SARCAT: Generative Span-Act Guided Response Generation Using Copy-Enhanced Target Augmentation

Jeong-Doo Lee[1], Hyeongjun Choi[1], Beomseok Hong[2], Youngsub Han[2], Byoung-Ki Jeon[2], Seung-Hoon Na[1]*,
{bo0od12, kaguya}@jbnu.ac.kr, bshong@lguplus.co.kr,
{yshan042, bkjeon}@lguplus.co.kr, nash@jbnu.ac.kr
[1] Department of Computer Science and Artificial Intelligence,
Jeonbuk National University
[2]LG Uplus

Abstract

In this paper, we present a novel extension to improve document-grounded response generation by proposing the *Generative Span-Act Guided Response Generation Using Copy-Enhanced Target Augmentation* (**SARCAT**), which consists of two major components: 1) *Copy*-enhanced *target-side input augmentation* is an extended data augmentation that handles the *exposure bias* problem by additionally incorporating the copy mechanism on top of target-side augmentation (Xie et al., 2021); 2) *Span-act guided* response generation first predicts the grounding spans and dialogue acts before generating a response. Experimental results on the validation set in MultiDoc2Dial show that the proposed SARCAT yields improvements over strong baselines in both seen and unseen settings, achieving start-of-the-art performance even with the *base* reader using the pretrained T5-*base* model.

1 Introduction

Recently, there has been a surge in research interest in developing dialogue systems grounded in knowledge and multiple documents (Zhao et al., 2022; Li et al., 2022c; Shuster et al., 2021; Zhao et al., 2020).

Existing approaches for document- and knowledge-grounded dialogue systems are based on the *open-book* approach, which is typically based on the retrieve-and-generate framework (Fan et al., 2021; De Bruyn et al., 2020; Li et al., 2022a), or the *close-book* approach, which relies on the scalability of pretrained language models.

This paper addresses the generation step in the open-book approach, i.e., *document-grounded response generation*, where the goal is to generate an appropriate response given a conversational history and retrieved contents.

In this paper, we propose the *Generative Span-Act Guided Response Generation Using Copy-Enhanced Target Argumentation* (**SARCAT**) method to improve the response generation module, i.e., consisting of two novel components, as follows:

- **Copy-enhanced target-side input augmentation (CAT)**:

  To further improve *target-side input augmentation* (TIA) as a promising method to relieve the *exposure bias* problem, (Bengio et al., 2015; Ranzato et al., 2016; Arora et al., 2022), we newly propose *copy-enhanced TIA* (CAT) by incorporating the *copy mechanism* into TIA, such that the resulting augmented sequence better matches the distribution at the inference time. The underlying motivation for CAT is that in document-grounded response generation, some parts of retrieved content are often required to be copied to a target sequence; thus, the synthetic soft words of conventional TIA might not be sufficiently close to the observed distributions at inference time.

- **Span-act guided response generation (SAR)**: Motivated by *chain-of-thought* prompting (Wei et al., 2022), we expect the prediction of *grounding spans* and *dialogue acts* to serve as a key intermediate reasoning chain for response generation. As an additional chain, we propose a two-step response generation: 1) *Span-act generation*, which predicts sequences of grounding spans and dialogue acts; and 2) *Response generation*, which generates a response by taking the predicted sequence as the reasoning chain.

Experimental results on the validation set in MultiDoc2Dial (Feng et al., 2021) demonstrate that SARCAT achieves state-of-the-art performance in

*Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14780–14787
November 12-16, 2024

both seen and unseen settings, even using the T5-base model, outperforming the best-performing system (Li et al., 2022a), which uses much larger parameters of pretrained models.

2 Related Work

Knowledge selection is the basic component guiding response generation in the existing document- and knowledge-grounded dialogue generation methods, including the long short-term memory (LSTM)-based sequential knowledge selector (Zhao et al., 2020), he sequential latent variable model (Zhao et al., 2020), and retrieval-based knowledge selection on an external database (De Bruyn et al., 2020).

In MultiDoc2Dial (Feng et al., 2021), e dense retrieval is typically performed as an initial step for knowledge selection under the *retriever-reader* framework (Li et al., 2022b; Bansal et al., 2022a; Zhang et al., 2022a), as in RAG(Lewis et al., 2020), which is the official baseline method. In particular, (Zhang et al., 2022a) further elaborated the method of guiding the response generation by relying not only on a set of retrieved passages, but also on the predicted spans which are much shorter than passages

Although (Zhang et al., 2022a) used a separate model for span prediction, their method employed a unified single T5 model to predict both spans and responses, similar to a strand of the chain-of-thought (Wei et al., 2022). Unlike (Zhang et al., 2022a) that directly predicts span tokens, our model instead predicts span markers with positional information, as we are largely motivated by the recent "marker"-based extensions of FiD (Izacard and Grave, 2020), such as FiD-Ex(Lakhotia et al., 2021) and PATHFID (Yavuz et al., 2022).

3 Methods

This section presents the details of the proposed components. Figure 1 illustrates an overall architecture of the proposed SARCAT framework, which consists of two components: copy-enhanced TIA and span-act guided generation.

3.1 Task Definition

Suppose that $\mathcal{H} = (u_1, \ldots, u_{T-1})$ is a dialogue history, $q = u_T$ is a given query at the current utterance time, and $\mathcal{P} = \{p_1, \ldots, p_m\}$ is a set of m passages retrieved in response to q and $\mathcal{H}$. The objective of document-grounded response generation is to produce an appropriate response u_{T+1}.

3.2 Copy-enhanced Target-Side Input Augmentation (CAT)

The core component of CAT is a modified way for constructing an augmented target sequence, being aware of the copy mechanism. To formally describe CAT, given a vocabulary set $\mathcal{V}$, let $x \subseteq (q, \mathcal{H}, \mathcal{P})$ be the encoder input.

Suppose that $y = (y_1, \ldots, y_n)$ is a *target* ground-truth response to generate, where $y_i \in \mathcal{V}$ and n is the length of tokens. TIA generates an augmented sequence $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)$, as in (Xie et al., 2021).

To generate $\tilde{y}_j$ at the j-th decoding time step, the decoder first emits the output probability as follows:

$$\mathbf{l}_j = \mathsf{T5}_{\text{dec}}\left(\mathsf{T5}_{\text{enc}}(x), y_{1:j-1}\right)$$

where $y_{1:j-1}$ indicates the previous target sequence (i.e. $y_1, \cdots, y_{j-1}$), $\mathsf{T5}_{\text{enc}}$ and $\mathsf{T5}_{\text{dec}}$ indicate the T5's encoder and decoder, respectively, and $\mathbf{l}_j \in \mathbb{R}^{|\mathcal{V}|}$ is a vector of logits, obtained before the softmax layer. We then obtain the *generate-mode soft word*, denoted as $\mathbf{p}_j^{gen} \in \mathbb{R}^{|\mathcal{V}|}$ as follows:

$$\mathbf{p}_j^{gen} = \text{softmax}\left(\mathbf{l}_t/\tau\right) \tag{1}$$

where τ is a temperature parameter.

Our novel part is to incorporates the *copy-mode soft word*, which is the probability distribution based on the copy mechanism, denoted as $\mathbf{p}_j^{copy}$, as follows:

$$\mathbf{p}_j^{copy} = \text{mean}_{l,h}\alpha_j^{l,h} \tag{2}$$

where $\alpha_j^{l,h} \in \mathbb{R}^{|\mathcal{V}|}$ denotes the attentive distribution over source tokens at the h-th head and l-th decoder layer.

We then interpolate the generate- and the copy-mode soft words to obtain the *copy-enhanced soft word*, denoted as $\mathbf{p}_j \in \mathbb{R}^{|\mathcal{V}|}$ as follows:

$$\mathbf{p}_j = (1-\beta)\mathbf{p}_j^{gen} + \beta\mathbf{p}_j^{copy} \tag{3}$$

We further apply the *sampling* to randomly choose between the copy-enhanced soft word $\mathbf{p}_j$ and ground-truth hard word $y_j \in \mathcal{V}$ with the probability γ. Formally, let z_j be a sample with a uniform distribution over $[0, 1]$. $\tilde{y}_j$ is thus defined as follows:

$$\tilde{y}_j = \begin{cases} \mathbf{p}_j & \text{if } z_j \leq \gamma \\ y_j & \text{otherwise} \end{cases} \tag{4}$$

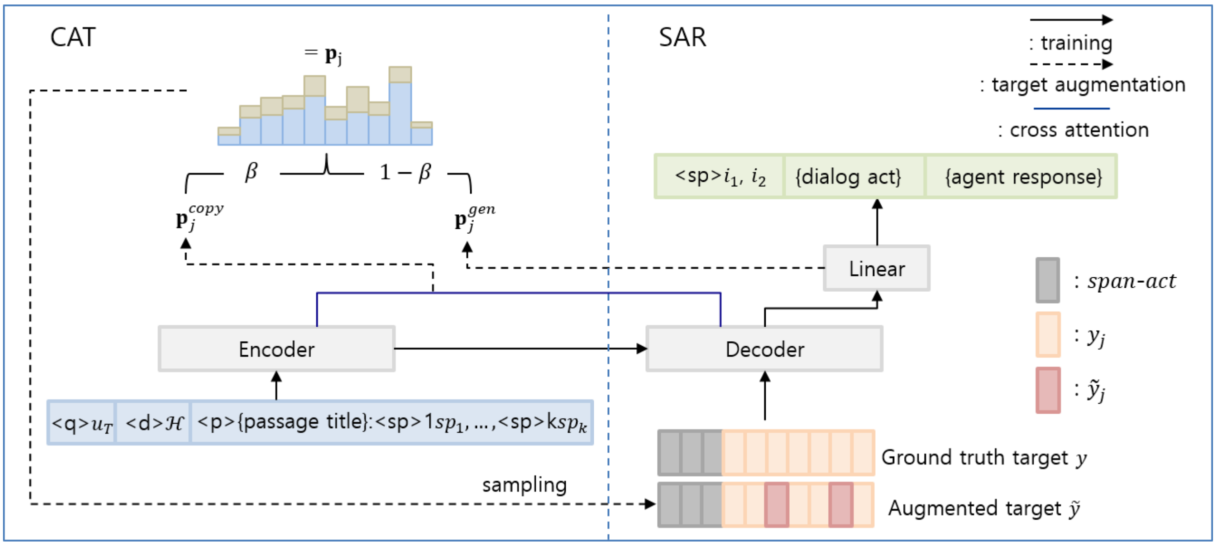

Figure 1: Overview of SARCAT architecture, consisting of two components – CAT and SAR; 1) CAT incorporates the *copy mechanism* into TIA based on the copy-enhanced soft word $\mathbf{p}_j$ (i.e., Eq. (3)) from the copy-mode soft word $\mathbf{p}_j^{copy}$ (i.e., Eq. (2)) and then applies the *sampling* mechanism to generate an augmented target $\tilde{y} = (\tilde{y}_j)_{i=1}^{n}$ (i.e,. Eq. (4))); 2) SAR performs the prediction of grounding spans and a dialog act as an auxiliary sequence generation task before generating a response (i.e., Section 3.3).

where $y_j \in \mathbb{R}^{|\mathcal{V}|}$ is treated as a probability vector with abuse of notation.

The remaining loss function is the same as that in (Xie et al., 2021), only being different in generating an augmented sequence $\tilde{y}$.

3.3 Span-Act Guided Response Generation (SAR)

In our method, the prediction of grounding spans and dialog acts is regarded as a sequence generation task. To be more specific, we introduce span markers for span prediction by prepending a span marker token "<sp>i" for i-th span sequence sp_i, as in (Yavuz et al., 2022; Lakhotia et al., 2021), while generating the corresponding gold tokens for dialog act prediction, thereby forming the input and output as follows:

input: "<q> u_T <d> $\mathcal{H}$ <p> {passage title}:<sp>1 sp_1, . . . , <sp>k sp_k"

output: "<sp> $i_1, \cdots, i_m$ {dialog act}: {agent response}"

where $i_1, \cdots, i_m$ are the predicted m span indices (i.e., $i_k \in \{1, \cdots, n\}$), <q>, <p>, <sp> are special tokens, {passage title} is the title of the retrieved passage, {dialog act} is the textual sequence of the dialog act to be predicted, and {agent response} is the response to be generated. Figure 2 shows an example of input and output sequences for SAR.

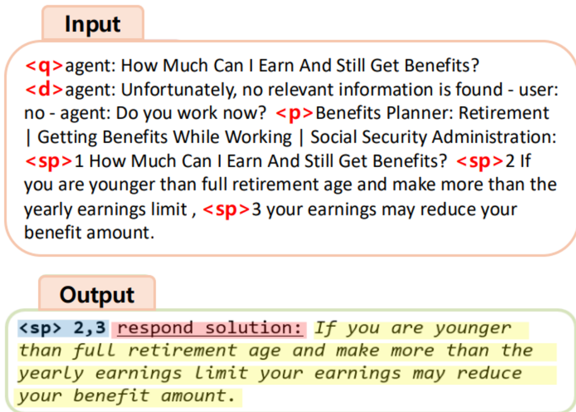

Figure 2: Sample input and output sequences for SAR. In the input at the top, **<·> text** indicates a special token. In the output at the bottom, **text** highlighted in blue is the output sequence for the *grounding span prediction*, text highlighted in red is the output sequence for the *dialog act prediction* and *text* highlighted in yellow is the original target sequence which is agent response

4 Experiments

4.1 Experimental Setup

Our experiments are conducted on the Multi-Doc2Dial shared task (Feng et al., 2021) with its official evaluation metrics (F1, SacreBLEU, METEOR and Rouge-L).

All experimental runs, including that for SARCAT, uses a T5-based model (Raffel et al., 2020) to train the response generator model. As the *base-*

Method	PLM(Size)	F1	SacreBLEU	METEOR	RougeL	Total
G4(Zhang et al., 2022a)	T5-base(220M)	44.60	31.24	42.41	42.68	160.93
R3(Bansal et al., 2022a)	T5-base(220M)	43.30	31.10	-	41.40	-
CPII-NLP(Li et al., 2022a)	BART-large(400M)	47.29	34.29	-	**46.04**	-
Baseline	T5-base(220M)	47.08	33.69	45.86	45.01	171.65
SARCAT	T5-base(220M)	**48.04**	**34.56**	**46.69**	45.93	**175.22**

Table 1: The results of response generation on the validation set of MultiDoc2Dial on the *seen* setting.

Method	PLM(Size)	F1	SacreBLEU	METEOR	RougeL	Total
CPII-NLP(Li et al., 2022a)	BART-large(400M)	36.74	24.20	-	35.49	-
Baseline	T5-base(220M)	35.68	20.38	32.64	33.95	122.65
SARCAT	T5-base(220M)	**36.85**	**24.55**	**35.11**	**35.61**	**132.12**

Table 2: The results of response generation on the validation set of MultiDoc2Dial on the *unseen* setting.

line run, we use our replicated version of (Li et al., 2022a), which deploys the grounding span prediction as an auxiliary task for the encoder, while excluding its passage dropout method. To obtain the retrieved content, we train the retrieval and reranking modules separately, obtaining a retrieval performance comparable to that of (Li et al., 2022b).

Method	Seen			Unseen		
	F1	SacreBLEU	RougeL	F1	SacreBLEU	RougeL
SARCAT	48.04	34.56	45.93	36.85	24.55	35.61
w/o **SAR**	47.03	33.42	44.89	35.63	23.12	34.20
w/o **CAT**	47.78	35.14	45.72	35.08	21.88	34.03

Table 3: Ablation results of SARCAT on the validation set on MultiDoc2Dial, obtained by excluding either **SAR** or **CAT**.

4.2 Main results

Tables 1 and 2 present the main results of SARCAT, comparing the baseline run and other previous methods, on both the seen and unseen settings. Here, the baseline run indicates SARCAT without either **SAR** or **CAT**.

As seen from Table 1, SARCAT consistently outperforms the baseline run, with increases of more than 0.8 in all evaluation metrics under the seen setting. These improvements are enlarged in the unseen setting, particularly in terms of SacreBLEU, METEOR, and RougeL, as shown in Table 2.

Notably, SARCAT achieves state-of-the-art performance by outperforming the CPII-NLP model (Li et al., 2022b), the previous best performing system in the MultiDoc2Dial shared task, under both seen and unseen settings.

Setting			F1	SacreBLEU	RougeL
Mixup	Sampling	Copy-enhanced			
✓			47.50	34.18	45.43
	✓		47.72	34.20	45.64
	✓	✓	48.02	34.34	45.94

Table 4: Ablation study of **CAT** on the validation set of the seen setting; 'Mixup' indicates the original TIA using the mixed distributions (not sampling);'Sampling' indicates the case of $\gamma < 1$ in Eq. (4); 'Copy-enhanced' indicates the case of $\beta > 0$ in Eq. (3).

4.3 Ablation studies

To examine the individual effects of CAT and SAR, Table 3 presents results obtained by SARCAT after excluding either CAT or SAR.

Interestingly, these effects vary between the seen and unseen settings; SAR performs particularly well in the seen setting, while CAT is stronger in the unseen setting.

One possible reason for the strong effect of CAT in the unseen setting is that copy-enhanced data augmentation helps relieve the lack of training data size in the unseen domain. In the case of SAR, the effect of predicting span grounding and dialogue acts may become stronger when a larger training dataset is used.

Ablation on Retrieval Component Table 5 and 6 present performance on in an oracle retrieval setting. Table 6 shows the results using the Wizard of Wikipedia(Dinan et al., 2018)(WoW) dataset. Since the WoW dataset does not have annotated dialog acts, the performance reported for SAR is without considering dialog acts.

Additionally, instead of providing gold sentences, we supply a gold passage composed of multiple sentences. We observe performance im-

Method	Seen				Unseen			
	F1	SacreBLEU	METEOR	RougeL	F1	SacreBLEU	METEOR	RougeL
baseline	54.82	41.85	53.89	53.00	43.10	26.84	40.06	41.71
w. SAR	55.89	**43.61**	54.29	54.10	43.43	29.39	40.51	41.91
w. CAT	55.26	42.29	54.33	53.36	44.92	29.03	42.46	43.12
w. SARCAT	**56.31**	43.48	**55.33**	**54.51**	**45.14**	**31.75**	**43.32**	**43.41**

Table 5: Performance using gold passage on the validation set of MultiDoc2Dial.

Method	F1	SacreBLEU	METEOR	RougeL
baseline	24.50	8.63	22.51	23.75
w. SAR	24.68	9.04	22.57	23.88
w. CAT	24.81	8.85	22.95	24.08
w. SARCAT	**25.14**	**9.35**	**23.23**	**24.31**

Table 6: Performance using gold passage on the test set of WoW.

provements on both the MultiDoc2Dial and WoW datasets when applying our proposed SARCAT model.

Analysis of CAT Table 4 presents the results of CAT without the copy-mode soft word or sampling mechanism., where the column labeled 'Mixup' indicates TIA (Xie et al., 2021) (i.e., $\beta = 0$ and $\tilde{y}_j = \mu \mathbf{p}_j + (1-\mu) y_j$ with $\mu = 0.5$). These results show that both sampling and the copy-mode soft word has similar effects on performance improvement.

Analysis of SAR Table 7 shows detailed results obtained by SAR after removing the intermediate steps. These results clearly show that SAR's performance gradually improves with the addition of more intermediate prediction steps.

Target	F1	SacreBLEU	RougeL
{**response**}	47.03	33.40	44.87
{**sp-response**}	47.34	34.10	45.32
{**sp-da-response**}	47.50	34.29	45.41

Table 7: Ablation study of **SAR** on the validation set of the seen setting; {**response**} is the case that directly generates a response without any span-act prediction; {**sp-response**} is the case that only adds the *grounding span prediction*; {**sp-da-response**} is the full-fledged generation including span-act prediction.

4.4 Case study

Figure 3 presents a case from the MultiDoc2Dial *seen* dataset, comparing responses generated by our model and the baseline model. Both SAR and SARCAT predict spans and dialog acts before generating a response, enabling them to formulate appropriate questions.

Case-1	
Dialogue	
User	Does the description of my issue come under borrower defense, do yo u know?
Agent	There are sections on the website dedicated to borrower defense and they will go through all questions and queries you may have
User	Regarding the discharge criteria, if I want to be 100 percent eligible for a discharge of my loans from the federal program, what requirements do I have to meet?
Knowledge	
Title	Discharge Criteria
SP1	You may be eligible for a 100 percent discharge of your William D. Ford Federal Direct Loan Direct Loan Program loans, Federal Family Educati on Loan FFEL Program loans, or Federal Perkins Loans if you were unab le to complete your program because your school closed, and if
SP2	you were enrolled when your school closed ;
SP3	you were on an approved leave of absence when your school closed ;
Response	
Gold	Were you enrolled when your school closed?
Baseline	You may be eligible for a 100 percent discharge of your William D. Ford Federal Direct Loan Direct Program loans, Federal Family Education Loan FFEL Program loans, or Federal Perkins Loans if you were unable to complete your program because your school closed
SAR	(SP2, query condition) Were you enrolled when your school closed?
CAT	You may be eligible for a 100 percent discharge of your loans if you were unable to complete your program because your school closed , and if you were enrolled when your school closed
SARCAT	(SP2, query condition) Were you enrolled when your school closed?

Figure 3: Case study. The blue ◯ denotes a gold span. We marked the selected span and the predicted dialog act in parentheses before each response.

5 Conclusion

In this paper, we presented **SARCAT** as novel fremework for improving document-grounded response generation. SARCAT consists of CAT, which explicitly incorporates a copy mechanism into TIA, enabling the infusion of context knowledge that is likely to be copied, and SAR, which initially generates a *span-act* sequence as an intermediate reasoning step to effectively guide the response generation process. Experimental results obtained using the MultiDoc2Dial dataset showed that the proposed framework outperformed the baseline run and achieved state-of-the-art performance.

6 Limitation

This paper proposes CAT by incorporating the copy mechanism into the TIA method to augment copy-enhanced target input, and subsequently evaluate CAT on a document-grounded response generation task, which represents a copy-aware generation task. However, CAT must also be validated on other generation tasks, such as the summarization task. It may also be interesting to determine whether CAT is extensible to tasks where the input might not appear in the target, such as machine translation tasks, It is therefore important to explore and validate CAT in various other types of text generation tasks.

In dialogue-grounded response generation, the performance of response generation systems such as SARCAT relies on a retrieval component. To minimize the effect from the retrieval component, it is also necessary to evaluate SARCAT on the "oracle" retrieval setting, where the gold retrieved content is assumed to be provided.

Furthermore, SAR has far been evaluated only for response generation. However, SAR also encompasses the subtask – span-act prediction – which must also be evaluated separately. In future work, it may be necessary to examine how effectively SAR predicts spans and dialog acts, and how strongly its impact on response generation is correlated with its span-act prediction performance.

Acknowledgments

This work was supported by LG Uplus Corporation. We would like to thank all anonymous reviewers for their valuable comments and suggestions.

References

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. 2022. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 700–710, Dublin, Ireland. Association for Computational Linguistics.

Srijan Bansal, Suraj Tripathi, Sumit Agarwal, Sireesh Gururaja, Aditya Srikanth Veerubhotla, Ritam Dutt, Teruko Mitamura, and Eric Nyberg. 2022a. R3 : Refined retriever-reader pipeline for multidoc2dial. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 148–154, Dublin, Ireland. Association for Computational Linguistics.

Srijan Bansal, Suraj Tripathi, Sumit Agarwal, Sireesh Gururaja, Aditya Srikanth Veerubhotla, Ritam Dutt, Teruko Mitamura, and Eric Nyberg. 2022b. R3: Refined retriever-reader pipeline for multidoc2dial. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 148–154.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2020. Bart for knowledge grounded conversations. In *Converse@ KDD*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2021. Augmenting transformers with KNN-based composite memory for dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. Multidoc2dial: Modeling dialogues grounded in multiple documents. *arXiv preprint arXiv:2109.12595*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Kushal Lakhotia, Bhargavi Paranjape, Asish Ghoshal, Scott Yih, Yashar Mehdad, and Srini Iyer. 2021. FiD-ex: Improving sequence-to-sequence models for extractive rationale generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3712–3727, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Kun Li, Tianhua Zhang, Liping Tang, Junan Li, Hongyuan Lu, Xixin Wu, and Helen Meng. 2022a. Grounded dialogue generation with cross-encoding re-ranker, grounding span prediction, and passage dropout. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 123–129, Dublin, Ireland. Association for Computational Linguistics.

Kun Li, Tianhua Zhang, Liping Tang, Junan Li, Hongyuan Lu, Xixin Wu, and Helen Meng. 2022b. Grounded dialogue generation with cross-encoding re-ranker, grounding span prediction, and passage dropout. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 123–129.

Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022c. Knowledge-grounded dialogue generation with a unified knowledge representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint*.

Shufang Xie, Ang Lv, Yingce Xia, Lijun Wu, Tao Qin, Tie-Yan Liu, and Rui Yan. 2021. Target-side input augmentation for sequence to sequence generation. In *International Conference on Learning Representations*.

Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, Nitish Shirish Keskar, and Caiming Xiong. 2022. Modeling multi-hop question answering as single sequence prediction. *arXiv preprint arXiv:2205.09226*.

Shiwei Zhang, Yiyang Du, Guanzhong Liu, Zhao Yan, and Yunbo Cao. 2022a. G4: Grounding-guided goal-oriented dialogues generation with multiple documents. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 108–114, Dublin, Ireland. Association for Computational Linguistics.

Shiwei Zhang, Yiyang Du, Guanzhong Liu, Zhao Yan, and Yunbo Cao. 2022b. G4: Grounding-guided goal-oriented dialogues generation with multiple documents. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 108–114.

Xueliang Zhao, Tingchen Fu, Chongyang Tao, Wei Wu, Dongyan Zhao, and Rui Yan. 2022. Learning to express in knowledge-grounded conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2258–2273, Seattle, United States. Association for Computational Linguistics.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. 2020. Knowledge-grounded dialogue generation with pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Appendix

A Data statistics

Split	Setting	Num of Instances	Num of Passages
Train	Seen	21451	3820
Validation	Seen	4181	3820
	Unssen	121	962
Development	Seen	199	3820
	Unseen	417	962
Test	Seen	661	3820
	Unseen	126	962

Table 8: Data statistics of MultiDoc2Dial dataset. We split a single dialogue into multiple instances of the train and validation set.

Table 8 presents statistics of the MultiDoc2Dial datasets for training, validation, blind development, and blind testing. We remove duplicate passages under both settings(i.e. *seen* and *unseen* set) and exclude repeated queries from the validation set, as in (Li et al., 2022b). Following official data preprocessing, the numbers of passages in the *seen* and *unseen* sets are 4110 and 963, respectively, with 4201 and 121, corresponding validation instances.

B Implementation Details

B.1 Passage retriever

We employ the standard *retriever* and *reranker* architectures for passage retrieval. Table 9 compares the results of passage retrieval in the *retriever-reranker* framework with those of previous works.

It is shown that our retriever model significantly outperforms all existing models and the reranker

Method	Seen			Unseen		
	R@1	R@5	R@10	R@1	R@5	R@10
Official-baseline(Feng et al., 2021)	49.0	72.3	80.0	-	-	-
G4(Zhang et al., 2022b)	39.5	68.5	77.3	-	-	-
R3(Bansal et al., 2022b)	-	-	78.6	-	-	-
+ reranker	-	-	85.7	-	-	-
CPII-NLP(Li et al., 2022b)	44.5	71.4	-	24.8	47.1	-
+ reranker	**69.6**	85.8	-	62.0	74.4	-
Ours	52.2	77.7	85.0	30.6	59.5	68.6
+ reranker	68.4	**86.2**	**90.8**	**62.0**	**74.4**	**84.3**

Table 9: Retrieval performance of the passage retrieval under the retriever-reranking framework, comparing to the existing models on the validation set. **R@K** represents Recall@K.

model shows comparable performances to CPII-NLP (Li et al., 2022b), the current best-performing system on MultiDoc2Dial.

B.2 Generator

We set the maximum input (i.e., query and passage) length to 512, and the max target length to 60. For training, we use a ground-truth passage. We set γ=0.15, τ=4, β=0.3 and use the double-round augmentation for CAT in Section 3.2. We employ AdamW as an optimizer with the linear scheduler, warmup proportion=0.06, peek learning rate=3e-4, batch size=32, and weight decay=0.01. All models were trained on two NVIDIA RTX A6000 GPUs over seven epochs. We use top-1 passages retrieved from the reranker module in Appendix B.1 for inference. We employ the beam-search with the beam size of 5.

C Ablation study of iterative CAT

Table 10 shows the effect of varying the number of applying CAT on the validation on the seen setting, where iteration 0 means no data augmentation The results exhibit that the overall score increases along with the number of iterations, for most evaluation metrics.

Iter	F1	SacreBLEU	RougeL
0	47.50	34.29	45.41
1	47.83	34.56	45.75
2	48.02	34.34	45.94

Table 10: Ablation study of the "iterative" data augmentation for the validation set on seen setting. **Iter** indicates the number of the iterations of CAT for data augmentation

D Ablation Study of Hyper-Parameters

Table 11-12 presents the results of additional experiments to examine the effects of β, τ.

β	0.0	0.1	0.2	0.3	0.4	0.5
F1	47.72	47.66	47.92	**48.02**	47.76	47.60
SacreBLEU	34.20	34.08	34.23	**34.34**	34.01	33.96
RougeL	45.64	45.61	45.92	**45.94**	45.62	45.53

Table 11: Ablation study of β on the validation set on the seen setting.

τ	2	4	6	8
F1	47.50	**48.02**	47.39	47.53
SacreBLEU	34.05	**34.34**	33.36	34.21
RougeL	45.50	**45.94**	45.30	45.55

Table 12: Ablation study of τ on the validation set on the seen setting.