Improving Multiple Documents Grounded Goal-Oriented Dialog Systems via Diverse Knowledge Enhanced Pretrained Language Model

Yunah Jang[1] **Dongryeol Lee**[1] **Hyungjoo Park**[1] **Taegwan Kang**[1]
Hwanhee Lee[1] **Hyunkyung Bae**[1] **Kyomin Jung**[2]
[1]Department of Electrical and Computer Engineering, Seoul National University, Seoul, Korea
[2]Automation and Systems Research Institute, Seoul National University, Seoul, Korea
{vn2209, drl123, harry0816, zd9370, wanted1007, hkbae, kjung}@snu.ac.kr

Abstract

In this paper, we mainly discuss about our submission to MultiDoc2Dial task, which aims to model the goal-oriented dialogues grounded in multiple documents. The proposed task is split into grounding span prediction and agent response generation. The baseline for the task is the retrieval augmented generation model, which consists of a dense passage retrieval model for the retrieval part and the BART model for the generation part. The main challenge of this task is that the system requires a great amount of pre-trained knowledge to generate answers grounded in multiple documents. To overcome this challenge, we adopt multi-task learning, data augmentation, model pre-training and contrastive learning to enhance our model's coverage of pretrained knowledge. We experiment with various settings of our method to show the effectiveness of our approaches. Our final model achieved 37.78 F1 score, 22.94 SacreBLEU, 36.97 Meteor, 35.46 RougeL, a total of 133.15 on DialDoc Shared Task at ACL 2022 released test set.

1 Introduction

Recently, deep learning-based dialog systems have attracted much attention from academia and the industry. The main challenge of dialog systems is to generate fluent responses consistent with the users' text input. As Pre-trained Language Models (PLMs) (e.g., BART (Lewis et al., 2019) and GPT2 (Radford et al., 2019)) have emerged, dialog systems have taken advantage of PLMs (Zhao et al., 2020; Wu et al., 2019; Budzianowski and Vulic, 2019), which can enhance the quality of dialog response by applying implicit language knowledge.

However, these systems lack knowledge of specific topics and thus show weakness in conducting an in-depth conversation with humans. There have been various works for knowledge-grounded dialogue systems to address this problem. (Kim et al., 2020; Zhan et al., 2021) Knowledge grounded dialogue models are capable of generating precise responses based on both the dialogue context and external sources. Therefore, researchers have usually constructed dialogue flows grounded in related documents (Dinan et al., 2018; Zhou et al., 2018b) or knowledge graphs (Moon et al., 2019; Zhou et al., 2018a; Tuan et al., 2019). In particular, Feng et al. (2020) have introduced the Doc2Dial tasks for goal-oriented document-grounded dialog systems. Compared to previous works, Doc2dial has introduced a more challenging setting with multi-turn queries and aims to generate natural language responses from relevant grounding document. On top of that, they also propose the MultiDoc2Dial dataset (Feng et al., 2021) ,which is built upon the Doc2Dial dataset. MultiDoc2Dial dataset is more closely related to real-life scenarios than the prior work since the agent generates responses based on multiple documents as grounding knowledge. Due to its multi-document setting, utilizing knowledge has become more complex.

To utilize external knowledge in dialogue, knowledge grounded models generally consist of a retrieval model and a generative model. Recently, the Retrieval Augmented Generation (RAG) model (Lewis et al., 2020a) has been proposed to leverage both parametric (Raffel et al., 2019; Lewis et al., 2019) and non-parametric memory (Lewis et al., 2020b; Xiao et al., 2020) methods by combining pre-trained seq2seq models and the dense vector index of grounding documents. However, the RAG model lacks knowledge related to question answering and dialogue generation.

In this paper, our team JPL proposes four approaches to enhance RAG's diverse knowledge: multi-task learning, data augmentation, pretraining and contrastive learning. Multi-task learning, extra pretraining on conversational question answering datasets, and data augmentation enhance the model's task-oriented knowledge. Contrastive

Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 136 - 141

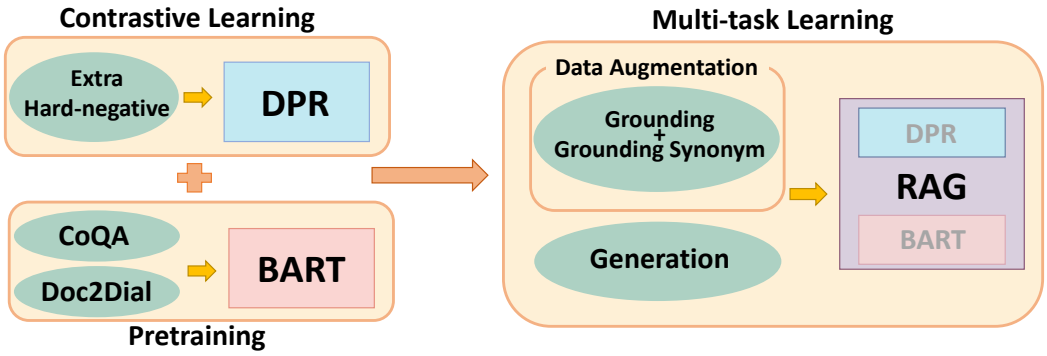

Figure 1: **Our training pipeline** We utilize four methods to cultivate the RAG model's diverse knowledge. To enhance model's task-agnositic knowledge, we add a hard negative sample for contrastive learning on the DPR retriever module. Pretraining BART with conversational QA datasets, data augmentation on grounding task, and multi-task learning improves task-specific knowledge for the final RAG model.

learning for the DPR retriever module strengthen task-agnostic knowledge. We participate in the second DialDoc shared task held by ACL, MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents (Feng et al., 2021). These methods cultivate the dialogue model's capability to use complex external knowledge on top of PLM's inherent power.

Splits	Train	Val	Test
Dialogues	3474	661	661
Queries	21453	4201	4094
Passages(struct)		4110	

Table 1: **Dataset Statistics** We split documents by using structural information from markup tags integrated in HTML files.

2 Shared Task

2.1 Dataset

In this shared task, we focus on the MultiDoc2Dial dataset (Feng et al., 2021), which contains conversations that are grounded in multiple documents. The dataset is constructed based on the Doc2Dial dataset, the dataset for the prior shared task at the DialDoc 2021 workshop. Unlike its predecessor, each dialogue in the MultiDoc2Dial dataset has multiple segments with different grounding documents for adjacent segments. The dataset consists of 4800 dialogues with an average of 14 turns that are grounded in 488 documents from four different domains (dmv, ssa, studentaid, va). Details of the MultiDoc2Dial dataset are given in Table 1.

2.2 Multidoc2dial

For the evaluations on MultiDoc2Dial dataset, two sub-tasks are proposed. Task 1 aims to predict the grounding span for the next agent response. For task 1, we get (1) current user turn, (2) dialogue history, (3) the entire set of documents from all domains as input. For the output, we aim to figure out the most relevant grounding text span from one document for the next agent response. Task 2 aims to generate agent response in natural language. For task 2, we get (1) current user turn, (2) dialogue history, (3) the entire set of documents from all domain as an input.

2.3 Baseline Model

In this shared task, the author proposed a baseline model based on the HuggingFace RAG. [1] For the retriever part, DPR (Karpukhin et al., 2020) was given in the form of both finetuned DPR encoders by author [2] and the original Facebook DPR. [3] The generator module of the baseline is BART-large from the HuggingFace.[4] Our final submission model is composed of our own fine-tuned DPR and Bart-large pretrained with conversational QA datasets.

3 Methodology

We use four methods to enhance the model's ability to efficiently utilize external grounding knowledge especially on dialogue modeling.

[1] https://huggingface.co/docs/transformers/master/model_doc/rag
[2] https://huggingface.co/sivasankalpp
[3] https://github.com/facebookresearch/DPR
[4] https://huggingface.co/facebook/bart-large

3.1 Multi-task Learning

Multi-task learning improves the model's performance when different tasks share information or semantics. If the tasks have a higher correlation, it is likely for the model to benefit more from multi-task learning. The final goal of the proposed task is to generate natural language responses, which corresponds to the generation task. Figure 2 presents the similarity between the ground truth of each task. From this statistic, it is clear that two tasks share much semantic information.

In order to implement multi-task learning, we first train the model on the grounding task with prefix "TASK1: " added to the input string for the generator. Then, using the last checkpoint, we continue training the model on the generation task with prefix "TASK2: " concatenated to each input string.

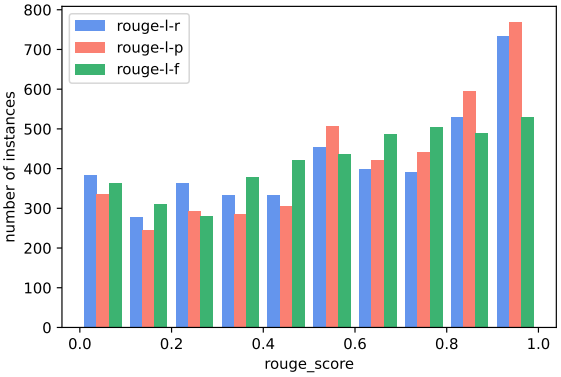

Figure 2: Similarity score of ground truth answer on grounding and generation task

3.2 Data Augmentation

To enhance the adaptability of the RAG model to the dataset, we attempt to increase the amount of data for finetuning. For each dialogue query in the original dataset, we apply the synonym augmenter from nlpaug[5]. The synonym augmenter randomly changes some words in the input to similar words based on WordNet[6]. We exclude '[SEP]', 'agent:' 'user:' since these words are special tokens for the task.

3.3 Pretraining on Conversational QA Datasets

To enhance the generative performance of the model, we pretrain the RAG generator on two datasets.

CoQA The first dataset is the CoQA dataset (Reddy et al., 2018), a conversational QA dataset grounded in a diverse range of documents. Because MultiDoc2Dial is not a large dataset, there is always a possibility of underfitting. CoQA, with its 127k questions, can provide us with much-needed extra data for our generator. As the format of the CoQA dataset (grounding document, then questions) is different from the input format of our BART model (query and dialogue context, followed by the grounding document), we reformat the dataset to fit our needs before training.

Doc2Dial The second dataset is the Doc2Dial dataset (Feng et al., 2020), a goal-oriented document-grounded dialogue dataset which is extremely similar to the MultiDoc2Dial dataset. As mentioned above, most of the instances in the MultiDoc2Dial dataset are formed by modifying Doc2Dial instances to fit a multi-document setting. Along with the existence of a single grounding document, this extreme similarity of content makes it an ideal candidate to train our generator without relying on the proper functioning of the retriever. Therefore, we can expect pretraining the generator on the Doc2Dial dataset to boost the generative capabilities of our model. As with CoQA, we reformat the dataset to fit the input of our BART model before training.

For both datasets, we do not cut down the grounding document to fit the maximum input length of our model. This may have resulted in truncation of the relevant span in some instances, and remains an area of possible improvement.

3.4 Contrastive Learning

To enhance the retrieval performance of the model, we adopt data augmentation to increase the number of hard negative contexts in the DPR training data. We apply the antonym augmenter from nlpaug[7]. The antonym augmenter takes positive contexts, which is the correct grounding document for the dialogue, as input. Based on WordNet antonym, the augmenter switches some words in the inputs to their respective antonyms and outputs the augmented sentences. We consider these outputs as the hard negative contexts and added them to the original dataset. We use the augmented dataset to finetune DPR.

[5]https://github.com/makcedward/nlpaug
[6]https://wordnet.princeton.edu/
[7]https://github.com/makcedward/nlpaug

4 Experiments

4.1 Training Details

We fine-tune RAG by following the default hyperparameter settings from the baseline code.[8] Due to hardware shortage, there are minor modifications; we set the gradient accumulation step as 2 and reduce the training and evaluation batch size to 4 and 1, respectively. We only report results of utilizing document structural information for segmentation since it shows better results in our experimental settings. The retrieved documents are not re-ranked since this method doesn't benefit the model performance.

4.2 Results and Analysis

Model	F1	EM	S_Bleu
baseline	34.69	3.86	20.63
+Multi-task learning	34.85	3.98	19.86
+Data Augmentation	33.55	3.28	19.01

Table 2: **RAG Fine-tuning Methods Results** Models are evaluated with F1, Exact Match, and sacreBLEU scores. The baseline model is composed of the released version of finetuned DPR[9] and BART-large on the HuggingFace.

4.2.1 RAG Fine-tuning Methods

In this section, the Facebook released version of DPR and BART-large in the HuggingFace constitute the baseline model.

Multi-task Learning We sequentially fine-tune the model on the grounding and generation tasks. Table 2 shows the results for multi-task learning. There are improvements in the F1 and EM score using multi-task learning, even though considering the fact that the model was trained on the generation task for a much shorter time. We expect the model to show better results with more extended training.

Data Augmentation For data augmentation, we apply synonym transformation to the original dataset, attaining twice the baseline size. Table 2 presents the result for data augmentation on generation task. We have observed that applying data augmentation to the generation task degraded the performance. However, by utilizing augmented data on the grounding task, the model achieves a 40.55 F1 score and a 23.49 exact match score. Compared to our baseline model implementation trained with the original grounding task data, training with augmented data improved +0.5 F1 score and +0.64 exact match score. These results demonstrate that synonym data augmentation on the generation task's gold answers does not provide the model with any informative knowledge for the generation task. Therefore, we include augmented data only on grounding task during multi-task learning.

Model	F1	EM	S_Bleu
baseline	34.69	3.86	20.63
+CoQA	35.08	4.02	20.37
+CoQA&Doc2Dial	35.34	4.09	20.63
DPR(adv_nq)	34.05	3.57	19.76
+DPR(+hard neg)	35.09	3.83	20.87

Table 3: **Module Specific Methods Results** We evaluate models with F1, Exact Match, and sacreBLEU scores. **+CoQA&Doc2Dial** reports results for BART-large pretrained on CoQA and Doc2Dial dataset. **DPR(adv_nq)** is the RAG model composed of our own fine-tuned DPR using shared task configuration. **+DPR(+hard_negative)** corresponds to results for RAG with our fine-tuned DPR version with an extra hard negative sample.

4.2.2 Module Specific Methods

This section mainly discusses results for module-specific training methods. We fine-tune RAG's retriever and pretrain generator, DPR and BART, with contrastive learning and conversational QA datasets. We set the baseline model as the same configuration with section 4.2.1.

Pretraining We pretrain BART-large on CoQA and Doc2Dial before integrating it into RAG. We train 10 epochs for each dataset using hyperparameters suggested by the DialDoc2021 baseline code on subtask2.[10] Table 3 shows the result for pretraining. We report two results; pretrained on CoQA only and pretrained on both CoQA and Doc2Dial. Both datasets enhanced the model performance in terms of F1 and EM scores. There is extra room for improvement since we pretrain BART only for a few epochs due to long training time and limited resources.

Contrastive Learning We fine-tune DPR using the settings implemented by the shared task. We fine-tune the recently released version of DPR, `checkpoint.retriever.single-adv-hn.nq.bert-base-encoder`, for 50 epochs

[8] https://github.com/IBM/multidoc2dial

[10] https://github.com/doc2dial/sharedtask-dialdoc2021

on our new DPR dataset with one extra hard negative sample generated by antonym augmentation. Table 3 reports the results for contrastive learning. Despite using the same hyperparameters for DPR, there is degradation in the score for fine-tuning on our setting. However, after adding another hard negative sample, the model shows better performance on the shared task.

4.2.3 Leaderboard Submission

Our final model for DialDoc shared task at ACL 2022 utilizes all four suggested methods in this paper. We only participate in MultiDoc2Dial-seen-domain task which training data and test data share the same domains for the grounding documents. Our best performing model achieves 37.78 F1 score, 22.94 SacreBLEU, 36.97 Meteor, 35.46 RougeL, a total of 133.15 on the officially released test set (MDD-SEEN).

5 Conclusion

In this paper, we explain our submissions to the MultiDoc2Dial shared task. We utilize various conversational QA datasets and methods to improve the given baseline model. Our RAG model is composed of DPR for the retriever and BART for the generator. We train DPR with contrastive learning with an extra hard negative sample. BART is pretrained on conversational QA datasets, CoQA and Doc2Dial. On the end-to-end level, we implement multi-task learning to utilize model knowledge obtained from the previous grounding task that is trained on augmented data. All of the mentioned techniques enhance the model performance compared to the suggested baseline model.

Acknowledgements

K. Jung is with ASRI, Seoul National University, Korea. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. 2021R1A2C2008855)

References

Pawel Budzianowski and Ivan Vulic. 2019. Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. *CoRR*, abs/1907.05774.

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*.

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

Song Feng, Hui Wan, R. Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A. Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. *CoRR*, abs/2011.06623.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Patrick S. H. 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. 2020a. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Patrick S. H. Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020b. Question and answer test-train overlap in open-domain question answering datasets. *CoRR*, abs/2008.02637.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2018. Coqa: A conversational question answering challenge. *CoRR*, abs/1808.07042.

Yi-Lin Tuan, Yun-Nung Chen, and Hung-yi Lee. 2019. Dykgchat: Benchmarking dialogue generation grounding on dynamic knowledge graphs. *arXiv preprint arXiv:1910.00610*.

Qingyang Wu, Yichi Zhang, Yu Li, and Zhou Yu. 2019. Alternating recurrent dialog model with large-scale pre-trained language models. *CoRR*, abs/1910.03756.

Jinfeng Xiao, Lidan Wang, Franck Dernoncourt, Trung Bui, Tong Sun, and Jiawei Han. 2020. Open-domain question answering with pre-constructed question spaces. *CoRR*, abs/2006.08337.

Haolan Zhan, Lei Shen, Hongshen Chen, and Hainan Zhang. 2021. CoLV: A collaborative latent variable model for knowledge-grounded dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2250–2261, Online and Punta Cana, Dominican Republic. 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. *CoRR*, abs/2010.08824.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018a. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018b. A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.