DS-TOD: Efficient Domain Specialization for Task-Oriented Dialog

Anonymous ARR submission

Abstract

Recent work has shown that self-supervised dialog-specific pretraining on large conversational datasets yields substantial gains over traditional language modeling (LM) pretraining in downstream task-oriented dialog (TOD). These approaches, however, exploit general dialogic corpora (e.g., Reddit) and thus presumably fail to reliably embed domain-specific knowledge useful for concrete downstream TOD domains. In this work, we investigate the effects of domain specialization of pretrained language models (PLMs) for TOD. Within our DS-TOD framework, we first automatically extract salient domain-specific terms, and then use them to construct DOMAINCC and DOMAINREDDIT – resources that we leverage for domain-specific pretraining, based on (i) masked language modeling (MLM) and (ii) response selection (RS) objectives, respectively. We further propose a resource-efficient and modular domain specialization by means of *domain adapters* – additional parameter-light layers in which we encode the domain knowledge. Our experiments with prominent TOD tasks – dialog state tracking (DST) and response retrieval (RR) – encompassing five domains from the MULTIWOZ benchmark demonstrate the effectiveness of DS-TOD. Moreover, we show that the light-weight adapter-based specialization (1) performs comparably to full fine-tuning in single domain setups and (2) is particularly suitable for multi-domain specialization, where besides advantageous computational footprint, it can offer better downstream performance.

1 Introduction

Task-oriented dialog (TOD), where conversational agents help users complete concrete tasks (e.g., book flights or order food), has arguably been one of the most prominent NLP applications in recent years, both in academia (Budzianowski et al., 2018; Henderson et al., 2019c; Liu et al., 2021a, *inter alia*) and industry (e.g., Yan et al., 2017; Henderson et al., 2019b). Like for most other NLP tasks, fine-tuning of pretrained language models (PLMs) like BERT (Devlin et al., 2019) and GPT-2 (Radford et al., 2019) pushed the state-of-the-art in TOD tasks (Budzianowski and Vulić, 2019; Hosseini-Asl et al., 2020), with LM pretraining at the same time alleviating the need for large labeled datasets (Ramadan et al., 2018).

More recent TOD work recognized the idiosyncrasy of dialog – that dialogs represent interleaved exchanges of utterances between two (or more) participants – and proposed pretraining objectives specifically tailored for dialogic corpora (Henderson et al., 2019c; Wu et al., 2020; Bao et al., 2020, *inter alia*). For instance, Wu et al. (2020) pretrain their TOD-BERT model on the concatenation of nine human-to-human multi-turn dialog datasets. Similarly, Henderson et al. (2019c, 2020) pretrain a general-purpose dialog encoder on a large corpus from Reddit by means of response selection objectives. Encoding dialogic linguistic knowledge in this way led to significant performance improvements in downstream TOD tasks.

While these approaches impart useful dialogic linguistic knowledge they fail to exploit the fact that individual task-oriented dialogs typically belong to one narrow domain (e.g., *food* ordering) or few closely related domains (e.g., booking a *train* and *hotel*; Budzianowski et al., 2018; Ramadan et al., 2018). Given the multitude of different downstream TOD domains (e.g., ordering *food* is quite different from booking a *flight*) it is, intuitively, unlikely that general dialogic pretraining reliably encodes domain-specific knowledge for all of them.

In this work, we propose **D**omain **S**pecialization for **T**ask **O**riented **D**ialog (DS-TOD), a novel domain specialization framework for task-oriented dialog. DS-TOD, depicted in Figure 1, has three steps: (1) we extract domain-specific terms (e.g., *taxi*-related terms) from the training portions of a task-specific TOD corpus; (2) we use the extracted terms to obtain domain-specific data from large

unlabeled corpora (e.g., Reddit); (3) we conduct intermediate training of a PLM (e.g., BERT) on the domain-specific data in order to inject the domain-specific knowledge into the encoder. As a result, we obtain a domain-specialized PLM, which can then be fine-tuned for downstream TOD tasks such as dialog state tracking or response retrieval.

Contributions. We advance the state-of-the-art in TOD with the following contributions: **(i)** Departing from general-purpose dialogic pretraining (e.g., Henderson et al., 2019a), we leverage a simple terminology extraction method to construct DOMAINCC and DOMAINREDDIT corpora which we then use for domain-specific LM and dialogic pretraining, respectively. **(ii)** We examine different objectives for injecting domain-specific knowledge into PLMs: we empirically compare Masked Language Modeling (MLM) applied on the "flat" domain dataset DOMAINCC against two different Response Selection (RS) objectives (Henderson et al., 2019c; Oord et al., 2018) applied on the dialogic DOMAINREDDIT corpus. We demonstrate the effectiveness of our specialization on two TOD tasks – dialog state tracking (DST) and response retrieval (RR) – for five domains from the MULTIWOZ dataset (Budzianowski et al., 2018; Eric et al., 2020). **(iii)** We propose modular domain specialization for TOD via *adapter* modules (Houlsby et al., 2019; Pfeiffer et al., 2020). Additional experiments reveal the advantages of adapter-based specialization in *multi-domain* TOD: combining domain-specific adapters via stacking (Pfeiffer et al., 2020) or fusion (Pfeiffer et al., 2021) (a) performs *en par* with or outperforms expensive multi-domain pretraining, while (b) having a much smaller computational footprint.[1]

2 Data Collection

We create large-scale domain-specific corpora in two steps: given a collection of in-domain dialogs we first extract salient domain terms (§2.1); we then use these domain terms to filter content from CCNet (Wenzek et al., 2020) as a large general corpus and Reddit as a source of dialogic data (§2.2).

[1]Assume N mutually close domains and a bi-domain downstream setup (any two domains). With an adapter-based approach, we pretrain one adapter for each domain (complexity: N) and then combine the adapters of the two domains intertwined in the concrete downstream setup. In contrast, multi-domain specialization would require one bi-domain pretraining for each two-domain combination (complexity: N^2).

2.1 Domain-Specific Ngrams

We start from Wizard-of-Oz, a widely used multi-domain TOD dataset (MultiWOZ; Budzianowski et al., 2018): we resort to the revised version 2.1 (Eric et al., 2020) and work with the five domains that have test dialogs: *Taxi*, *Attraction*, *Train*, *Hotel*, and *Restaurant*. Table 1 shows the statistics of domain-specific MultiWOZ subsets.

To obtain large domain-specific corpora for our intermediate training, we first construct sets of domain-specific ngrams for each domain. To this end, we first compute TF-IDF scores[2] for all {1,2,3}-grams found in single-domain dialogs from MultiWOZ training sets.[3] We then select N ngrams with the largest TF-IDF scores[4] and manually eliminate from that list ngrams that are not intrinsic to the domain (e.g., weekdays, named locations). Finally, since MultiWOZ terms follow the British spelling (e.g., *centre*, *theatre*), we add the corresponding American forms (e.g., *center*, *theater*). The resulting ngram sets are given in Table 2.

2.2 Domain-Specific Corpora

We next use the extracted domain ngrams to retrieve two types of in-domain data for domain specialization: (i) flat text and (ii) dialogic data.

DOMAINCC. For each of the five MultiWOZ domains, we create the corresponding flat text corpus for MLM training by filtering out 200K sentences from the English portion of CCNet (Wenzek et al., 2020)[5] that contain one or more of the previously extracted domain terms. We additionally clean all DOMAINCC portions by removing email addresses and URLs, and lower-casing all terms.

DOMAINREDDIT. Being constructed from CommonCrawl, DOMAINCC portions do not exhibit any natural conversational structure, encoding of which has been shown beneficial for downstream TOD (Henderson et al., 2019c; Wu et al., 2020). We thus additionally create a dialogic corpus for each domain: we employ the Pushshift API (Baumgartner et al., 2020) to extract dialogic data from

[2]TF: total ngram frequency in all domain dialogs; IDF: inverse of the proportion of dialogs containing the ngram.

[3]E.g., for the *Taxi* domain, we collect all training dialogs that span only that domain (i.e., only taxi ordering) and omit dialogs that besides *Taxi* involve one or more other domains (e.g., taxi ordering and hotel booking in the same dialog).

[4]In all our experiments, we set $N = 80$.

[5]A high-quality collection of monolingual corpora extracted from CommonCrawl that has been used for pretraining multilingual PLMs (Conneau et al., 2020; Liu et al., 2020).

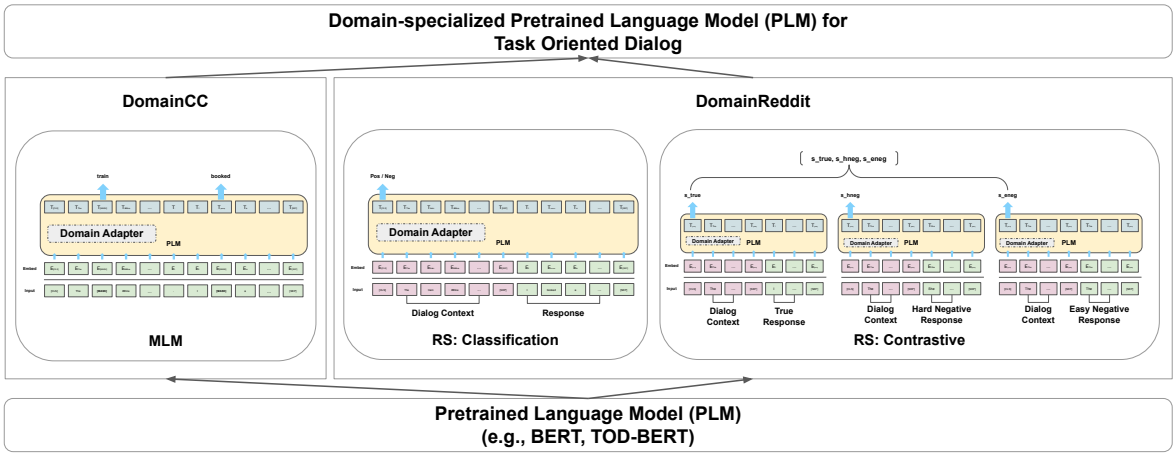

Figure 1: Overview of DS-TOD. Three different specialization objectives for injecting domain-specific knowledge into PLMs (see §3.1): (1) Masked Language Modeling (MLM) on the "flat" domain corpus DOMAINCC, (2) Response Selection (RS) via Classification, and (3) Response Selection via Contrastive Learning operating on the dialogic DOMAINREDDIT. Domain specialization performed either via (a) full fine-tuning or (b) adapters (see §3.2).

	Taxi	Restaurant	Hotel	Train	Attraction
Slot names	*destination*, *departure*, *arriveBy*, *leaveAt*	*pricerange*, *area*, *day*, *people*, *food*, *name*, *time*	*pricerange*, *parking*, *internet*, *stars*, *area*, *type*, *people*, *day*, *stay*, *name*	*destination*, *departure*, *day*, *people*, *arriveBy*, *leaveAt*	*area*, *type*, *name*
# Total (tr., dev, test)	1654, 207, 195	3813, 438, 437	3381, 416, 394	3103, 484, 494	2717, 401, 395
# Multi-domain (tr., dev, test)	1329, 150, 143	2616, 388, 375	2868, 360, 327	2828, 454, 461	2590, 390, 383
# Single domain (tr., dev, test)	325, 57, 52	1197, 50, 62	513, 56, 67	275, 30, 33	127, 11, 12
% Single domain	24.62%	19.00%	15.21%	7.25%	3.49%

Table 1: Statistics for MultiWOZ 2.1 dataset. For each domain, we report slot names, the total number of dialogs as well as the number of single-domain and multi-domain dialogs.

Reddit (period 2015–2019). To this end, we select subreddits related to *traveling* (listed in Table 3) which we believe align well with the content of MultiWOZ, which was created by simulating conversations between tourists and clerks in a tourist information center. Each of the subreddits contains threads composed of a series of comments, each of which can serve as a *context* followed by a series of *responses*. For DOMAINREDDIT we select context-response pairs where either the context utterance or the response contains at least one of the domain-specific terms. To construct examples for injecting conversational knowledge, we follow Henderson et al. (2019a) and couple each true context-response pair (i.e., a comment and its immediate response) with a *false response* – a non-immediate response from the same thread. Table 4 provides an example context with its true and one false response. Finally, we also clean DOMAINREDDIT by removing email addresses and URLs as well as comments having fewer than 10 characters. The total number of Reddit triples *(context, true response, false response)* that we extract this way for the MultiWOZ domains is as follows: *Taxi* – 120K; *Attraction* – 157K; *Hotel* – 229K; *Train* – 229K; and *Restaurant* – 243K.

3 Domain Specialization Methods

The next step in DS-TOD is the injection of domain-specific knowledge through intermediate model training on DOMAINCC and DOMAINREDDIT. To this end, we train a PLM (1) via Masked Language Modeling on DOMAINCC and (2) using two different Response Selection objectives on DOMAINREDDIT. Finally, for all objectives, we compare full domain fine-tuning (i.e., we update all PLM parameters) against adapter-based specialization where we freeze the PLM parameters and inject domain knowledge into new adapter layers.

3.1 Training Objectives

Masked Language Modeling (MLM). Following successful work on domain-adaptive pretraining via LM (Gururangan et al., 2020; Aharoni and Goldberg, 2020; Glavaš et al., 2020), we investi-

Domain	Ngrams
Taxi	*taxi, contact number, book a taxi, booked, time schedule, pickup, leaving, booked type, booking completed, departing, destination, cab, completed booked, honda, ford, audi, lexus, toyota, departure, skoda, lexus contact, toyota contact, ford contact, volvo, train station, departure site, tesla, audi contact, honda contact, skoda contact, picking, departing, volkswagen*
Attraction	*museum, college, entrance, attraction, information, centre town, center town, entertainment, swimming pool, gallery, sports, nightclub, pounds, park, postcode, architecture, centre area, center area, cinema, church, trinity college, entrance free, jello gallery, post code, town centre, town center, downing college*
Train	*train station, travel time, leaving, pounds, train ticket, departing, payable, train leaving, cambridge, london, reference id, arrive, destination, kings cross, total fee, departure, arriving, book a train, booked, stansted, stansted airport, peterborough, traveling, trip, airport, booking successful, norwich*
Hotel	*hotel, nights, parking, free parking, wifi, star hotel, price range, free wifi, guesthouse, guest house, internet, guest, hotel room, star rating, expensive room, priced, rating, book room, moderately priced, moderate price, stay for, reservation, breakfast available, book people, fully booked, booking, reference*
Restaurant	*restaurant, food, price range, expensive, cheap, priced, chinese food, italian food, moderately priced, south town, book table, city, north town, serving, city centre, city center, european food, reservation, food type, phone address, centre town, center town, expensive restaurant, moderate price, cuisine, restaurant center, restaurant centre, south town, expensive price, east town, cheap restaurant, indian food, asian food, british food, book people*

Table 2: Salient domain ngrams extracted from the single-domain training portions of MultiWOZ.

Subreddit	# Members	Domains
travel	5.8M	Taxi, Attraction, Train, Hotel, Restaurant
backpacking	2.5M	Taxi, Attraction, Train, Hotel, Restaurant
solotravel	1.7M	Taxi, Attraction, Train, Hotel, Restaurant
CasualUK	797K	Taxi, Attraction, Train, Hotel, Restaurant
unitedkingdom	553K	Taxi, Attraction, Train, Hotel, Restaurant
restaurant	81.6K	Restaurant
trains	64.8K	Train, Attraction
hotel	1.8K	Hotel
hotels	4.9K	Hotel
tourism	3.9K	Taxi, Attraction, Train, Hotel, Restaurant
uktravel	1.5K	Taxi, Attraction, Train, Hotel, Restaurant
taxi	0.6K	Taxi

Table 3: Subreddits and associated domains selected for creating DOMAINREDDIT.

Field	Example
Subreddit	restaurant
Context	*Hosts don't get tips? That's news to me. Most host positions in my area get at least 1% of sales; they make anywhere between 60−100 per night in tips!*
Response	*We get tips but definitely not that much (in my experience). The tip out in my restaurant is 1% split between shift leaders, food runners, and any other FOH other than servers/bartenders. Full time hosts get about 50-75 every other week*
False response	*Wow that's terrible. Then again, my restaurant is in CA, so wages and guest check averages are usually higher.*

Table 4: Example from DOMAINREDDIT dataset.

gate the effect of running standard MLM on the domain-specific portions of DOMAINCC.

Response Selection (RS). RS objectives force the model to recognize the correct response utterance given the context – pretraining with such objectives is particularly useful for conversational settings, including TOD tasks (Henderson et al., 2019c, 2020). We consider two RS objectives. The first is a simple pairwise binary classification formulation (**RS-Class**): given a context-response pair, predict whether the response is a true (i.e., immediate) response to the context. We straightforwardly use pairs of contexts and their true responses from DOMAINREDDIT as positive training instances. Next, we create negative samples for each positive instance as follows: (a) we use the crawled *false response* from DOMAINREDDIT,[6] which represents a relevant but non-consecutive response from the same thread; (b) we additionally randomly sample k utterances from the same domain but different threads (the *easy negatives*).[7]

The second response selection objective (**RS-Contrast**) that we adopt is a type of loss function used for contrastive model training based on the representational similarities between sampled positive and negative pairs (Oord et al., 2018). It has been used for pretraining cross-lingual language models (Chi et al., 2021) and shown to be useful in information retrieval (Reimers and Gurevych, 2021; Thakur et al., 2021). The idea is to estimate the mutual information between pairs of variables by discriminating between a positive pair and its associated N negative pairs. Given a true context-response pair and N corresponding negatives (the same as for RS-Class), the noise-contrastive estimation (NCE) loss is computed as:

$$L_{NCE} = -\log \frac{\exp\left(f(c, r_{+})\right)}{\sum_{i=1}^{N+1} \exp\left(f(c, r_i)\right)},$$

where c is the context, r_{+} is the true response and r_i iterates over all responses for the context – the true response r_{+} and N false responses; a function f produces a score that indicates whether the response r is a true response of the context c.

[6] Non-immediate responses from the same thread represent the so-called *hard negatives* introduced to prevent the model from learning simple lexical cues and similar heuristics that poorly generalize.

[7] k is uniformly sampled from the set $\{1, 2, 3\}$.

By learning to differentiate whether the response is true or false for a given context (RS-Class) or to produce a higher score for a true response than for false responses (RS-Contrast), RS objectives encourage the PLM to adapt to the underlying structure of the conversation. By feeding only in-domain data to it, we encode domain-specific conversational knowledge into the model.

3.2 Adapter-Based Domain Specialization

Fully fine-tuning the model requires adjusting all of the model's parameters, which can be undesirable due to large computational effort and risk of catastrophic forgetting of the previously acquired knowledge (McCloskey and Cohen, 1989; Pfeiffer et al., 2021). To alleviate these issues, we investigate the use of adapters (Houlsby et al., 2019), additional parameter-light modules that are injected into a PLM before fine-tuning. In adapter-based fine-tuning only adapter parameters are updated while the pretrained parameters are kept frozen (and previously acquired knowledge thus preserved). We adopt the adapter-transformer architecture proposed by Pfeiffer et al. (2020), which inserts a single adapter layer into each transformer layer and computes the output of the adapter, a two-layer feed-forward network, as follows:

$$\textit{Adapter}(\boldsymbol{h}, \boldsymbol{r}) = U \cdot g(D \cdot \boldsymbol{h}) + \boldsymbol{r},$$

with $\boldsymbol{h}$ and $\boldsymbol{r}$ as the hidden state and residual of the respective transformer layer. $D \in \mathbb{R}^{m \times h}$ and $U \in \mathbb{R}^{h \times m}$ are the linear down- and up-projections, respectively (h being the transformer's hidden size, and m as the adapter's bottleneck dimension), and $g(\cdot)$ is a non-linear activation function. The residual $\boldsymbol{r}$ is the output of the transformer's feed-forward layer whereas $\boldsymbol{h}$ is the output of the subsequent layer normalization. The down-projection D compresses token representations to the adapter size $m \ll h$, and the up-projection U projects the activated down-projections back to the transformer's hidden size h. The ratio h/m captures the factor by which the adapter-based fine-tuning is more parameter-efficient than full fine-tuning.

For multi-domain TOD scenarios (i.e., dialogs covering more than a single domain), we further experiment with combinations of individual domain adapters: (1) sequential stacking of adapters one on top of the other (Pfeiffer et al., 2020) and (2) adapter fusion, where we compute a weighted average of outputs of individual adapter, with fusion weights as parameters to be tuned in the final task-specific fine-tuning (Pfeiffer et al., 2021).

4 Experiments

Evaluation Task and Measures. We evaluate our domain-specialized models and baselines on two prominent downstream TOD tasks: *dialog state tracking (DST)* and *response retrieval (RR)*. DST is treated as a multi-class classification task based on a predefined ontology, where given the dialog history, the goal is to predict the output state, i.e., (domain, slot, value) tuples. For our implementation, we follow Wu et al. (2020), and represent the dialog history as a sequence of utterances. The model then needs to predict slot values for each (domain, slot) pair at each dialog turn. We report the *joint goal accuracy*, in which the predicted dialog states are compared to the ground truth slot values at each dialog turn. The ground truth contains slot values for all the (domain, slot) candidate pairs. A prediction is considered correct if and only if all predicted slot values exactly match its ground truth values. RR is a ranking problem, relevant for retrieval-based TOD systems (Wu et al., 2017; Henderson et al., 2019c). Following Henderson et al. (2020) and Wu et al. (2020), we adopt recall at top rank given 100 randomly sampled candidates (R_{100}@1) as the evaluation metric for RR.

Data. In the pretraining procedure, we use the domain-specific portions of our novel DOMAINCC and DOMAINREDDIT resources (§2). For the MLM training, we randomly sample 200K domain-specific contexts from DOMAINCC and dynamically mask 15% of the subword tokens. For RS-Class and RS-Contrast, we randomly sample 200K instances from DOMAINREDDIT. We evaluate the efficacy of the methods on DST and RR using MultiWOZ 2.1 (Eric et al., 2020). Since we aim to understand the effect of the domain specialization, we construct domain-specific training, development, and testing portions from the original data set by assigning them all dialogs that belong to a domain (i.e., both single- and multi-domain dialogs) from respective overall (train, dev, test) portions.

Models and Baselines. We experiment with two PLMs: BERT (Devlin et al., 2019) and its TOD-sibling, TOD-BERT (Wu et al., 2020).[8] As baselines, we report the performance of the non-

[8] We use the pretrained models `bert-base-cased` and `TODBERT/TOD-BERT-JNT-V1` from HuggingFace.

	Dialog State Tracking						Response Retrieval					
Model	**Taxi**	**Restaurant**	**Hotel**	**Train**	**Attraction**	Avg.	**Taxi**	**Restaurant**	**Hotel**	**Train**	**Attraction**	Avg.
BERT	23.87	35.44	30.18	41.93	29.77	32.24	23.25	37.61	38.97	44.53	48.47	38.57
TOD-BERT	30.45	43.58	36.20	48.79	42.70	40.34	45.68	57.43	53.84	**60.66**	60.26	55.57
BERT-MLM	23.74	37.09	32.77	40.96	36.66	34.24	31.37	53.08	45.41	51.66	52.23	46.75
TOD-BERT-MLM	29.94	43.14	36.11	47.61	41.54	39.67	41.77	55.27	50.60	55.17	54.62	51.49
TOD-BERT-RS-Class	**36.39**	43.38	37.89	48.82	43.31	41.96	47.01	58.21	**57.05**	59.70	57.72	55.94
TOD-BERT-RS-Contrast	35.03	**44.81**	**38.74**	**49.04**	42.73	**42.07**	48.04	59.82	54.49	60.06	60.63	56.61
BERT-MLM-adapter	22.52	40.49	31.90	42.17	35.05	34.43	32.84	44.01	39.15	38.43	45.05	39.90
TOD-BERT-MLM-adapter	32.06	44.06	36.74	48.84	**43.50**	41.04	49.08	58.18	55.55	59.46	60.26	56.51
TOD-BERT-RS-Class-adapter	33.10	42.57	38.61	49.03	42.35	41.13	**49.59**	**61.26**	56.87	58.88	60.00	**57.32**
TOD-BERT-RS-Contrast-adapter	34.90	44.42	37.52	48.71	42.83	41.68	47.97	58.97	55.41	59.15	**61.95**	56.69

Table 5: Results of DS-TOD models on two downstream tasks: Dialog State Tracking (DST) and Response Retrieval (RR) with joint goal accuracy (%) as the metric for DST and $R_{100}@1$ (Henderson et al., 2020) (%) for RR.

specialized variants and compare them against our domain-specialized PLM variants, obtained after intermediate MLM-training on DOMAINCC or RS-Class/RS-Contrast training on DOMAINREDDIT.

Hyperparameters and Optimization. During domain-specific pretraining, we fix the maximum sequence length to 256 subword tokens (for RS objectives, we limit both the context and response to 128 tokens). We train for 30 epochs, in batches of 32 instances and search for the optimal learning rate among the following values: $\{1 \cdot 10^{-4}, 5 \cdot 10^{-5}, 1 \cdot 10^{-5}, 1 \cdot 10^{-6}\}$. We apply early stopping based on development set performance (patience: 3 epochs). We minimize the cross-entropy loss using Adam (Kingma and Ba, 2015). For downstream evaluation, we train for 300 epochs in batches of 6 (DST) and 24 instances (RS) with the learning rate fixed to $5 \cdot 10^{-5}$. We also apply dev-set-based early stopping (patience: 10 epochs).

5 Results and Discussion

Overall performance. We report downstream DST and RR results in Table 5, which is segmented in three parts: (1) at the top we show the baseline results (BERT, TOD-BERT) without any domain specialization; (2) in the middle of the table we show results of PLMs specialized for domains via full fine-tuning; (3) the bottom of the table contains results for our adapter-based domain specialization.

In both DST and RR, TOD-BERT massively outperforms BERT due to its conversational knowledge. Domain specialization brings gains for both PLMs across the board. The only exception is full MLM-fine-tuning of TOD-BERT (i.e., TOD-BERT-MLM vs. TOD-BERT; -4% for RR and -0.8% for DST): we believe that this is an example of negative interference – while TOD-BERT is learning domain knowledge, it is – because of MLM-based domain training – forgetting the conversational knowledge obtained in dialogic pretraining (Wu et al., 2020). This hypothesis is further supported by the fact that adapter-based MLM specialization of TOD-BERT – which prevents negative interference by design – brings slight performance gains (i.e., TOD-BERT-MLM-adapter vs. TOD-BERT; +0.8% for DST and +1.0% for RR) and is consistent with the concurrent findings of Qiu et al. (2021).

Overall, domain specialization with RS seems to be more robust than that via MLM-ing, with the two variants (RS-Class and RS-Contrast) exhibiting similar average performance across evaluation settings. This points to the importance of injecting both the knowledge of dialogic structure as well as domain knowledge for performance gains in TOD tasks in the domain of interest.

Interestingly, the gains from domain specialization are significantly more pronounced for *Taxi* than for other domains. We relate this to the proportion of the single-domain dialogs for a given domain in MultiWOZ, which is by far the largest (24%, see Table 1) for the *Taxi* domain. Consequently, successful specialization for that domain is *a priori* more likely to show substantial gains on MultiWOZ (i.e., less multi-domain influence).

An encouraging finding is that, on average, adapter-based specialization yields similar gains as specialization via full fine-tuning: given that adapter fine-tuning is substantially more efficient, this holds the promise of more sustainable TOD.

Sample Efficiency. To further understand the effect of the injected domain-specific knowledge, we conduct an additional few-shot analysis (Figure 2) on DST. To this end, we select the *Taxi* domain, since we witnessed the largest gains for that domain. We analyse the differences in performance between baseline and domain-specialized PLMs when they are exposed to downstream training por-

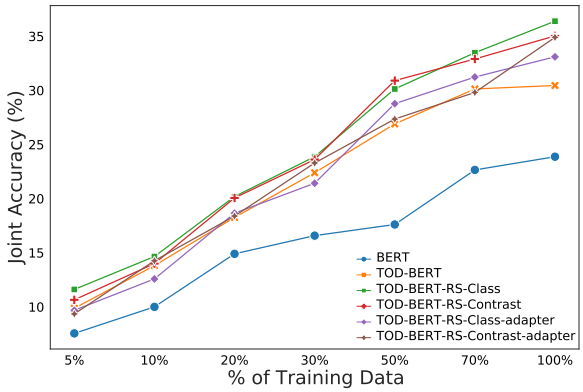

Figure 2: Sample efficiency of DS-TOD for DST: joint goal accuracy (%) for randomly sampled sub-portions (5%, 10%, 20%, 30%, 50%, 70%, and 100%) of the downstream training data from the *Taxi* domain.

tions of different sizes, ranging from 5 to 100% of the whole training dataset.[9] TOD-BERT retains a sizable performance gap over BERT for all settings, pointing to the power of dialogic pretraining. Importantly, for all dataset sizes, the performances of the domain-specialized variants of TOD-BERT-RS-{Class, Contrast} surpass the one of the non-specialized TOD-BERT. Even more interestingly, specialized variants exposed to only 50% of the DST training data manage to surpass the performance of TOD-BERT fine-tuned on all of the training data (100%). This suggests that self-supervised domain specialization has the potential to substantially reduce the amount of annotated TOD data required to reach some performance level.

Cross-Domain Transfer. MultiWOZ domains are mutually quite related: some are similar, i.e., share vocabulary and slots (e.g., *Taxi* and *Train*) whereas others often appear together in a dialog (e.g., *Train* and *Hotel*; see Table 1 for the number of multi-domain MultiWOZ dialogs). We thus next investigate whether intermediate training for one domain benefits other, closely related domains. To this end, we expose models specialized for one domain (e.g., *Taxi*) to downstream fine-tuning and evaluation in the other domain (e.g., *Restaurant*). Figure 3 summarizes the deltas in performance between the non-specialized TOD-BERT and TOD-BERT-RS-Contrast for all domain pairs. Encouragingly, the specialization for one domain seems to generally lead to downstream gains in related domains too: the gains are most prominent for pairs of domains that frequently co-occur in dialogs – *Hotel* pretraining for the *Restaurant* downstream (and vice versa) and *Taxi* pretraining for downstream tasks in the *Restaurant* and *Attraction* domains.

Multi-Domain Specialization. In many real-world scenarios, a single model needs to be able to handle multiple domains because (a) multi-domain (MD) dialogs exist and (b) simultaneous deployment of multiple single-domain (SD) models may not be feasible. To simulate this scenario, we conduct an additional analysis, in which we concatenate dialogs from respective MultiWOZ portions that cover concrete combinations of two or three domains. We choose three domain combinations with the largest number of MD dialogs, namely the two largest 2-domain combinations and the largest 3-domain combination: *Hotel+Train*, *Attraction+Train*, and *Hotel+Taxi+Restaurant*.

As baselines, we report the performance of BERT and TOD-BERT fine-tuned on the respective MD TOD training sets. We test the effect of MD specialization in two variants: (1) *fully specialized model trained for multiple domains* (**Full-FT**): as RS-Class has proven to be effective in our SD-specialization experiments, we run RS-Class training on the concatenation of the selected domains from DOMAINREDDIT that correspond to the domains of the joint training sets. Accordingly, the training data is roughly twice (or three times) as big as that used for SD specialization; (2) *composition of SD adapters for multiple domains*: while for Full-FT, a new intermediate training is necessary for each domain combination, with adapter-based specialization we can simply combine the adapters of relevant domains in downstream fine-tuning. In this setup, we combine the SD adapters by sequentially stacking them (Pfeiffer et al., 2020) (**Stacking**) or by fusing them, i.e., interpolating between their outputs (Pfeiffer et al., 2021) (**Fusion**).

The MD specialization results are shown in Table 6. Interestingly, combining SD adapters in downstream training (via Stacking or Fusion) performs *en par* with full-sized two-domain specialization on DOMAINREDDIT by means of RS-Class training. In contrast to TOD-BERT-RS-Class (Full-FT), which requires full retraining of the model on the unlabelled domain-specific corpora for each combination of the domains, combining SD adapters is much more efficient as it does not require any further intermediate domain training for domain combinations. In the 3-domain setup (*Hotel+Taxi+Restaurant*), the Fusion approach even

[9] Note that 5% of the training data in the *Taxi* domain amounts to 83 dialogs.

Figure 3: Relative improvements (TOD-BERT-RS-Contrast vs. TOD-BERT) in cross-domain DST transfer.

Model		Hotel+ Train	Attraction+ Train	Hotel+Taxi+ Restaurant
BERT		42.66	45.06	37.00
TOD-BERT		46.38	46.40	42.47
TOD-BERT-RS-Class	Full-FT	**47.39**	**47.33**	42.39
	Stacking	47.19	46.68	42.15
	Fusion	44.25	45.57	**44.02**

Table 6: DS-TOD performance on DST in multi-domain scenarios. We compare the fully multi-domain-specialized variant (Full-FT) of the TOD-BERT-RS-Class model with its variant that combines readily available single-domain adapters (Stacking and Fusion) on three multi-domain evaluation sets.

outperforms the full 3-domain specialization (TOD-BERT-RS-Class Full-FT) by 2 points.

Overall, we find that the adapter compositions provide a simple and effective way to combine information from several domain-specialized adapters, removing the need for additional MD specialization in the face of MD dialogs downstream.

6 Related Work

TOD Datasets. Datasets for task-oriented dialog can be divided into single-domain (Wen et al., 2017; Mrkšić et al., 2017) and multi-domain ones (Budzianowski et al., 2018; Rastogi et al., 2020). The latter are generally seen as closer to real-world situations and intended usages of personal assistants, where strict adherence to a single domain is unlikely. While downstream TOD datasets exist for specific domains, corresponding large(er)-scale datasets that would enable domain-specific pretraining have been limited to the general domain (Henderson et al., 2019a). We address this gap in this work by creating large-scale domain-specific corpora – flat as well as dialogic – for the five domains of the MultiWOZ dataset.

Pretrained Language Models in Dialog. The advantages of large-scale pretraining of deep language models on massive amounts of text (Devlin et al., 2019; Radford et al., 2019; Lewis et al., 2020), ubiquitous in natural language tasks, have also spilled over to task-oriented dialog. Recent research focused on either (1) leveraging general-domain dialogic resources (e.g., Reddit, Twitter) in order to improve downstream TOD tasks (Henderson et al., 2019c, 2020; Zhang et al., 2020; Bao et al., 2020; Liu et al., 2021b) or (2) using TOD datasets to inject dialogic structure into PLMs (Wu et al., 2020; Peng et al., 2021; Su et al., 2021). Neither of the two, however, considers domain adaptation or domain-specific pretraining.

Domain Adaptation and Knowledge Reuse. Intermediate training is the prevalent approach for injecting domain knowledge into PLMs, either as a step before the downstream task-specific fine-tuning (Glavaš et al., 2020) or in parallel with it (i.e., in a multi-task training setup) (Gururangan et al., 2020). In the narrower context of TOD, Whang et al. (2020) present the lone effort on domain specialization for TOD: they focus on easier, single-domain TOD and investigate the specialization effect with a single task, response retrieval. In this work, in contrast, we focus on dialogic domain-specific pretraining and show its effectiveness in multi-domain TOD. For efficiency and to avoid catastrophic forgetting, adapter modules have been widely used for parameter-efficient fine-tuning of PLMs for new tasks (Houlsby et al., 2019) and languages (Pfeiffer et al., 2020). Non-destructive adapter compositions (e.g., stacking or fusion) can be beneficial if multiple knowledge facets, stored in separate adapters, need to be leveraged (Pfeiffer et al., 2020, 2021).

7 Conclusion

We introduced **DS-TOD** – a novel framework for domain specialization of PLMs for task-oriented dialog. Given a collection of in-domain dialogs, we extract domain terms and use them to filter in-domain dialogic corpora. Our experimental study, on five domains of the MultiWOZ dataset, shows that domain specialization, especially by means of response selection objectives on the dialogic in-domain corpora, leads to consistent gains in TOD tasks: dialogue state tracking and response retrieval. We hope that our domain-specific resources (which we make available at `[URL-ANONYMOUS]`) catalyze research on domain specialization for TOD, especially for multi-domain setups. Our future efforts will focus on the joint domain- and language-specialization for task-oriented dialog.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 85–96, Online. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 15–22, Hong Kong. Association for Computational Linguistics.

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*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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*, pages 8342–8360, Online. Association for Computational Linguistics.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019a. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Matthew Henderson, Ivan Vulić, Iñigo Casanueva, Paweł Budzianowski, Daniela Gerz, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019b. PolyResponse: A rank-based approach to task-oriented dialogue with application in restaurant search and booking. 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): System Demonstrations*, pages 181–186, Hong Kong, China. Association for Computational Linguistics.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019c. Training neural response

selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404, Florence, Italy. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiexi Liu, Ryuichi Takanobu, Jiaxin Wen, Dazhen Wan, Hongguang Li, Weiran Nie, Cheng Li, Wei Peng, and Minlie Huang. 2021a. Robustness testing of language understanding in task-oriented dialog. 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)*, pages 2467–2480, Online. Association for Computational Linguistics.

Qi Liu, Lei Yu, Laura Rimell, and Phil Blunsom. 2021b. Pretraining the Noisy Channel Model for Task-Oriented Dialogue. *Transactions of the Association for Computational Linguistics*, 9:657–674.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. 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)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Buildingtask bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Yao Qiu, Jinchao Zhang, and Jie Zhou. 2021. Different strokes for different folks: Investigating appropriate further pre-training approaches for diverse dialogue tasks. *arXiv preprint arXiv:2109.06524*.

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):9.

Osman Ramadan, Paweł Budzianowski, and Milica Gašić. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 432–437, Melbourne, Australia. Association for Computational Linguistics.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Nils Reimers and Iryna Gurevych. 2021. The curse of dense low-dimensional information retrieval for large index sizes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2:*

Short Papers), pages 605–611, Online. Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2021. Multi-task pre-training for plug-and-play task-oriented dialogue system. *arXiv preprint arXiv:2109.14739*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Taesun Whang, Dongyub Lee, Chanhee Lee, Kisu Yang, Dongsuk Oh, and HeuiSeok Lim. 2020. An effective domain adaptive post-training method for bert in response selection. In *Proc. Interspeech 2020*, pages 1585–1589.

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)*, pages 917–929, Online. Association for Computational Linguistics.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4618–4625. AAAI Press.

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*, pages 270–278, Online. Association for Computational Linguistics.