

UniGDD: A Unified Generative Framework for Goal-Oriented Document-Grounded Dialogue *

Chang Gao, Wenxuan Zhang, and Wai Lam

The Chinese University of Hong Kong

{gaochang, wxzhang, wlam}@se.cuhk.edu.hk

Abstract

The goal-oriented document-grounded dialogue aims at responding to the user query based on the dialogue context and supporting document. Existing studies tackle this problem by decomposing it into two sub-tasks: knowledge identification and response generation. However, such pipeline methods would unavoidably suffer from the error propagation issue. This paper proposes to unify these two sub-tasks via sequentially generating the grounding knowledge and the response. We further develop a prompt-connected multi-task learning strategy to model the characteristics and connections of different tasks and introduce linear temperature scheduling to reduce the negative effect of irrelevant document information. Experimental results demonstrate the effectiveness of our framework.

1 Introduction

Recent years have seen significant progress in goal-oriented dialogues (Loshchilov and Hutter, 2017; Wen et al., 2017; Wu et al., 2019; Hosseini-Asl et al., 2020; Peng et al., 2021), which aim at assisting end users in accomplishing certain goals via natural language interactions. However, due to the lack of external knowledge, most goal-oriented dialogue systems are restricted to providing information that can only be handled by given databases or APIs (Kim et al., 2020) and completing certain tasks in a specific domain such as restaurant booking. To address this challenge, goal-oriented document-grounded dialogue has been proposed to leverage external documents as the knowledge source to assist the dialogue system in satisfying users' diverse information needs (Feng et al., 2020; Wu et al., 2021).

* The work described in this paper is substantially supported by a grant from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project Code: 14200620).

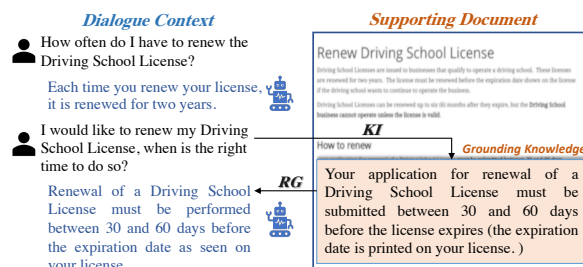


Figure 1: An example of the goal-oriented document-grounded dialogue problem.

As shown in Figure 1, the goal-oriented document-grounded dialogue problem is commonly formulated as a sequential process including two sub-tasks: knowledge identification (KI) and response generation (RG) (Feng, 2021). Given the dialogue context and supporting document, knowledge identification aims to identify a text span in the document as the grounding knowledge for the next agent response, which is often formulated as a conversational reading comprehension task (Feng, 2021; Wu et al., 2021). Response generation then aims at generating a proper agent response according to the dialogue context and the selected knowledge. Therefore, one straightforward solution for this problem is to use two models to conduct KI and RG in a pipeline manner (Daheim et al., 2021; Kim et al., 2021; Xu et al., 2021; Chen et al., 2021; Li et al., 2021). However, such pipeline methods fail to capture the interdependence between KI and RG. As a result, error propagation is a serious problem. The problem is more pronounced in low-resource scenarios, where accurate knowledge identification is difficult due to limited data, making it harder to generate appropriate responses.

To address the aforementioned issue, we propose a **Unified generative framework for Goal-oriented Document-grounded Dialogue (UniGDD)**. Given the dialogue context and associated document, instead of treating KI and RG as two separate processes, we tackle them simultaneously via sequen-

tially generating the grounding knowledge and the agent response. Therefore, the inherent dependencies between these two sub-tasks can be naturally modeled. On one hand, the generation of the agent response depends not only on the dialogue context and external document but also on the identified knowledge, forcing the model to focus on the specific knowledge. On the other hand, the generation of the grounding knowledge receives the supervision signal from the agent response when training, leading to more accurate knowledge identification.

Although KI and RG can be unified with the proposed generative method, they have different characteristics. Generating the grounding knowledge is similar to copying appropriate sentences from the document, while generating the response needs more effort to make the response coherent with the dialogue and consistent with the grounding knowledge. Therefore, in addition to the main task that uses the concatenation of the grounding knowledge and response as the target sequence, we introduce the generation of the grounding knowledge and the generation of the response as two auxiliary tasks in the same framework to force the model to capture their characteristics so as to perform well on them as well. Moreover, inspired by the recent success in prompt learning for pre-trained models (Li and Liang, 2021; Lester et al., 2021; Liu et al., 2021), we design prompts for these three tasks to guide the model on what to generate for each task. These prompts can naturally connect these tasks via indicating the model that each auxiliary task aims to generate a part of the target sequence of the main task. Through this prompt-connected multi-task learning strategy, the model can capture the characteristics of different tasks as well as exploit the connections between them.

In addition, for a particular user query in the goal-oriented dialogue, the selected knowledge and generated response need to be specific, while the generation conditions on a relatively long document. Thus, much information in the input document is irrelevant. To tackle this problem, we introduce linear temperature scheduling to make the attention distribution to the input document gradually sharper during the training process in order to enable the model to learn to pay more attention to the relevant content.

Our contributions are summarized as follows: (1) We propose a unified generative framework for the goal-oriented document-grounded dialogue. (2)

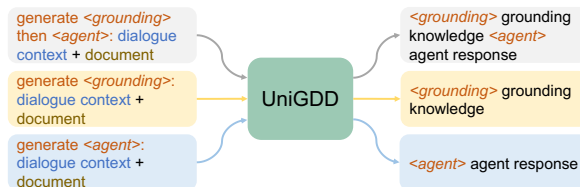


Figure 2: Overview of our framework.

We develop a prompt-connected multi-task learning strategy to exploit the characteristics and connections of different tasks and introduce linear temperature scheduling to enable the model to pay more attention to relevant information. (3) Our framework advances state-of-the-art methods on the concerned task, especially in low-resource scenarios.

2 Our UniGDD framework

UniGDD is a multi-task generative framework for the goal-oriented document-grounded dialogue problem.

Main Task Given the dialogue context $C = (u_1, a_1, \dots, u_{t-1}, a_{t-1}, u_t)$ and grounding document D , where u_i is the i -th user utterance and a_i is the i -th agent utterance, our main task aims to generate the target sequence $Y = (k_t, a_t)$, where k_t is the grounding knowledge from D and a_t is the response to u_t . Specifically, for the example in Figure 1, the input and output of the main task are as follows:

Input: generate <grounding> then <agent>:
 <user> I would like to renew ... ? <agent>
 Each time you ... <user> How often do ...
 ? <title> Renew Driving School License
 </title> ... Your application for renewal ...
Output: <grounding> Your application for
 ... <agent> Renewal of a Driving ...

We use different special tokens to identify different elements in the input and output. For example, we add "<user>" in front of each user utterance, "<agent>" in front of each agent utterance, and "<grounding>" in front of the grounding knowledge. The prompt "generate <grounding> then <agent>:" is added to the dialogue context and supporting document to form the input and guide the model to generate the grounding knowledge and the response in order. The input-to-target generation can be modeled with a pre-trained encoder-decoder model $\mathcal{M} : (C, D, TP) \rightarrow (k_t, a_t)$ such as T5 (Raffel et al., 2020), where TP is the task prompt.

Prompt-Connected Multi-Task Learning We introduce two auxiliary tasks to steer our framework to model the respective characteristics of knowledge identification and response generation. Given the dialogue context C and grounding document D , these two tasks aim to generate the grounding knowledge k_t and the response a_t with the same model \mathcal{M} . As depicted in Figure 2, we construct prompts "generate <grounding>:" and "generate <agent>:" for them. These prompts indicate the model that the goals of the two auxiliary tasks are to generate the first part and the second part of the target sequence of the main task, respectively. As a result, the connections between different tasks are naturally modeled. Instead of using discrete language phrases, we randomly initialize the embeddings of those special tokens in the prompts and train them end-to-end to better encode the characteristics and connections of these tasks.

Linear Temperature Scheduling For a specific user query in the dialogue, many document contents are actually irrelevant. To force the model to pay less attention to the irrelevant parts, we propose a linear temperature scheduling strategy to make the attention distribution of cross-attention gradually sharper during the training process. Specifically, we design the `softmax` function in the cross-attention module of each decoder layer as follows:

$$a_i = \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)} \quad (1)$$

$$\tau = (\tau_e - \tau_s) \frac{S_c}{S_{total}} + \tau_s \quad (2)$$

where a_i is the attention weight for the i -th input token, z_i is the logit for the i -th input token, S_c is the current training step, S_{total} is the total training steps, τ_s and τ_e are the starting and ending temperature respectively, $\tau_e < \tau_s$, and $0 < \tau_e < 1$. Compared with the original cross-attention module, the ending temperature $0 < \tau_e < 1$ leads to a sharper attention distribution, giving more attention weight to the relevant content.

Training The model is trained with a maximum likelihood objective. Given the training example $e = (C, D, TP, Y)$, the objective L_θ is defined as

$$\mathcal{L}_\theta = - \sum_{i=1}^n \log P_\theta(Y_i | Y_{<i}, C, D, TP) \quad (3)$$

where θ is the model parameters, TP is the task prompt, Y is the target sequence, and n is the

Models	EM	F1
BERTQA	42.2	58.1
BERT-PR-large	56.3	70.8
RoBERTa-PR-large	65.6	77.3
Multi-Sentence	59.5	68.8
DIALKI (\mathcal{L}_{next} only)	60.4	71.2
DIALKI	65.9	74.8
UniGDD-base	65.6	76.8
UniGDD-large	66.9	77.5

Table 1: Results on knowledge identification.

Models	BLEU
DIALKI+BART-base	25.8
RoBERTa-PR-large+BART-base	39.6
RoBERTa-large+T5-base	40.7
UniGDD-base	42.8
UniGDD-large	42.9

Table 2: Results on response generation.

length of Y . We mix the data of the main task and two auxiliary tasks for training.

Inference After training, for each pair of dialogue context and document (C, D) , we generate the target sequence of the main task for obtaining the grounding knowledge k_t and the response a_t .

3 Experiments

3.1 Experimental Setup

Dataset We conduct experiments on the goal-oriented document-grounded dialogue dataset Doc2Dial (Feng, 2021), which is adopted by the DialDoc21 shared task¹. It contains 3,474 dialogues with 44,149 turns for training and 661 dialogues with 8539 turns for evaluation².

Evaluation Metrics Following Feng (2021), we use Exact Match (EM) and token-level F1 for knowledge identification and BLEU (Papineni et al., 2002; Post, 2018) for response generation.

Baselines For knowledge identification, we compare UniGDD with several strong baselines, including BERTQA (Devlin et al., 2019), BERT-PR (Daheim et al., 2021), RoBERTa-PR (Daheim et al., 2021), Multi-Sentence (Wu et al., 2021), and DIALKI (Wu et al., 2021). These models formulate knowledge identification as the machine reading comprehension task and extract the grounding span

¹<https://github.com/doc2dial/sharedtask-dialdoc2021>

²Since we cannot access the test set, we report results on the development set for comparison.

from the document. For response generation, we compare UniGDD with several pipeline methods, including DIALKI+BART (Wu et al., 2021) that uses DIALKI to conduct knowledge identification, followed by BART (Lewis et al., 2020) to conduct response generation and RoBERTa-PR+BART (Daheim et al., 2021). We also build a strong baseline model RoBERTa+T5 which uses the same pre-trained generative model as ours.

Implementation Details We report results of UniGDD with two model sizes: UniGDD-base and UniGDD-large, which are initialized with pre-trained T5-base and T5-large models (Raffel et al., 2020), respectively. We adopt the implementation from Hugging Face Transformers (Wolf et al., 2020). We set the max input length to 2560. Any sequence over 2560 tokens will be truncated. For training, we use the AdamW (Loshchilov and Hutter, 2019) optimizer with an initial learning rate of 10^{-4} and a linear learning rate decay scheduler. We train 10 epochs for single-task learning and 5 epochs for multi-task learning. For decoding, we use beam search, and the beam size is 2. For linear temperature scheduling, we set the starting temperature $\tau_s = 1$ and choose the best ending temperature from $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. For our constructed baseline RoBERTa+T5 for response generation, we use RoBERTa-large and T5-base and adopt the implementation from the DialDoc21 shared task.

3.2 Results

The results on knowledge identification and response generation are shown in Table 1 and Table 2, respectively. Our UniGDD framework outperforms all the baselines on two sub-tasks. On the knowledge identification task, UniGDD-base can obtain comparable results to previous state-of-the-art methods. With a larger model size, UniGDD-large achieves new state-of-the-art performance. On the response generation task, UniGDD obtains a marked improvement over all pipeline methods. This verifies our assumption that our unified generative framework can alleviate the error propagation problem of pipeline approaches.

Effect of Prompt-Connected Multi-task Learning (PCMTL) and Linear Temperature Scheduling (LTS) To verify the effectiveness of PCMTL and LTS, we first remove PCMTL (i.e., training with the main task only), and the performance of UniGDD-base on two tasks decreases

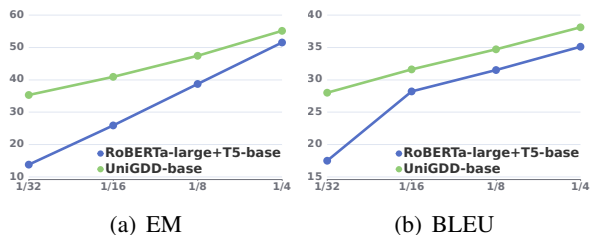


Figure 3: Experimental results on knowledge identification and response generation in low-resource scenarios


to 65.2 EM, 76.3 F1, and 42.3 BLEU, showing that PCMTL endows the model with the ability of modeling the characteristics and connections of different tasks and achieving better generation. Further removing LTS, the performance drops to 64.7 EM, 76.0 F1, and 41.7 BLEU. This indicates that LTS can guide the model to pay more attention to relevant content during generation and bring improvements on two sub-tasks.


Effect of Connected Prompts (CP) To examine whether CP can capture the connections of different tasks, we use an alternative approach that employs task-independent prompts "<Task1>", "<Task2>", and "<Task3>" to specify each task for comparison. As in the case of CP, we randomly initialize the embeddings of these three special tokens. With these prompts, UniGDD-base obtains 64.9 EM, 76.2 F1, and 42.3 BLEU, which performs worse than using CP. This indicates that CP enables the model to take advantage of the connections between the three tasks.


Low-Resource Setting To evaluate the model in low-resource scenarios, we randomly shuffle the training set and then take 1/32, 1/16, 1/8, and 1/4 of the data for training. Figure 3 shows the results of UniGDD-base and the best-performing pipeline baseline RoBERTa-large+T5-base on the four low-resource training splits. Generally, our framework performs substantially better than the pipeline method on both tasks. Particularly, when there is only 1/32 training data, UniGDD-base obtains more than 20 and 10 absolute points improvement over the pipeline approach on EM and BLEU, respectively.

Case Study Figure 4 shows a real case including the dialogue context, supporting document, and the responses generated by the pipeline method and our proposed UniGDD framework. It can be observed that our framework identifies accurate knowledge from the supporting document and thus provides a

Dialogue Context

 I filled out all of the information in the Retirement Estimator and it took a long time. When I came back from answering the door, all of the information was gone. What happened?


 Oh that's too bad. Were you gone for a long time?


 Yes I guess I was.

Supporting Document

..... How Long Can You Stay On Each Page? For security reasons, there are time limits for viewing each page. You will receive a warning after 25 minutes without doing anything, and you will be able to extend your time on the page. After the third warning on a page, you must move to another page. If you do not, your time will run out and your work on that page will be lost.

Response

 RoBERTa-large+T5-base
Do you have any more questions about the Retirement Estimator?

 UniGDD-base
For security reasons, there are time limits for viewing each page. You will receive a warning after 25 minutes without doing anything and you will be able to extend your time on the page.


 Ground Truth
For reasons of security, there are time limits for viewing each page.

Figure 4: A case from the development set.

proper and informative response about the reasons for the problem the user encounters. In contrast, the pipeline method only gives a relatively general response that is not suitable in this case.

3.3 Human Evaluation

We randomly sample 100 evaluation instances. For each instance, given the dialogue context and grounding document, three human annotators are asked to conduct a pairwise comparison between the response generated by UniGDD-base and the one generated by the pipeline baseline RoBERTa-large+T5-base in terms of two aspects: (1) *Relevance*: which response is more relevant and appropriate to the user query? (2) *Informativeness*: which response is more informative? Results are shown in Table 3. Compared with the pipeline method, our framework can reduce error propagation, resulting in more relevant and appropriate responses. Moreover, our framework has a clear advantage over the baseline in terms of Informativeness since it can utilize rich document context during the generation.

	Win	Tie	Lose
Relevance	26	64	10
Informativeness	23	69	8

Table 3: UniGDD-base vs RoBERTa-large+T5-base. The numbers indicate how many instances there are in each case.

4 Conclusion

Our UniGDD framework unifies knowledge identification and response generation and models their characteristics via a multi-task generative modeling strategy. Both automatic evaluation and human evaluation demonstrate the effectiveness of our framework.

References

- Xi Chen, Faner Lin, Yeju Zhou, Kaixin Ma, Jonathan Francis, Eric Nyberg, and Alessandro Oltramari. 2021. [Building goal-oriented document-grounded dialogue systems](#). In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 109–112, Online. Association for Computational Linguistics.
- Nico Daheim, David Thulke, Christian Dugast, and Hermann Ney. 2021. [Cascaded span extraction and response generation for document-grounded dialog](#). In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 57–62, 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.
- Song Feng. 2021. [DialDoc 2021 shared task: Goal-oriented document-grounded dialogue modeling](#). In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 1–7, Online. Association for Computational Linguistics.
- Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. [doc2dial: A goal-oriented document-grounded dialogue dataset](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. 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.
- Boeun Kim, Dohaeng Lee, Sihyung Kim, Yejin Lee, Jin-Xia Huang, Oh-Woog Kwon, and Harksoo Kim. 2021. [Document-grounded goal-oriented dialogue](#)

- systems on pre-trained language model with diverse input representation. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 98–102, Online. Association for Computational Linguistics.
- Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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.
- Jiapeng Li, Mingda Li, Longxuan Ma, Wei-Nan Zhang, and Ting Liu. 2021. Technical report on shared task in DialDoc21. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 52–56, Online. Association for Computational Linguistics.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.
- Ilya Loshchilov and Frank Hutter. 2017. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayan-deh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building Task Bots at Scale with Transfer Learning and Machine Teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.
- Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Tsung-Hsien Wen, Yishu Miao, Phil Blunsom, and Steve Young. 2017. Latent intention dialogue models. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3732–3741. PMLR.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.
- Zequ Wu, Bo-Ru Lu, Hannaneh Hajishirzi, and Mari Ostendorf. 2021. DIALKI: Knowledge identification in conversational systems through dialogue-document contextualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1852–1863, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yan Xu, Etsuko Ishii, Genta Indra Winata, Zhaojiang Lin, Andrea Madotto, Zihan Liu, Peng Xu, and Pascale Fung. 2021. [CAiRE in DialDoc21: Data augmentation for information seeking dialogue system](#). In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 46–51, Online. Association for Computational Linguistics.