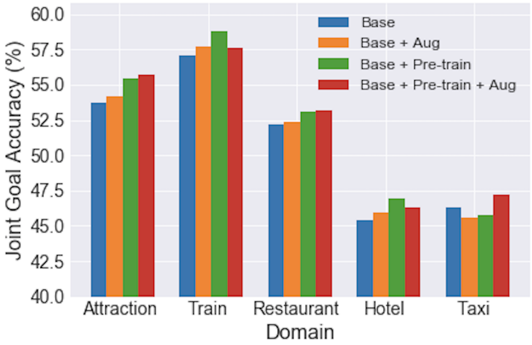

Figure 3: Performance comparison across different domains of MultiWOZ dataset. Here, we use the best pre-training strategy (TAPT) and data augmentation (Rotate) for BERT.

How does the performance vary with different levels of augmentation? (refer Figure 2) In this experiment, we gradually increase the amount of augmented data in training set and observe performance variation for 3 of the best performing augmentation strategies. We notice that augmenting data always performs better (in terms of JGA) than the non-augment counterpart. However, the performance does not increase steadily as we increase the training data. There is a dip/saturation as we use more amount of training data, which might be because of the increased noise or change in data distributions when we augment with more data.

Putting it all together: Pre-training + Data Augmentation (refer Table 5) We have demonstrated how pre-training and augmentations can independently help in improving the performance of dialogue state tracking. Now, we put all our findings together, and report the results for the combination of the best pre-training and best data augmentation techniques for BERT and T5 respectively. We observe that the combination significantly boosts the performance over the baseline models.

Model	JGA	S F1	SA
BERT	53.5	75.8	92.2
+ TAPT + Rotate	**54.4**	**76.1**	**92.3**
T5	50.3	90.2	96.3
+ TAPT + Paraphrase	**51.2**	**90.3**	**96.7**

Table 5: DST performance on MultiWOZ 2.1 with the best pre-training and augmentations combined.

How does the performance vary across different domains of MultiWOZ? (refer Figure 3) The impact of augmentation and pre-training is significant in all the domains. We notice that pre-training alone performs better than augmentation alone in all the domains, leading us to believe that pre-training is a more important factor than augmentation. Nonetheless, a combined approach is beneficial for most of the domains. Comparing across domains, there are differences in accuracy scores primarily due to the skewed distribution of dialogues per domain in the MultiWOZ dataset.

7 Qualitative Analysis

We tried to analyse the model errors by qualitatively probing into the predicted and the ground truth dialogue states for different approaches. The example in Table 6 presents an interesting instance where pre-training has helped in better understanding of the user utterance. Presence of the phrase “includes free wifi" in the user dialogue triggers the detection of *hotel-internet=yes* slot when using a baseline model with task-adaptive pre-training. Notice that the baseline model alone fails to detect this slot, which might be because of the lack of explicit token for the slot-value and pre-training helped to capture the implicit meaning of the sentence. Our models struggle to identify the "don’t care" values for slots as depicted by the example in Table 7. It fails to interpret the user meaning of “any choice" and doesn’t explicitly predict the *dontcare* token, which is present in the ground truth dialogue state. Additionally, we also noticed that the MultiWOZ dataset, despite multiple corrections, is still very noisy. The ground truth data of some utterances is incorrect which impacts the accuracy scores even when our models predict the appropriate dialogue state. For instance, in Table 6, the predicted *hotel-name* is more comprehensive with additional details like the branch name, and doest not match directly with the ground truth name. We present some additional examples in the Appendix.

User:	I am looking for a hotel to stay in that is expensive and on the east side.
Sys:	Express by Holiday Inn cambridge is on the east side and expensive.
User:	That sounds good, but can you tell me if it includes free wifi ?
Sys:	Yes it does. Would you like me to book that for you?
Base:	*hotel-area=east; hotel-pricerange=expensive; hotel-name=express by holiday inn cambridge*
Base + PT:	*hotel-area=east; hotel-pricerange=expensive; hotel-name=express by holiday inn cambridge ; hotel-internet=yes*
GT:	*hotel-area=east; hotel-pricerange=expensive; hotel-name=holiday inn ; hotel-internet=yes*

Table 6: Example dialogue where the baseline T5 model with task-adaptive pre-training (PT) helps in better understanding of the users’ ask for internet. GT represents the ground truth dialogue state.

User:	I am looking for a restaurant in the centre.
Sys:	There are over 60 restaurants to choose from in the centre. Is there a type of food you are interested in?
User:	I would like it to be expensive. Any choose is fine . I ’ll need the postcode, also, please.
Sys:	Kymmoy meets your criteria. It serves Asian oriental food. The postcode is cb12as.
Base + PT + DA:	*restaurant-area=centre; restaurant-pricerange=expensive*
GT:	*restaurant-area=centre; restaurant-food=dontcare ; restaurant-pricerange=expensive; restaurant-name=dontcare*

Table 7: Example dialogue where the baseline model with Pre-training (PT) and Data Augmentation (DA) fails to identify the *dontcare* slot-values in the Ground truth (GT) state.

8 Conclusion

To sum up, we proposed and experimented with multiple pre-training and data augmentations strategies for the task of dialogue state tracking. The results corroborate our hypothesis that dialogue data differs from web data to a great extent. We showed that the proposed multi-phase adaptive pre-training provides substantial improvement for dialogue tasks. We attempted to demonstrate how augmenting the training data can solve the data scarcity problem, but this is an active research area and requires more careful investigations in the future. We believe that these approaches can potentially assist downstream generation tasks as well, and a plausible future direction would be to extend them to the end-to-end task of response generation in dialogue systems.

References

[1] Bing Liu and Ian Lane. Attention-based recurrent neural network models for joint intent detection and slot filling. *Interspeech 2016*, pages 685–689, 2016.

[2] Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve 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)*, 2017.

[3] Christopher Tegho, Pawel Budzianowski, and Milica Gašić. Benchmarking uncertainty estimates with deep reinforcement learning for dialogue policy optimisation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[4] Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

[5] Jason D Williams, Antoine Raux, and Matthew Henderson. The dialog state tracking challenge series: A review. *Dialogue & Discourse*, 7(3):4–33, 2016.

[6] Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica 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*, 2020.

[7] Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, and Pascale Fung. Mintl: Minimalist transfer learning for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[8] Giovanni Campagna, Agata Foryciarz, Mehrad Moradshahi, and Monica Lam. Zero-shot transfer learning with synthesized data for multi-domain dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[9] Tao Yu, Rui Zhang, Alex Polozov, Christopher Meek, and Ahmed Hassan Awadallah. Score: Pre-training for context representation in conversational semantic parsing. In *International Conference on Learning Representations*, 2020.

[10] Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2018.

[11] Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*, 2019.

[12] Matthew Henderson, Blaise Thomson, and Steve Young. Word-based dialog state tracking with recurrent neural networks. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 292–299, 2014.

[13] Shuyang Gao, Abhishek Sethi, Sanchit Agarwal, Tagyoung Chung, and Dilek Hakkani-Tur. Dialog state tracking: A neural reading comprehension approach. *arXiv preprint arXiv:1908.01946*, 2019.

[14] Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. A simple language model for task-oriented dialogue. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc., 2020.

[15] Paweł Budzianowski and Ivan Vulić. Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*, 2019.

[16] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 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*, pages 270–278, Online, July 2020. Association for Computational Linguistics.

[17] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online, July 2020. Association for Computational Linguistics.

[18] Samuel Coope, Tyler Farghly, Daniela Gerz, Ivan Vulić, and Matthew Henderson. Span-convert: Few-shot span extraction for dialog with pretrained conversational representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[19] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[20] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, 2021.

[21] Samuel Louvan and Bernardo Magnini. Simple is better! lightweight data augmentation for low resource slot filling and intent classification. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 167–177, Hanoi, Vietnam, October 2020. Association for Computational Linguistics.

[22] Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

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

[24] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[25] Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. Pretraining methods for dialog context representation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[26] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 2020.

[27] Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, GP Shrivatsa Bhargav, Dinesh Garg, and Avirup Sil. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782, 2020.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[30] Li Zhou and Kevin Small. Multi-domain dialogue state tracking as dynamic knowledge graph enhanced question answering. *arXiv preprint arXiv:1911.06192*, 2019.

[31] Gözde Gül Şahin and Mark Steedman. Data augmentation via dependency tree morphing for low-resource languages. *arXiv preprint arXiv:1903.09460*, 2019.

[32] Jarana Manotumruksa, Jeff Dalton, Edgar Meij, and Emine Yilmaz. Improving dialogue state tracking with turn-based loss function and sequential data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1674–1683, 2021.

[33] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

[34] Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*, 2020.

[35] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 2020.

[36] Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[37] Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*. Springer, 2021.

[38] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 2020.

[39] Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

A Appendix

A.1 Data Augmentation Statistics

Augmentation	# Dialogues	# Turns	# Turns / Dialogue	Utterance Length	BLEU Score
Original	$8,434$	$56,747$	6.728	13.549	-
Entity Replacement	$5,225$	$32,348$	6.191	13.181	0.205
Crop	$4,789$	$35,300$	7.371	13.034	0.187
Rotate	$5,519$	$40,321$	7.306	12.997	0.193
Sequential	$8,434$	$39,883$	4.728	41.531	0.086
Paraphrase	$8,434$	$56,747$	6.728	12.122	0.046
Translate	$8,434$	$56,747$	6.728	13.544	0.205

Table 8: Statistics and BLEU score (upto bigrams) for different data augmentation techniques. Note that we report numbers only for the augmented dialogues, not including the original data samples. Number of augmented dialogues vary across different methods due to specific limitations in the data or method. For example, crop and rotate give reasonable outputs only when the parse tree has the required sub-tree components and relations that enable those operations. Entity replacement has the highest BLEU score and this is expected since the sentence structure doesn't change, only the entities. On the other hand, paraphrase has the lowest BLEU score, a subtle indication of most linguistic variation from original dataset.

A.2 Additional Examples

We provide some more interesting examples from our qualitative analysis:

User:	I am looking for a restaurant named the Lucky Star in Cambridge.
Sys:	This is a Chinese restaurant. It is located at Cambridge leisure park clifton way cherry hinton.
User:	Can you book a table for me? Just for 1 on Saturday at 12:45.
Sys:	I have your table booked. The reference number is opjjx9xa.
User:	Great. I ’ll also need a train to get me there from London King’s Cross station.
Sys:	The tr7309 leaves at 11:17 and arrives at 12:08 , would that be 1 you would like to book?
Pred State:	*restaurant-name=the lucky star; restaurant-time=12:45; restaurant-day=saturday; restaurant-people=1; train-destination=cambridge ; train-departure=Londons king’s cross*
GT State:	*restaurant-name=the lucky star; restaurant-time=12:45; restaurant-day=saturday; restaurant-people=1; train-day=saturday; train-arrive=12:30 ; train-departure=London king’s cross*
User:	As long it arrives by 08:45 going to Cambridge I should be good yes ticket for 1 please.
Sys:	Reference number is : ab2fx8kz. the total fee is 18.88 gbp payable at the station
Pred State:	*restaurant-name=the lucky star; restaurant-time=12:45; restaurant-day=saturday; restaurant-people=1; train-destination=cambridge; train-departure=Londons kings cross; train-arrive=08:45*
GT State:	*restaurant-name=the lucky star; restaurant-time=12:45; restaurant-day=saturday; restaurant-people=1; train-day=saturday; train-arrive=12:30 ; train-departure=london kings cross; train-people=1*

Table 9: Example where the predicted state is correct, but the ground truth is noisy and inconsistent with the dialogue history. Notice that the our model with pre-training and data-augmentation correctly identifies the slot *train-destination=Cambridge* based on the dialogue history of the restaurant location, which is missing in the ground truth state. Also notice the inconsistency and mismatch in the train arrival time at different turns of the dialogue and the ground truth state.

User:	Hi there. I am looking to visit a great cinema in town. can you recommend 1?
Sys:	There are 2, 1 in the south and 1 in centre. do you have a location preference?
State:	*attraction-type=cinema*
User:	Any place if fine. Please give me the phone number and entrance fee?
Sys:	Cineworld cinema 00872208000. I am sorry but the entrance fee is not listed. is there anything else i can help you with today?
State:	*attraction-type=cinema; attraction-area=dontcare*
User:	Yes please. can you provide info about a restaurant called Prezzo. What kind of food, price range, and what area if you can find it. thanks.
Sys:	They serve Italian food, are located in the west, and have a moderate price range
State:	*attraction-type=cinema; attraction-area=dontcare; restaurant-name=prezzo*
User:	Sounds good. please make a reservation for 3 people at 19:30 on Thursday.
Sys:	Booking was successful. The table will be reserved for 15 minutes. Your reference number is : h0y7r1no. Is there anything else I could help you with?
State:	*attraction-type=cinema; attraction-area=dontcare; restaurant-name=prezzo;restaurant-people=3; restaurant-day=thursday; restaurant-time=19:30*
User:	No that is it! Thanks for all your help!
Sys:	Have a good day!
State:	*attraction-type=cinema; attraction-area=dontcare; restaurant-name=prezzo;restaurant-people=3; restaurant-day=thursday; restaurant-time=19:30*

Table 10: Example dialogue spanning multiple domains, where all the slots in the ground truth dialogue state are correctly predicted by our model.