End-to-End Document-Grounded Conversation with Encoder-Decoder Pre-Trained Language Model

Jinhyeon Kim,[1] **Donghoon Ham,**[2] **Jeong-Gwan Lee,**[2] **Kee-Eung Kim** [1,2]

[1] Graduate School of AI, KAIST, Daejeon, Republic of Korea
[2] School of Computing, KAIST, Daejeon, Republic of Korea
{jhkim, dhham, jglee}@ai.kaist.ac.kr, kekim@kaist.ac.kr

Abstract

The first track of the Ninth Dialog System Technology Challenge (DSTC9), "Beyond Domain APIs: Task-Oriented Conversational Modeling with Unstructured Knowledge Access," encourages the participants to build goal-oriented dialog systems with access to unstructured knowledge, thereby making it possible to handle diverse user inquiries outside the scope of API/DBs. It consists of three sub-tasks: knowledge-seeking turn detection, knowledge selection, and knowledge-grounded response generation. We claim that tackling these sub-tasks separately is neither parameter-efficient nor of better performance. In this paper, we present an end-to-end document-grounded conversation system that utilizes a pre-trained language model with an encoder-decoder structure. In the human evaluation, our dialog system achieved the accuracy score of 4.3082 and the appropriateness score of 4.2665, which ranked 9th out of 24 participant teams. Furthermore, we conduct an ablation study and show that the end-to-end encoder-decoder scheme enables more efficient use of parameters in the document-grounded conversation setting.

Introduction

Goal-oriented dialog systems assist users to fulfill their purposes, such as booking, by properly understanding and responding to the user queries. Traditionally, they have relied on internal databases and APIs, which are often insufficient for diverse user demands; for example, a database may not know specific information like if a hotel allows pets. However, such knowledge may be found in the FAQs and reviews, etc. From this motivation, the first track of the Ninth Dialog System Technology Challenge (DSTC9), "Beyond Domain APIs: Task-Oriented Conversational Modeling with Unstructured Knowledge Access" (Kim et al. 2020), encourages the development of goal-oriented dialog systems with access to the unstructured knowledge.

The research on the dialog systems with groundings to the natural language documents has largely focused on the chit-chat conversation. Many previous works (Ghazvininejad et al. 2018; Moghe et al. 2018) had utilized RNN-based sequence-to-sequence models. More recently, Transformer-based models have been adopted for the document-grounded conversation (Dinan et al. 2019; Gopalakrishnan et al. 2019).

Meanwhile, the transfer learning with the language models pre-trained on large unlabeled data has become a popular approach in many natural language processing tasks. Many recent works on the chit-chat conversation (Wolf et al. 2019b; Zhang et al. 2020) and task-oriented conversation (Budzianowski and Vulic 2019; Ham et al. 2020) have opted for this framework as well.

Recently, the pre-trained language models with encoder-decoder structure such as T5 (Raffel et al. 2020) have emerged and achieved the state-of-the-art on several benchmarks including SuperGLUE (Wang et al. 2019). They combine the generative capability of the auto-regressive decoder such as GPT-2 (Radford et al. 2019) with the bi-directional auto-encoder such as BERT (Devlin et al. 2019).

We propose an end-to-end document-grounded conversation system that utilizes a pre-trained language model with an encoder-decoder structure. Specifically, the encoder receives the given dialog history paired with each candidate knowledge, and the decoder generates the response based on the encoding of the most relevant pair, if any. This model stood at 9th place out of 24 participant teams in the final evaluation. To analyze which components are more meaningful, we further present an ablation study and show that the end-to-end encoder-decoder scheme enables more efficient use of parameters in the document-grounded conversation. The code for reproducing our results is available at our GitHub repository[1].

Task Description

Beyond Domain APIs: Task-Oriented Conversational Modeling with Unstructured Knowledge Access

We participated in the first track of DSTC9, "Beyond Domain APIs: Task-Oriented Conversational Modeling with Unstructured Knowledge Access" (Kim et al. 2020). It aims for a goal-oriented dialog system that can handle user requests outside the scope of the API/DBs by accessing the external unstructured knowledge.

The track consists of three tasks, as illustrated in Figure 1. The first task is *Knowledge-seeking Turn Detection*, where the model determines if the user's inquiry can be answered

[1] https://github.com/kaist-ailab/End-to-End-Enc-Dec-DSTC9

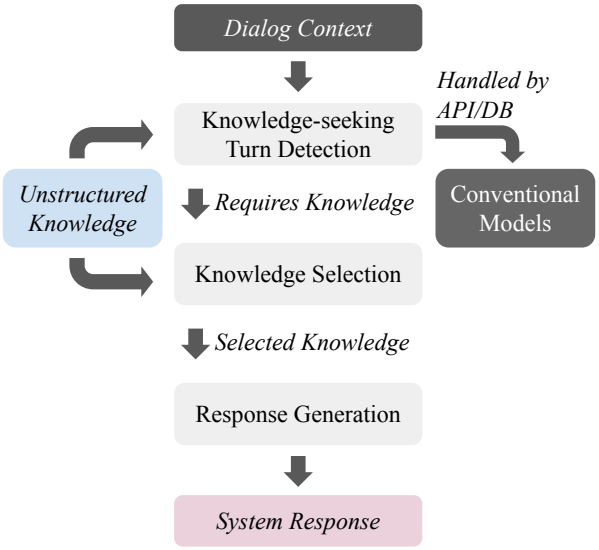

Figure 1: An overview of the first track of DSTC9

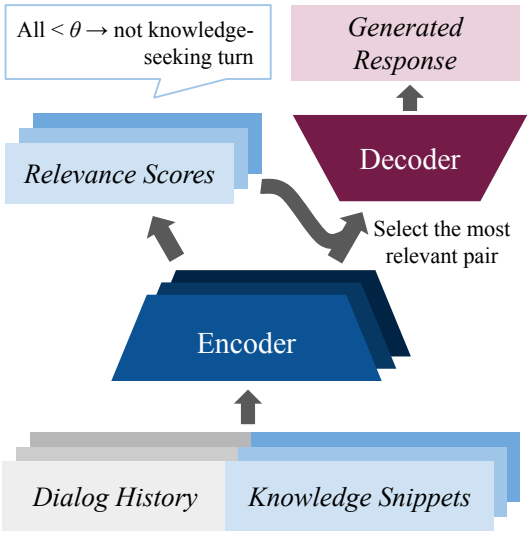

Figure 2: The illustration of our end-to-end encoder-decoder scheme for document-grounded conversation.

with the API/DBs. If not, it moves to the second task, *Knowledge Selection*, where the model chooses adequate knowledge. In the third task, *Response Generation*, the response is generated from the retrieved knowledge. The details of the track can be found in (Kim et al. 2020; Gunasekara et al. 2020).

Dataset

The track used the following two datasets. The first dataset is a modified version of MultiWOZ 2.1 dataset (Budzianowski et al. 2018; Eric et al. 2019) augmented with additional knowledge-seeking turns. A knowledge-seeking turn contains a question and an answer regarding a specific *knowledge snippet* about the domains and the entities of the MultiWOZ dataset. The first dataset is split into three subsets for train, validation, and test. The labels contain the ground-truth answers for the three tasks: whether the turn is knowledge-seeking or not, the relevant knowledge snippet, and the human response for the user utterance.

The second dataset is a newly collected dataset for the test phase, differing from the first dataset in the domain, entity, and locale, to evaluate the generalizability. It contains spoken conversation as well as written conversation, which may affect the robustness of the dialog system (Gopalakrishnan et al. 2020). The test phase evaluation was conducted both on the test split of the augmented MultiWOZ 2.1 and this new dataset.

Baseline

Kim et al. (2020) proposed several baseline methods for this track. They reported that leveraging the pre-trained language models was the best-performing approach. They trained binary classifiers based on BERT (Devlin et al. 2019) for Knowledge-seeking Turn Detection and Knowledge Selection tasks, and a language model based on GPT-2 (Radford et al. 2019) for Response Generation task. The baseline for Knowledge-seeking Turn Detection uses only the dialog context as the input, whereas the knowledge snippet is also given to the baselines for the other tasks.

End-to-End Document-Grounded Conversation with Encoder-Decoder Pre-Trained Language Model

In this section, we describe our end-to-end approach for the three tasks. The encoder solves both the Knowledge-seeking Turn Detection task and the Knowledge Selection task; the decoder solves the Response Generation task. Figure 2 illustrates the workflow of our approach.

Input Representation

Each input begins with `<bos>` token followed by the turns in the dialog history, each starting with the corresponding speaker token `<user>` or `<system>`. A candidate knowledge snippet follows the dialog history, with `<knowledge>` token in-between. The knowledge snippet is formatted as the concatenation of its three constituents, separated by `<knowledge-sep>` token: the name of the relevant entity (or domain if not applicable), the title of the knowledge snippet, and the body of the knowledge snippet. Both the dialog history and the knowledge snippet were truncated up to the length of 128 tokens.

Fine-tuning

T5 (Raffel et al. 2020) utilizes the encoder-decoder structure of Transformer (Vaswani et al. 2017). The encoder consists of several encoder blocks, each containing a self-attention layer and a fully connected layer. The decoder is shaped similarly except for the additional attention layer over the encoder output.

We fine-tuned the pre-trained T5 model in the following way. We put an additional fully-connected layer which takes

Recall@1	Recall@2	Recall@5	Recall@10
0.7355	0.9020	0.9895	0.9955

Table 1: Validation performance of knowledge snippet filtering method. Note that the scores are measured for the entities rather than the individual snippets.

the encoder output at the <bos> token to classify if the dialog history and the provided knowledge snippet are relevant. This binary classifier is trained with the binary cross-entropy loss L_{mc}. When the provided knowledge snippet is relevant to the dialog, the decoder is trained to generate the response based on the encoder outputs with the cross-entropy loss L_{lm}. The encoder and decoder are jointly fine-tuned under the multi-task loss $L = L_{lm} + \lambda L_{mc}$, where λ is the balancing coefficient between the two losses.

Inference

At inference time, we pass each of the candidate knowledge snippets to the encoder and sort them by their predicted probabilities of relevance. Only when some of them exceed the threshold $\theta = 0.5$, the dialog is classified as the knowledge-seeking turn, and the knowledge snippet with the highest probability is selected.

From the decoder, the tokens are sampled using the nucleus sampling with the probability threshold of $p = 0.9$, the temperature of $T = 0.7$, and the maximum length of 40 tokens, following the scheme used for the GPT-2 baseline of Task 3 (Kim et al. 2020).

Knowledge Snippet Filtering

We filtered the candidate knowledge snippets in the following three steps. Firstly, the entities are extracted from the dialog history using a rule-based algorithm. Secondly, the extracted entities from the first step are then sorted by the bi-gram TF-IDF retriever of DrQA (Chen et al. 2017). In order to measure the relevance of each entity to the dialog, we used the dialog history as the query, and the concatenation of all the snippets of the entity as the document. Finally, we select all the snippets pertaining to the top-10 relevant entities as the candidate knowledge snippets.

The rule-based entity extraction algorithm is as follows. Above all, we tokenized both the dialog history and the entity names by space and punctuation, and removed all the tokens that are included in the NLTK (Loper and Bird 2002) stop words for English. Note that the entity names are given as part of the knowledge snippets.

After that, we extracted *fingerprints* for each entity name, which are consecutive subsequences of tokens that are not part of any other entity names. We did not use any fingerprints containing only a single token that is one of the most frequent 250 tokens in the dialog corpus. When an entity name has no fingerprints, we used the entire entity name as the fingerprint.

Finally, we performed approximate string matching for fingerprints to recognize all the entities that appear in the dialog history. The approximate string matching covers the case where a single letter is added, dropped, or replaced in a token, and the case where two letters in the token are swapped. The reason behind this is that many entity names were misspelled in the training and validation dialogs; for example, *Huntingdon* Mariott Hotel was often confused with *Huntington* Mariott Hotel.

Entry ID	Description
0	Ensemble
1	$\lambda = 2, \eta = 5 \times 10^{-5}$
2	$\lambda = 1, \eta = 6.25 \times 10^{-5}$
3	$\lambda = 2, \eta = 6.25 \times 10^{-5}$
4	Our baseline

Table 2: Description of our submission entries.

Priority	Description		
Highest	$N = 10$	$\lambda = 2$	$\eta = 6.25 \times 10^{-5}$
	$N = 10$	$\lambda = 2$	$\eta = 5 \times 10^{-5}$
	$N = 10$	$\lambda = 1$	$\eta = 6.25 \times 10^{-5}$
	$N = 8$	$\lambda = 2$	$\eta = 6.25 \times 10^{-5}$
Lowest	$N = 8$*	$\lambda = 2$	$\eta = 6.25 \times 10^{-5}$

Table 3: Description of our ensemble entry. The priority is used for tie-breaking. The asterisk indicates that during training, TF-IDF ranking was applied to all the entities rather than those extracted from the dialog history.

Experiments

In this section, we present the objective and human evaluation results at the first track of DSTC9. Note that only the best submissions of the top-12 teams in the objective evaluation were subject to the human evaluation. We ranked at 10th place in the objective evaluation and 9th place in the human evaluation. We also analyze how each component of our approach affects performance via the ablation study.

Training Details

We used *transformers* library (Wolf et al. 2019a) for the pre-trained language model, where we used a `t5-large` model of 770M parameters. Unless noted otherwise, the experiments used AdamW optimizer (Loshchilov and Hutter 2019) with an effective batch size of 16, an initial learning rate of $\eta = 6.25 \times 10^{-5}$, linear learning rate decay for five epochs, and a multi-task balancing coefficient of $\lambda = 2.0$.

Each instance contained $N = 10$ knowledge snippets, among which the ground-truth knowledge snippet is included if it is a knowledge-seeking turn. We applied the knowledge snippet filtering before sampling the knowledge snippets. The validation performance of the knowledge snippet filtering is shown in Table 1.

Each team was allowed to submit at most five predictions for all three tasks. Table 2 describes the details of our submission entries. The ensemble model (Entry ID 0) used hard voting on the five T5-large models with varying training details as described in Table 3. We assigned priorities on the constituent models based on the validation performance to break any ties in the process. The next three entries (Entry

Entry ID	Task 1: Turn Detection			Task 2: Knowledge Selection			Task 3: Response Generation		
	Precision	Recall	F1	MRR@5	Recall@1	Recall@5	BLEU-1	METEOR	ROUGE-L
0	0.9984	**0.9278**	**0.9618**	**0.9233**	**0.8959**	**0.9555**	0.3523	0.3527	0.3500
1	**0.9989**	0.9122	0.9536	0.9161	0.8908	0.9478	0.3457	0.3496	0.3437
2	0.9842	0.9147	0.9482	0.9094	0.8802	0.9430	0.3485	0.3497	0.3469
3	0.9978	0.9011	0.9469	0.9137	0.8886	0.9427	**0.3534**	**0.3565**	**0.3519**
4	0.9905	0.8935	0.9395	0.8146	0.7622	0.8832	0.3181	0.3154	0.3200
Baseline	0.9933	0.9021	0.9455	0.7263	0.6201	0.8772	0.3031	0.2983	0.3039

Table 4: The objective evaluation results of our submission entries. Bold indicates the best scores.

ID 1, 2, 3) were the best single models in terms of validation at the time of submission.

The last entry (Entry ID 4) is our baseline model, which is similar to the baseline model of the track (Kim et al. 2020) except that our knowledge snippet filtering method is used for Task 2. We used `bert-base-uncased` of 109M parameters[2] for Task 1 and 2, and `gpt2-medium` of 345M parameters for Task 3.

Objective Evaluation

The objective evaluation results of our submissions are shown in Table 4. Note that since Task 2 and 3 are only evaluated on the true positives of Task 1, the scores in Task 2 and 3 are adjusted in a way roughly equivalent to multiplying the F1 score of Task 1. The ensemble model (Entry ID 0) performed better than or comparably to our other entries across all the tasks. Meanwhile, our baseline model (Entry ID 4) outperforms the track baseline on Task 2 even though it slightly under-performs on Task 1. This gap in the Task 2 performance probably measures the effect of our knowledge snippet filtering method. When we compare the performance of our single models (Entry ID 1, 2, 3) to that of our baseline (Entry ID 4), we see that the performance gain is significant even considering the model size difference. We inspect this aspect in the ablation study.

Human Evaluation

Table 5 shows the human evaluation results. For the human evaluation, crowd workers were asked to read the conversation between a user and an agent and rate the accuracy of the response with respect to the reference knowledge, and the appropriateness of the response in connection with the conversation, on a scale of 1-5. The scores are weighted with the knowledge-seeking turn detection performance, similarly to the objective evaluation. Only the best submissions of the top-12 teams in the objective evaluation were subject to the human evaluation. Our proposed model (ensemble) ranked 9th place out of 24 participants with the accuracy score of 4.3082, the appropriateness score of 4.2665, and the average score of 4.2874.

Ablation Study

In this section, we analyze how each component of our approach affects performance. Table 6 shows the validation

[2]We tried larger BERT models but could not fine-tune them successfully during the competition.

Rank	Team ID	Accuracy	Appropriateness	Average
	Ground-truth	4.5930	4.4513	4.5221
1	19	4.3917	4.3922	4.3920
2	3	4.3480	4.3634	4.3557
3	10	4.3544	4.3201	4.3373
4	15	4.3793	4.2755	4.3274
5	17	4.3360	4.3076	4.3218
6	7	4.3308	4.2989	4.3149
7	18	4.3309	4.2859	4.3084
8	13	4.3763	4.2360	4.3061
9	23 (Ours)	4.3082	4.2665	4.2874
10	11	4.2722	4.2619	4.2670
11	20	4.2283	4.2486	4.2384
12	21	4.1060	4.1560	4.1310
	Baseline	3.7155	3.9386	3.8271

Table 5: Overall results of the human evaluation.

performances of several ablated models. Figure 3 demonstrates some randomly sampled responses on the validation set from the ablated models, conditioned on the ground-truth knowledge snippet. The scores for Task 2 and 3 are evaluated both on the ground-truth and on the prediction of the previous tasks.

For easier comparison, we trained a smaller `t5-base` model of 220M parameters and labeled it as T5 in Table 6. To examine the effectiveness of the encoder-decoder structure, we also fine-tuned GPT-2 (Radford et al. 2019), which is an auto-regressive decoder-only model. We trained a `gpt2-medium` model of 345M parameters to jointly solve Task 1, 2, and 3 under a similar training scheme. We did not fine-tune an encoder-only model since it was hard to solve all the tasks jointly in an end-to-end manner.

As shown in Table 6 and Figure 3, GPT-2 significantly under-performed T5 on Task 3 while only producing comparable results on Task 1 and 2, even with 57% more parameters. This may be due to the lack of bi-directionality enjoyed by the encoder-decoder structure. This result provides supporting evidence for the advantage of the encoder-decoder structure in the knowledge-grounded conversation domain.

On the other hand, to assert the merit of end-to-end multi-task learning, we trained two additional models. T5 (Task 1 & 2) is the same as T5 except that it is trained on the multiple-choice loss L_{mc} only. It utilizes only the encoder part of the model. T5 (Task 3), on the other hand, is trained

Models	Task 1: Turn Detection			Task 2: Knowledge Selection			Task 3: Response Generation		
	Precision	Recall	F1	MRR	Recall@1	Recall@5	BLEU-1	METEOR	ROUGE-L
T5	0.9882	**0.9731**	0.9806	0.9622 (**0.9788**)	0.9470 (**0.9611**)	0.9791 (**0.9989**)	**0.3971** (**0.4065**)	**0.3997** (**0.4100**)	**0.3871** (**0.3981**)
Without encoder-decoder structure									
GPT-2	0.9931	0.9727	**0.9828**	**0.9635** (0.9727)	**0.9484** (0.9536)	**0.9802** (0.9951)	0.2675 (0.2675)	0.2752 (0.2761)	0.2841 (0.2850)
Without end-to-end multi-task training									
T5 (Task 1&2)	**0.9958**	0.9693	0.9824	0.9619 (0.9726)	0.9445 (0.9514)	0.9801 (0.9959)		-	
T5 (Task 3)		-			-		(0.4033)	(0.4082)	(0.3928)

Table 6: Validation performances of the ablated models. The scores evaluated on the ground-truth of the previous tasks are written in parentheses. Bold indicates the best scores.

on the language modeling loss L_{lm} only, conditioned on the ground-truth knowledge snippets. It takes advantage of both the encoder and decoder part of the model solely for the response generation. These two combined make up a 50% larger two-stage model for the three tasks. However, this two-stage model failed to produce a significant improvement over the end-to-end model. This demonstrates the benefit of end-to-end multi-task learning. We speculate that the three tasks are highly interrelated so that sharing the weight enabled the model to take advantage of the knowledge from other tasks and thereby resulted in more efficient use of parameters.

Related Works

Generating response grounded on unstructured knowledge has been an active research area in the field of non-goal-oriented conversation systems. Ghazvininejad et al. (2018) extended the sequence-to-sequence model with an additional recurrent network encoder for external facts in a similar manner to Memory Network (Sukhbaatar et al. 2015). Moghe et al. (2018) adapted a copy-or-generate model to the knowledge-grounded conversation. Gopalakrishnan et al. (2019) utilizes the Transformer model (Vaswani et al. 2017) where the decoder attends to the concatenation of the dialog history and the selected knowledge, encoded by a shared encoder.

The most similar to our work is Transformer Memory Network (Dinan et al. 2019). They compared the two-stage model and the end-to-end multi-task model for the two tasks of knowledge selection and response generation. The two-stage model employs two separate encoders and one decoder. The first encoder encodes each knowledge and each dialog context independently. After an attention-based knowledge selection, the encodings of the dialog context and the selected knowledge are concatenated, on which the second encoder and the decoder operates. The end-to-end model, on the other hand, does not introduce the second encoder, and the decoder directly processes the concatenated encoding.

Dinan et al. (2019) reported that the two-stage model outperformed the end-to-end model. However, we showed in our ablation study that the end-to-end training scheme performs better in our approach. We speculate that the lack of bi-directional incorporation of the dialog context and the knowledge was critical to the poorer performance of the end-to-end version of Transformer Memory Network. In our work, we fully utilize the capability of the bi-directional encoder.

Meanwhile, many recent works on dialog systems have adopted the transfer learning approach using large-scale pre-trained language models. TransferTransfo (Wolf et al. 2019b) and DialoGPT (Zhang et al. 2020) proposed the transfer learning schemes suited for the dialog systems, based on GPT-2 (Radford et al. 2019). Meena (Adiwardana et al. 2020) pre-trained a sequence-to-sequence Evolved Transformer (So, Le, and Liang 2019) on large-scale data to produce a sensible and specific conversation.

In the goal-oriented setting, Budzianowski and Vulic (2019) and Ham et al. (2020) fine-tuned GPT-2 models on the MultiWOZ dataset; Budzianowski and Vulic (2019) generated the response from the dialog context containing the belief state and the database state as well as the dialog history, whereas Ham et al. (2020) generated the response and intermediate states from the dialog history in an end-to-end manner. Our work lies in line with these transfer learning approaches, except that we utilize pre-trained language models with an encoder-decoder structure, which are more suited to the document-grounded conversation setting as we demonstrated in the ablation study.

Conclusion

In this paper, we presented an end-to-end encoder-decoder model for document-grounded conversation. Our method can be applied to any pre-trained models with an encoder-decoder structure. In the official evaluation results, we could observe that our proposed models outperform the baselines across all the tasks by significant margins. The human evaluation result showed that our approach can respond with reasonable accuracy and appropriateness. We further demonstrated by ablation study that the end-to-end encoder-decoder scheme enables more efficient use of parameters. One limitation of our methodology is the dependency on the heuristics and traditional information retrieval systems for knowledge snippet filtering. Overcoming this dependency

by integrating neural retrievers can serve as a future research direction.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00940, Foundations of Safe Reinforcement Learning and Its Applications to Natural Language Processing).

References

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* .

Budzianowski, P.; and Vulic, I. 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@EMNLP-IJCNLP 2019*.

Budzianowski, P.; Wen, T.-H.; Tseng, B.-H.; Casanueva, I.; Stefan, U.; Osman, R.; 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*.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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*.

Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2019. Wizard of Wikipedia: Knowledge-powered Conversational Agents. In *7th International Conference on Learning Representations*.

Eric, M.; Goel, R.; Paul, S.; Sethi, A.; Agarwal, S.; Gao, S.; and Hakkani-Tur, D. 2019. MultiWOZ 2.1: Multi-Domain Dialogue State Corrections and State Tracking Baselines. *arXiv preprint arXiv:1907.01669* .

Ghazvininejad, M.; Brockett, C.; Chang, M.-W.; Dolan, B.; Gao, J.; Yih, S. W.-t.; and Galley, M. 2018. A Knowledge-Grounded Neural Conversation Model. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Gopalakrishnan, K.; Hedayatnia, B.; Chen, Q.; Gottardi, A.; Kwatra, S.; Venkatesh, A.; Gabriel, R.; and Hakkani-Tür, D. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *20th Annual Conference of the International Speech Communication Association*.

Gopalakrishnan, K.; Hedayatnia, B.; Wang, L.; Liu, Y.; and Hakkani-Tür, D. 2020. Are Neural Open-Domain Dialog Systems Robust to Speech Recognition Errors in the Dialog History? An Empirical Study. In *21st Annual Conference of the International Speech Communication Association*.

Gunasekara, C.; Kim, S.; D'Haro, L. F.; Rastogi, A.; Chen, Y.-N.; Eric, M.; Hedayatnia, B.; Gopalakrishnan, K.; Liu, Y.; Huang, C.-W.; Hakkani-Tür, D.; Li, J.; Zhu, Q.; Luo, L.; Liden, L.; Huang, K.; Shayandeh, S.; Liang, R.; Peng, B.; Zhang, Z.; Shukla, S.; Huang, M.; Gao, J.; Mehri, S.; Feng, Y.; Gordon, C.; Alavi, S. H.; Traum, D.; Eskenazi, M.; Beirami, A.; Eunjoon; Cho; Crook, P. A.; De, A.; Geramifard, A.; Kottur, S.; Moon, S.; Poddar, S.; and Subba, R. 2020. Overview of the Ninth Dialog System Technology Challenge: DSTC9. *arXiv preprint arXiv:2011.06486* .

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*.

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. *arXiv preprint arXiv:2006.03533* .

Loper, E.; and Bird, S. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations*.

Moghe, N.; Arora, S.; Banerjee, S.; and Khapra, M. M. 2018. Towards Exploiting Background Knowledge for Building Conversation Systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

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. In *Journal of Machine Learning Research*.

So, D. R.; Le, Q. V.; and Liang, C. 2019. The Evolved Transformer. In *Proceedings of the 36th International Conference on Machine Learning*.

Sukhbaatar, S.; szlam, a.; Weston, J.; and Fergus, R. 2015. End-To-End Memory Networks. In *Advances in Neural Information Processing Systems*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Lan-

Example #1	
Dialog Context	
...	...
User	Yes, please. I'm going to need a taxi to take me from the hotel to the restaurant.
System	what time do you want to leave
User	I need to arrive at the restaurant by 12:45. I also need the contact number and car type please.
System	look out for a yellow skoda with contact number 07860378071
User	Will I be able to cancel my taxi booking later?
Relevant Knowledge	
Q:	What is the cancellation policy?
A:	You are able to cancel for free up to 24 hours in advance, we will charge 10 GBP after the cancellation time frame.
Generated Responses	
Ground-truth	You can cancel the taxi booking up to 24 hours in advance, after that it is 10 GBP. Do you have any other questions I can answer today?
T5	You can cancel for free up to 24 hours in advance. After that there is a 10 GBP charge. Is there anything else I can help you with?
GPT-2	nope, you may cancel up to 24 hours for free, 10 gbp after this time frame. Anything else for you today?
T5 (Task 3)	You can cancel up to 24 hours in advance. After that there is a 10 GBP fee. Do you have any other questions?

Example #2	
Dialog Context	
...	...
User	I'd like to know how I can get written confirmation. Is that something passengers can obtain?
System	Once the booking is complete, you will get a confirming email. I hope that's o.k.; any other questions about the train?
User	Will we be able to use wireless internet during the ride
Relevant Knowledge	
Q:	Is there WiFi available on the train?
A:	There will be basic WiFi service on the train, the ID is GreatNorthern_WiFi.
Generated Responses	
Ground-truth	Basic wi-fi service is available for passengers. Can I be of any other assistance?
T5	Yes, you'll be able to use wifi on the train. The ID is GreatNorthern_WiFi. Anything else I can do for you?
GPT-2	This train service provides basic WiFi services. Do you require assistance with anything else?
T5 (Task 3)	Yes, you can use wifi on the train. Is there anything else I can do for you?

Example #3	
Dialog Context	
...	...
User	Yes could you give me the phone number for The Nirala?
System	Their number is 01223 360966.
User	Which part of town it is in?
System	it is located in the north part of town
User	Does this restaurant allow customers to BYOB?
Relevant Knowledge	
Q:	Do you have BYOB?
A:	BYOB is not available at The Nirala.
Generated Responses	
Ground-truth	No, BYOB isn't allowed for guest. Do you need anything else?
T5	BYOB is not currently available at The Nirala. Do you have any other questions?
GPT-2	Is that okay?
T5 (Task 3)	No, BYOB is not an option at this restaurant. Do you have any other questions about the restaurant?

Figure 3: Randomly sampled responses on the validation set from the ablated models, conditioned on the ground-truth knowledge snippet. We only show the last few turns of the dialog contexts due to space constraints.

guage Understanding Systems. In *Advances in Neural Information Processing Systems*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2019a. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771* .

Wolf, T.; Sanh, V.; Chaumond, J.; and Delangue, C. 2019b. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *arXiv preprint arXiv:1901.08149* .

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2020. DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.