# External Knowledge Selection with Weighted Negative Sampling in Knowledge-grounded Task-oriented Dialogue Systems

**Janghoon Han, Joongbo Shin, Hosung Song, Hyunjik Jo, Gyeonghun Kim, Yireun Kim, Stanley Jungkyu Choi**
LG AI Research
Seoul, Korea
{janghoon.han, jb.shin, hosung.song, hyunjik.jo, ghkayne.kim, yireun.kim, stanleyjk.choi}@lgresearch.ai

## Abstract

Constructing a robust dialogue system on spoken conversations bring more challenge than written conversation. In this respect, `DSTC10-Track2-Task2` is proposed, which aims to build a task-oriented dialogue (TOD) system incorporating unstructured external knowledge on a spoken conversation, extending `DSTC9-Track1`. This paper introduces our system containing four advanced methods: data construction, weighted negative sampling, post-training, and style transfer. We first automatically construct a large training data because `DSTC10-Track2` does not release the official training set. For the knowledge selection task, we propose weighted negative sampling to train the model more fine-grained manner. We also employ post-training and style transfer for the response generation task to generate an appropriate response with a similar style to the target response. In the experiment, we investigate the effect of weighted negative sampling, post-training, and style transfer. Our model ranked 7 out of 16 teams in the objective evaluation and 6 in human evaluation. [1]

## Introduction

Task-oriented dialogue (TOD) systems, which aim to assist users with specific tasks through conversation, have received much attention in research and industry due to their applicability in various services such as personal assistants and customer chatbots (Chen et al., 2017). In general, TOD systems are able to respond to the user based on a given DB or API. In reality, however, a user often requests detailed information that exceeds the DB or API coverage, such as whether a companion with a pet is allowed in a restaurant may not be included in the DB or API.

Recently, `DSTC9-Track1` (Kim et al., 2020) tackled this issue by proposing a new task that integrates external knowledge sources into TOD systems. By making the systems utilize frequently-asked questions (FAQs), which is a typical in-domain unstructured knowledge, we anticipate that TOD systems are able to respond to requests beyond the DB or API with no friction. Specifically, this track aims at developing a pipeline of three successive sub-tasks: 1) detecting whether external knowledge is required for a given dialog, 2) selecting a proper knowledge snippet from the entire knowledge, and 3) generating a response based on the retrieved knowledge. Various studies (He et al., 2021; Mi et al., 2021; Tang et al., 2021) have been conducted on unstructured knowledge-grounded TOD systems.

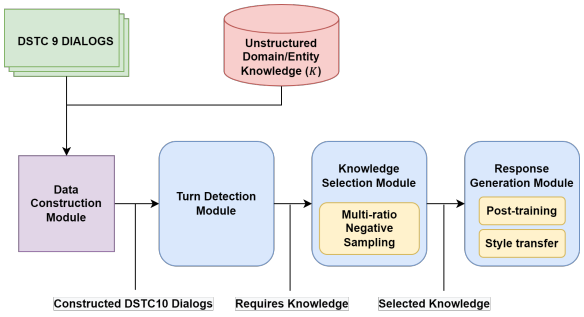


Figure 1: Overall architecture of our knowledge-grounded task-oriented dialogue system.

In reality, many TOD services include spoken conversation scenarios such as customer service and call centers. However, training such a system is much more difficult because spoken conversations contain speech recognition noise. In this respect, the Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations task or `DSTC10-Track2-Task2` has been proposed by Kim et al. (2021). The task aims at the robustness of the systems against the gaps between written and spoken conversations.

In this work, we build a knowledge-grounded TOD system that solves the `DSTC10` task. The outline of our system is shown in Figure 1. First, we automatically construct new spoken conversation data using `DSTC9` conversation and `DSTC10` knowledge FAQ because the `DSTC10` task only provides new external knowledge and no spoken dialogues for training. And then, we use our synthetic dialogue data for training the detection, selection, and generation modules. Furthermore, we propose a new weighted negative sampling method in the selection module to improve. Finally, we apply the post-training method, which trains the pre-trained model on a large TOD corpus before fine-tuning, and the style transfer method, which learns style from responses similar to target system utterances in the generation

[1] https://github.com/hanjanghoon/Weighted_NS

step. This allows the system to generate a response appropriate for the dialogue contexts and similar to the target.

In summary, our contributions are as follows.

1. **Automatic data construction:** We automatically construct new conversations required for model training using DSTC9 data and new knowledge of DSTC10.
2. **Weighted negative sampling:** We improve the model's performance by suggesting and applying a new weighted negative sampling in the selection module.
3. **Post-training and style transfer:** We generate an appropriate response with a similar style to the target through the post-training and transfer process in the generation module.

## Related Works

The traditional pipeline approach (Williams and Young, 2007) in TOD systems consists of several subdivided components (NLU: natural Language understanding, DST: dialogue state tracking, DM: dialogue management, NLG: natural language generation). With the development of deep learning technology, TOD systems are gradually progressing from pipeline approach to end-to-end approach (Shi and Yu, 2018; Ham et al., 2020) that combines some or all components of an existing pipeline.

As the architecture of TOD systems has changed, a large body of TOD benchmarks has also been proposed. Henderson, Thomson, and Williams (2014); Wen et al. (2017) provide TOD datasets for restaurant domain. Recently, as large-scale multi-domain dataset such as MultiWOZ (Budzianowski et al., 2018; Eric et al., 2019) have emerged, research on multi-domain TOD are being actively conducted (Wu et al., 2019; Hosseini-Asl et al., 2020; Yang, Li, and Quan, 2021).

DSTC9-Track1 (Kim et al., 2020) is a task that focuses on generating an appropriate response in a turn that requires external knowledge resources. Data is an augmented version of Multiwoz 2.1 (Eric et al., 2019) with out-of-API-coverage turns grounded on external knowledge sources beyond the original database entries. This task consists of three subtasks: turn detection, knowledge selection, and knowledge-grounded response generation. In this track, He et al. (2021) applies the multi-scale (random, in-domain, in-entity, cross-entity) negative sampling and shows the best performance. The second best system (Mi et al., 2021) improves the robustness of the system by introducing data augmentation, joint task learning, and an error-fixing ensemble algorithm.

Style Transfer is a method that has been successfully employed in the vision field (Gatys, Ecker, and Bethge, 2016). It changes original data into data of the desired style. Since data for style transfer is expensive and difficult to obtain, augmentation methods are applied to construct data. In the field of natural language processing, a back-translation (Sennrich, Haddow, and Birch, 2016; Prabhumoye et al., 2018) method is widely used. This method converts the input to another language or another domain and then back to the original domain to secure the diversity of sentences. Specially back-translation is used not only in the natural language processing field but also in the automatic speech recognition (ASR) field as a method to acquire parallel text data for speech (Hayashi et al., 2018).

## Problem Formalization

DSTC10-Track2-Task2 focuses on accessing external knowledge and generating appropriate answers, assuming that a conventional API-based TOD system already exists. Dialogue context is a sequence of $m$ utterances which is $C_i = \{u_1, s_2, ..., u_m\}$, where $u_t$ and $s_t$ denote user utterance and system utterance respectively at the $t_{th}$ turn. A knowledge set $K \in \{k_1, ..., k_n\}$ contains $n$ knowledge faq pairs $k_j$ composed of domain, entity, title and body. The final goal is to generate a knowledge reflected response $r_i$ by selecting the knowledge suitable for the conversation context when an utterance requires external knowledge. We define three sub-tasks as follows.

- **Detection:** Build a model to predict whether external knowledge access is needed to respond for a given dialogue context $C_i$. $y_i \in \{0, 1\}$ denote truth label, $y_i = 1$ indicate accessing external knowledge is required; otherwise, $y_i = 0$. We define detection task formulation as follow:

$$g_{detection}(C_i) \in \{0, 1\} \tag{1}$$

- **Selection:** Select the most relevant knowledge for given the dialogue context $C_i$ and knowledge set $K$. The matching degree between $C_i$ and knowledge $k_j$ is denoted as $g_{selection}(C_i, k_j)$, and the knowledge with the highest matching degree becomes the related knowledge. we define selection task formulation as follow:

$$g_{selection}(C_i, k_j) \in \{0, 1\} \tag{2}$$

  where $y_i = 1$ indicate $k_j$ is relevant with context $C_i$; otherwise, $y_i = 0$.

- **Generation:** Find a generative model $g_{generation}$ that generates a response $r_i$ suitable for given context $C_i$ and related knowledge $k_r$. We define generation task formulation as follow:

$$g_{generation}(C_i, k_r) = r_i \tag{3}$$

## Methodology

### Automatic Data Construction

In the DSTC10 knowledge-grounded TOD task, no training conversation data are given. However, the existing DSTC9 conversation data ($D_9$) is inappropriate for training because the knowledge of DSTC9 ($K_9$) and DSTC10 ($K_{10}$) is different and $D_9$ is not spoken conversation data. To address this issue, we developed a module that automatically constructs new DSTC10 conversation data ($D_{10}$) based on $D_9$ and $K_{10}$.

To construct conversations about $K_{10}$, we replace the last user turn $u_m$ that requires knowledge from the $K_9$ to the new utterance that requires $K_{10}$. The steps are as follows: 1) create a conversation session template by removing the last user turn $u_m$, which requires knowledge, from the $D_9$.

2) select a knowledge snippet from $K_{10}$. 3) replace the dialogue template with the corresponding entity of the selected knowledge snippet, and substitute the last user-turn and target response with query and answer, which match the title and body of chosen knowledge snippet. Query and answer are the utterances with high similarity scores on knowledge title and body respectively from the candidate set; candidate set comprises $D_9$, $K_9$, the paraphrased `DSTC10` knowledge and $K_{10}$. We used sentence BERT (Reimers and Gurevych, 2019) to measure sentence similarity. We also train T5 (Raffel et al., 2020) to paraphrase knowledge.

Finally, we intentionally add some ASR-like noises such as disfluencies and barge-ins into the generated conversations for imitating spoken conversations. We employ a method similar to back translation among data augmentation methods to train the noise injector module. The learning process detail is as follows. First, we train a wav2vec2.0 (Baevski et al., 2020) based ASR model using the common voice dataset (Ardila et al., 2020) for data augmentation. Afterward, we train the BART (Lewis et al., 2020) based noise injector module that transfers the written to spoken style, including ASR noise.

## Knowledge-seeking Turn Detection

For a given dialogue context, we need to check whether it requires external knowledge or not. To address this problem, we approach this task to the binary classification task. We use GPT2 (Radford et al., 2019) base model, and the model input $x$ is as follows:

$$x = [BOS]\ [user]\ u_1\ [sys]\ s_2\ [user]\ u_3\ ...\ u_m\ [EOS] \tag{4}$$

where $[BOS], [EOS]$ are begin of the sentence token and the end of the sentence token, respectively. We also insert speaker tokens $[user], [sys]$ to distinguish user and system utterances. After that, the output state of the last token $T_{[LAST]}$ is used as the aggregate representation and passed through the single linear function as follow:

$$g_{detection}(C_i) = \sigma(W_{detection} T_{[LAST]} + b) \tag{5}$$

where $W_{detection}$ is a task-specific trainable parameter. Eventually, the model weights are fine-tuned by using the cross-entropy loss function.

$$Loss = -\sum_i y_i \log(g_{detection}(C_i)) + (1 - y_i) \log(1 - g_{detection}(C_i)) \tag{6}$$

## Knowledge Selection

The system selects the appropriate knowledge from the entire knowledge set if the dialogue context requires external knowledge. We train the matching model $g_{selection}$ between the conversation history and each knowledge pair to select appropriate knowledge. We use a RoBERTa (Liu et al., 2019) base model for the knowledge selection. The input format $x$ of RoBERTa is as follows:

$$x = [CLS]\ u_1\ s_2\ u_3\ ...\ u_m\ [SEP]\ k_j\ [EOS] \tag{7}$$

where $[CLS], [SEP]$ are class token and separator token, respectively. The final hidden vector of CLS token $T_{[CLS]}$ is used as the aggregate representation of the matching model and passed through the single linear function. If the given knowledge is related, the value is 1; if not related, the value is 0. Finally, the model is trained through cross-entropy loss for multi-class between related knowledge and negative samples as follow:

$$Loss = -\sum_i \sum_j y_j \log(g_{selection}(C_i, k_j)) \tag{8}$$

**Weighted Negative Sampling** In general, random knowledge other than target knowledge is used as a negative sample when training a selection model. However, since most random negative samples are easily distinguished from target knowledge, the model has a problem learning only on easily distinguishable samples. In this respect, multi-scale negative sampling (He et al., 2021) train the model by classifying negative samples into multiple categories is proposed. However, we argue that the multi-scale negative sampling method overlooks that the difficulty of negative samples is different per category. To this end, we propose a weighted negative sampling method that trains the model more fine-grained manner than previous methods. Our method gives different weight probabilities to each negative sample category to make the model focus on learning the negative samples that are difficult to distinguish.

Negative sample categories in our method are as follows :
**Random:** knowledge randomly selected from the entire knowledge set.
**In-entity:** knowledge randomly chosen from among the knowledge in the same entity.
**In-domain:** knowledge chosen from among the knowledge in the same domain.
**Semantically-similar:** knowledge arbitrarily selected from among similar knowledge. The similarity between knowledge is obtained through BERT similarity (Reimers and Gurevych, 2019).
Furthermore, the model may not train enough if the number of negative samples is inadequate for the weighted negative sampling method. Conversely, if there are many negative samplings, the false prediction of the model might increase. We set the appropriate number of negative samples to four through the experiments.

## Knowledge-grounded Generation

Given the knowledge related to the conversation history, the system needs to generate an appropriate response. Specifically, the response should maintain coherency and context flow for conversation context and contain the user's information from the selected external knowledge. For this sub-task, we use GPT2 base as the model and input the conversation history and one related knowledge to the model as follows:

$$x = [BOS][know]\ k_r\ [user]\ u_1\ ...\ u_m\ [EOS] \tag{9}$$

where [know] is the knowledge tag to mark the start of knowledge and $k_r$ is the most related knowledge from the selection module. The training objective for generation is

| Subtask | Task1: Turn Detection | | | Task2: Knowledge Selection | | | Task3: Response Generation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Precison | Recall | F1 | MRR@5 | Recall@1 | Recall@5 | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-1 | ROUGE-2 | ROUGE-L |
| Baseline | 0.985 | 0.634 | 0.771 | 0.540 | 0.444 | 0.690 | 0.133 | 0.053 | 0.023 | 0.012 | 0.145 | 0.165 | 0.044 | 0.118 |
| Knover | 0.970 | 0.625 | 0.760 | 0.578 | 0.526 | 0.666 | 0.130 | 0.063 | 0.032 | 0.011 | 0.159 | 0.157 | 0.053 | 0.124 |
| **Ours** | **0.990** | **0.961** | **0.975** | **0.831** | **0.780** | **0.887** | **0.414** | **0.327** | **0.259** | **0.194** | **0.491** | **0.492** | **0.312** | **0.469** |

Table 1: Experimental results on the validation set

| Subtask | Task1: Turn Detection | | | Task2: Knowledge Selection | | | Task3: Response Generation | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | Precison | Recall | F1 | MRR@5 | Recall@1 | Recall@5 | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-1 | ROUGE-2 | ROUGE-L |
| Baseline | 0.901 | 0.711 | 0.795 | 0.523 | 0.458 | 0.625 | 0.115 | 0.051 | 0.018 | 0.007 | 0.121 | 0.144 | 0.041 | 0.114 |
| Knover | 0.896 | 0.673 | 0.769 | 0.557 | 0.495 | 0.647 | 0.124 | 0.061 | 0.029 | 0.015 | 0.136 | 0.151 | 0.051 | 0.122 |
| **Ours (B12)** | **0.869** | **0.938** | **0.902** | **0.732** | **0.691** | **0.79** | **0.332** | **0.235** | **0.156** | **0.098** | **0.403** | **0.384** | **0.2** | **0.374** |

Table 2: Experimental results on the test set

| Model | Accuracy | Appropriateness | Average |
|---|---|---|---|
| Baseline: DSTC9 | 2.7425 | 2.7894 | 2.7659 |
| Baseline: Knover | 2.7793 | 2.7435 | 2.7614 |
| **Ours (B12)** | **3.2546** | **3.2336** | **3.2441** |
| Ground-truth | 3.5769 | 3.4814 | 3.5292 |

Table 3: Final human evaluation on the test set

| Method | MRR@5 | Recall@1 | Recall@5 |
|---|---|---|---|
| Random | 0.77 | 0.682 | 0.878 |
| Multi-scale | 0.783 | 0.712 | 0.887 |
| **Weighted (ours)** | **0.831** | **0.78** | **0.887** |

Table 4: Comparison of weighted negative sampling method on validation set.

the same with language modeling objective (Bengio et al., 2003), which maximize next word prediction probability as follow:

$$Loss = -\sum_{i} \log(g_{generation}(r|C_i, k_r)) \qquad (10)$$

**Post-training** Since pre-trained models (PLMs) are trained with general corpus, the context representation for the specific task may be insufficient. To address this issue, post-training methods (Gururangan et al., 2020; Han et al., 2021) that train PLMs once more with in-domain data before fine-tuning have been proposed. From this point of view, we post-train the model through the large TOD data, which contains generated `DSTC10` conversation, `DSTC9` conversation, `DSTC9` knowledge, `DSTC10` knowledge, and MultiWOZ. For post-training, we use the same training objective as pre-training, next-word prediction.

**Style Transfer** When we fine-tune the generation task only with constructed `DSTC10` training data, generated responses have a different style with a response from validation data. Therefore the performance of the automatic evaluation metric decreases even if the accuracy or consistency of the response with dialogue context is well enough. To make the model generate a response that has a similar style to the validation set, we additionally train the model using an extra dataset for style transfer.

# Experiments

## Settings

We evaluated the model with the validation set and test set provided by `DSTC10-Track2-Task2`. As mentioned in the Automatic Data Construction Section, we use our automated constructed `DSTC10` dataset for training.

In the detection task, `DSTC10` training set, which does not apply the noise injection process, was used. Precision, recall, and F1 were used as evaluation metrics. For the selection task, we trained the model through `DSTC10` train set. We measured performance through evaluation metrics such as $MRR@5$ (Voorhees, 1999), $recall@1$, and $recall@5$. For post-training in the generation task, `DSTC9` dialogs, generate `DSTC10` dialogs, `DSTC10` knowledge FAQ, and MultiWOZ are used. For style transfer, we learned the validation step and the test step with different data. For the validation step, generated `DSTC10` data and `DSTC9` test set were used, For test step, the `DSTC10` validation set was additionally used for training. We used $BLEU - 1/2/3/4$ (Papineni et al., 2002), $Meteor$ (Denkowski and Lavie, 2014), and $ROUGE - 1/2/L$ (Lin, 2004) for evaluation metric.

The `DSTC9` baseline (Kim et al., 2020) and the `DSTC9` top ranked model, $Knover$ (He et al., 2021), were used as baseline. These models are trained with `DSTC9` conversation data and knowledge.

## Experimental Result

Table 1 and Table 2 are the evaluation set and test set performance for each subtask. Except for the precision of the detection task in the test set, our model shows significant improvement in performance for all metrics compared to the baseline. The main reason is baseline models are trained with previous knowledge (`DSTC9` knowledge), while our model is trained with conversation data to reflect new knowledge (`DSTC10` knowledge). In addition, new methods such as weighted negative sampling, post-training, and style transfer, improved the performance in the selection task and the generation task. Table 3 shows the results of human evaluation. Our model shows enhanced performance in both accuracy and appropriateness compared to baselines.

| Method | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|---|---|---|---|---|
| Baseline | 0.132 | 0.071 | 0.035 | 0.017 | 0.198 | 0.237 | 0.096 | 0.207 |
| +Style_transfer | 0.360 | 0.275 | 0.201 | 0.140 | 0.429 | 0.445 | 0.260 | 0.418 |
| **+Post-training+style transfer** | **0.414** | **0.327** | **0.259** | **0.194** | **0.491** | **0.492** | **0.312** | **0.469** |

Table 5: Ablation study for generation task on the validation set.

# Further Analysis

## Effectiveness of Weighted Negative Sampling

We compared multiple negative sampling techniques to investigate the effect of our weighted negative sampling method in Table 4. Random denotes the random selection of negative sampling from the whole knowledge. Multi-scale retrieves negative samples by dividing them into several categories (random, in-entity, in-domain, semantically-similar). Weighted indicate a method of varying the probability of negative samples for each category. Weight for each category is manually set through the experiments. The ratio of the number of weighted negative samples in categories is 2:1:2:2 for random, in-entity, in-domain, and semantically similar. As shown in Table 4, a multi-scale negative sampling method is more effective than selecting random knowledge FAQ pairs as negative samples. Moreover, our weighted negative sampling method, which learns with different ratios for each category, shows significant performance improvement. This is because our method strengthens the ability to select relevant knowledge by seeing more difficult categories to distinguish.

## Effectiveness of Post-training and Style Transfer

We have experimented to identify the effects of post-training and style transfer for the generation task in Table 5. Baseline is trained for the generation task without post-training and style transfer. It shows the lowest performance across automatic evaluation metrics because the style of system response in the `DSTC10` training data is different from the validation set. On the other hand, the model with style transfer has significantly enhanced performance compared to baseline. It is because the model learns the responses style by training additional `DSTC9` test data, which has a similar response style. The post-trained and style transferred model shows improved performance compared to the model that performed style transfer alone. This is because the model learns in-domain representation through post-training and is optimized for the TOD data distribution in advance.

# Conclusion

In this work, we address `DSTC10-Track2-Task2`. To enhance the knowledge-grounded TOD system, we employ an automatic data construction method containing a noise injector module to generate spoken conversation data about `DSTC10` knowledge. We also propose weighted negative sampling to improve the knowledge selection model and apply post-training and style transfer for the generation task. The validity of our methods has been checked through experiments. As a final result, our approach ranked 6 on objective human evaluation.

We find the selection module has a significant effect on overall system performance. Therefore we plan to research the advanced knowledge selection model as future work.

# References


Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F.; and Weber, G. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4218–4222.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33.

Bengio, Y.; Ducharme, R.; Vincent, P.; and Janvin, C. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3:1137–1155.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Ultes, S.; Ramadan, O.; and Gašić, M. 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*, 5016–5026. Brussels, Belgium: Association for Computational Linguistics.

Chen, H.; Liu, X.; Yin, D.; and Tang, J. 2017. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter* 19(2):25–35.

Denkowski, M., and Lavie, A. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 376–380. Baltimore, Maryland, USA: Association for Computational Linguistics.

Eric, M.; Goel, R.; Paul, S.; Kumar, A.; Sethi, A.; Ku, P.; Goyal, A. K.; Agarwal, S.; Gao, S.; and Hakkani-Tur, D. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Gururangan, S.; Marasovic, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020,*

*Online, July 5-10, 2020*, 8342–8360. Association for Computational Linguistics.

Ham, D.; Lee, J.-G.; Jang, Y.; and Kim, K.-E. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 583–592. Online: Association for Computational Linguistics.

Han, J.; Hong, T.; Kim, B.; Ko, Y.; and Seo, J. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tür, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 1549–1558. Association for Computational Linguistics.

Hayashi, T.; Watanabe, S.; Zhang, Y.; Toda, T.; Hori, T.; Astudillo, R.; and Takeda, K. 2018. Back-translation-style data augmentation for end-to-end asr. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 426–433. IEEE.

He, H.; Lu, H.; Bao, S.; Wang, F.; Wu, H.; Niu, Z.; and Wang, H. 2021. Learning to select external knowledge with multi-scale negative sampling.

Henderson, M.; Thomson, B.; and Williams, J. D. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 263–272. Philadelphia, PA, U.S.A.: Association for Computational Linguistics.

Hosseini-Asl, E.; McCann, B.; Wu, C.; Yavuz, S.; and Socher, R. 2020. A simple language model for task-oriented dialogue. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kim, S.; Eric, M.; Gopalakrishnan, K.; Hedayatnia, B.; Liu, Y.; and Hakkani-Tur, D. 2020. Beyond domain apis: Task-oriented conversational modeling with unstructured knowledge access.

Kim, S.; Liu, Y.; Jin, D.; Papangelis, A.; Gopalakrishnan, K.; Hedayatnia, B.; and Hakkani-Tur, D. 2021. "how robust r u?": Evaluating task-oriented dialogue systems on spoken conversations.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 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*, 7871–7880.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81. Barcelona, Spain: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR* abs/1907.11692.

Mi, H.; Ren, Q.; Dai, Y.; He, Y.; Sun, J.; Li, Y.; Zheng, J.; and Xu, P. 2021. Towards generalized models for beyond domain api task-oriented dialogue. *Proceedings of the 9th Dialog System Technology Challenge*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, 311–318. ACL.

Prabhumoye, S.; Tsvetkov, Y.; Salakhutdinov, R.; and Black, A. W. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 866–876.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21:140:1–140:67.

Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, 3980–3990. Association for Computational Linguistics.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86–96.

Shi, W., and Yu, Z. 2018. Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1509–1519. Melbourne, Australia: Association for Computational Linguistics.

Tang, L.; Shang, Q.; Lv, K.; Fu, Z.; Zhang, S.; Huang, C.; and Zhang, Z. 2021. Radge relevance learning and generation evaluating method for task-oriented conversational system-anonymous version. In *AAAI-21 DSTC9 Workshop*.

Voorhees, E. M. 1999. The TREC-8 question answering track report. In Voorhees, E. M., and Harman, D. K., eds., *Proceedings of The Eighth Text REtrieval Conference, TREC 1999, Gaithersburg, Maryland, USA, November 17-19, 1999*, volume 500-246 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Wen, T.-H.; Vandyke, D.; Mrkšić, N.; Gašić, M.; Rojas-Barahona, L. M.; Su, P.-H.; Ultes, S.; and Young, S. 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*, 438–449. Valencia, Spain: Association for Computational Linguistics.

Williams, J. D., and Young, S. 2007. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language* 21(2):393–422.

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

Yang, Y.; Li, Y.; and Quan, X. 2021. UBAR: towards fully end-to-end task-oriented dialog system with GPT-2. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 14230–14238. AAAI Press.