CAiRE in DialDoc21: Data Augmentation for Information-Seeking Dialogue System

Yan Xu[*], **Etsuko Ishii**[*], **Genta Indra Winata, Zhaojiang Lin,**
Andrea Madotto, Zihan Liu, Peng Xu and Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{yxucb,eishii,giwinata}@connect.ust.hk, pascale@ece.ust.hk

Abstract

Information-seeking dialogue systems, including knowledge identification and response generation, aim to respond to users with fluent, coherent, and informative responses based on users' needs, which. To tackle this challenge, we utilize data augmentation methods and several training techniques with the pre-trained language models to learn a general pattern of the task and thus achieve promising performance. In DialDoc21 competition, our system achieved 74.95 F1 score and 60.74 Exact Match score in subtask 1, and 37.72 Sacre-BLEU score in subtask 2. Empirical analysis is provided to explain the effectiveness of our approaches.

1 Introduction

Recent progress in research has opened up real-life applications of dialogue systems (Winata et al., 2021; Ishii et al., 2021), of which information-seeking dialogue systems are one of the major types. The goal of such dialogue systems is to provide fluent and coherent responses with sufficient information to users based on their needs, retrieving information using the dialogue history. The performance of an information-seeking dialogue system can be evaluated from three aspects: (1) user utterance understanding, (2) relevant knowledge retrieval, and (3) agent response generation (Feng et al., 2020).

This paper presents work on the DialDoc-21 Shared Task, which is to teach a dialogue system to identify the most relevant knowledge in the associated document for generating agent responses in natural language. It is composed of two subtasks: Knowledge Identification (KI) to retrieve the knowledge from the document, and Response Generation (RG) to generate an agent utterance utilizing the retrieved knowledge.

To tackle this problem, we leverage the pre-trained language models from Liu et al. (2019a) and Lewis et al. (2020) and explore data augmentation methods with several training techniques so as to avoid over-fitting to the DialDoc datasets and to teach the model the general pattern of the task. Ensemble and post-processing are conducted to further improve the model performance. Experimental results show that data augmentation is a simple but effective approach for knowledge identification in information-seeking dialogue systems (Madotto et al., 2020a), while bringing improvement to response generation at the same time. In the DialDoc-21 competition, our system achieved 74.95 of F1 score and 60.74 of Exact Match in subtask 1, and 37.72 SacreBLEU score (Post, 2018) in subtask 2[1].

2 Datasets

Doc2Dial dataset In this shared task, we mainly focus on the Doc2Dial dataset (Feng et al., 2020). Doc2Dial addresses the challenge of modeling different dialogue scenes with documents and providing free-form responses while allowing follow-up questions from the agent. The shared task evaluation is divided into a testdev phase and a test phase. The main difference between these is that in the test phase, out-of-domain (OOD) data samples are included by selecting documents from the domain which is unseen in the training process. The testdev phase only covers 30% of the data samples in the final test phase.

Besides Doc2Dial, several other datasets are leveraged for augmentation, as follows:

MRQA 2019 Shared Task dataset is a collection of multiple reading comprehension datasets for evaluating the generalization ability of QA models. Six datasets are assigned to the training split,

---

[*] These two authors contributed equally.

[1]The code is available at: https://github.com/HLTCHKUST/CAiRE_in_DialDoc21.

Model	Initialization	Training	
		Data	Method
$\mathrm{RoBERTa}_{\mathrm{mrqa}}$	$\mathrm{RoBERTa}_{\mathrm{large}}$	MRQA	PT
$\mathrm{RoBERTa}_{\mathrm{mrqa_s}}$	$\mathrm{RoBERTa}_{\mathrm{large}}$	MRQA_{small}	PT
$\mathrm{RoBERTa}_{\mathrm{cqa}}$	$\mathrm{RoBERTa}_{\mathrm{large}}$	Doc2Dial, CQA	FT
$\mathrm{RoBERTa}_{\mathrm{f(cqa)}}$	$\mathrm{RoBERTa}_{\mathrm{cqa}}$	Doc2Dial	FT
$\mathrm{RoBERTa}_{\mathrm{f(mrqa)}}$	$\mathrm{RoBERTa}_{\mathrm{mrqa}}$	Doc2Dial	FT
$\mathrm{RoBERTa}_{\mathrm{cqa(mrqa)}}$	$\mathrm{RoBERTa}_{\mathrm{mrqa}}$	Doc2Dial, CQA	FT
$\mathrm{RoBERTa}_{\mathrm{cqa(mrqa_s)}}$	$\mathrm{RoBERTa}_{\mathrm{mrqa_s}}$	Doc2Dial, CQA	FT
$\mathrm{RoBERTa}_{\mathrm{f(cqa(mrqa_s))}}$	$\mathrm{RoBERTa}_{\mathrm{cqa(mrqa_s)}}$	Doc2Dial	FT
$\mathrm{RoBERTa}_{\mathrm{all}}$	$\mathrm{RoBERTa}_{\mathrm{large}}$	Doc2Dial, CQA, and MRQA	FT

Table 1: The combinations of the experimental settings for the KI subtask. Two-stage training consists of two stages: pre-training (PT) and fine-tuning (FT).

which is not included in the evaluation. Among them, SearchQA (Dunn et al., 2017) and TriviaQA (Joshi et al., 2017) differ from the others by the data resource and have the least generalization ability compared to the other four datasets as reported in (Su et al., 2019). In this shared task, we consider two settings when leveraging the MRQA dataset: MRQA and MRQA$_{small}$ which excludes SearchQA and TriviaQA.

Conversational QA (CQA) datasets We also introduce three CQA datasets, CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), and DoQA (Campos et al., 2020), in the shared task because of their similar settings to the KI process.

Wizard-of-Wikipedia (WoW) is a commonly-used knowledge-grounded dialogue dataset (Dinan et al., 2018). It aims at providing content-full responses to user utterances based on Wikipedia documents.

3 Methodology

We utilize a series of data-augmentation approaches to enable the model to obtain better representations on both dialogue context and document context and learn a general pattern of the task with less domain bias. Namely, we have a two-stage training paradigm, the first step is pretraining (PT) to have a better model initialization, and the second step is fine-tuning (FT) to adapt to DialDoc task. For each step, we can apply the multi-task learning (MTL) strategy if we have multiple datasets by making the datasets format uniform and treat samples equally. As reported in Fisch et al. (2019), a model trained on multiple dataset under similar tasks, is supposed to provide a better initialization for further fine-tuning and is capable of generalizing to the data samples in other domains. Thus, we

expect a model trained with MTL in the first step to offer a better initialization and in the second step to reduce the domain bias and avoid overfitting.

3.1 Knowledge Identification

In the KI task, we conduct experiments on a large pre-trained model, RoBERTa-large (Liu et al., 2019a), which has shown its effectiveness on many QA datasets (Ju et al., 2019). The MRQA dataset and three CQA above datasets are leveraged for data augmentation. The combinations of the experimental settings are considered as follows:

We consider using CQA datasets to enrich the data source. **RoBERTa$_{\mathrm{cqa}}$** is fine-tuned on Doc2Dial and three CQA datasets using MTL method. **RoBERTa$_{\mathrm{f(cqa)}}$** leverages the pre-trained $\mathrm{RoBERTa}_{\mathrm{cqa}}$ model and is fine-tuned on Doc2Dial dataset for better performance.

We train the RoBERTa model on MRQA dazataset and MRQA$_{small}$ dataset described in § 2 using MTL respectively (denoted as $\mathrm{RoBERTa}_{\mathrm{mrqa}}$ and $\mathrm{RoBERTa}_{\mathrm{mrqa_s}}$). These models could be further fine-tuned while providing a better initialization (Fisch et al., 2019). **RoBERTa$_{\mathrm{f(mrqa)}}$** is to further fine-tune $\mathrm{RoBERTa}_{\mathrm{mrqa}}$ on Doc2Dial dataset. The corresponding settings are also applied to **RoBERTa$_{\mathrm{f(mrqa_s)}}$** model. While **RoBERTa$_{\mathrm{cqa(mrqa)}}$** is initialized with $\mathrm{RoBERTa}_{\mathrm{mrqa}}$ and fine-tuned on Doc2Dial and three CQA datasets using MTL. **RoBERTa$_{\mathrm{cqa(mrqa_s)}}$** follows the same setting as the former model, but use $\mathrm{RoBERTa}_{\mathrm{mrqa_s}}$ model for initialization instead. **RoBERTa$_{\mathrm{f(cqa(mrqa_s))}}$** is to further fine-tune $\mathrm{RoBERTa}_{\mathrm{cqa(mrqa_s)}}$ on Doc2Dial dataset.

RoBERTa$_{\mathrm{all}}$ is trained on Doc2Dial, MRQA dataset and CQA datasets using MTL method.

For better readability, we summarize the model settings in Table 1. We also explore more combinations of the experimental settings, such as other combinations of the datasets and other pre-trained language models. However, those fail to bring the improvements as much as those we mentioned above.

Post-processing We further conduct post-processing on the model predictions based on our observation that the ground truths of the data samples are annotated by document splits which are provided together with the dataset. We consider including the whole split of the document once the prediction covers λ percent of it, where λ is set

as 0.1. In addition, for better performance in the shared task, we also slightly extend the predictions when there is a "Yes" or "No" shown right in front of the predicted spans.

Ensemble To further boost the model performance, we build an ensemble of our existing models. We consider one prediction containing the start position and the end position of the document as a unit and conduct voting over all the predictions of each data sample. The most frequent one will be selected as the final prediction. We denote the ensemble result as $\text{RoBERTa}_{\text{ensemble}}$.

Knowledge Identification		Response Generation	
max input length	512	max input length	300
max answer length	50	max target length	200
batch size	120	batch size	60
document stride	128	beam size	4
learning rate	3e-5	learning rate	3e-5

Table 2: The hyper-parameter settings in the shared task.

3.2 Response Generation

To obtain natural and relevant responses, we take advantage of the evidence to the query identified from § 3.1 and focusing on paraphrasing the corresponding knowledge sentences based on the dialogue context. We leverage the large pre-trained model $\text{BART}_{\text{large}}$ (Lewis et al., 2020). The process of training and inference can be summarized as three steps:

Pre-training on WoW dataset. We first pre-train the BART model on the WoW dataset for better initialization because of its similarity with the RG task. In the training process, the gold grounded knowledge sentences are concatenated with the dialogue context and fed into the model as the inputs.

Fine-tuning on Doc2Dial dataset. In the Doc2Dial dataset, the labels of the gold document splits are also provided in the training and validation set. The model is further fine-tuned on the Doc2Dial dataset using the same components for the input sequences in the first step. The model could be evaluated under two scenarios: (1) **Gold mode** ($\text{BART}_{\text{gold}}$), leveraging the gold labels of the knowledge evidence in the dataset as the knowledge inputs; (2) **Prediction mode** ($\text{BART}_{\text{pred}}$), leveraging the prediction of the KI process as the inputs.

Model	# of ckpt	EM	F1
Testdev Phase			
$\text{RoBERTa}_{\text{large}}$ (baseline)	-	58.08	72.17
$\text{RoBERTa}_{\text{cqa}}$	2	59.09($\pm$1.01)	72.90($\pm$0.25)
$\text{RoBERTa}_{\text{f(cqa)}}$	2	58.08($\pm$1.01)	72.23($\pm$0.18)
$\text{RoBERTa}_{\text{f(mrqa)}}$	1	58.08	72.30
$\text{RoBERTa}_{\text{f(mrqa}_s\text{)}}$	16	59.37($\pm$1.89)	73.51($\pm$1.60)
$\text{RoBERTa}_{\text{cqa(mrqa)}}$	1	58.08	72.59
$\text{RoBERTa}_{\text{cqa(mrqa}_s\text{)}}$	6	59.60($\pm$1.35)	73.76($\pm$1.57)
$\text{RoBERTa}_{\text{f(cqa(mrqa}_s\text{))}}$	1	60.10	75.02
$\text{RoBERTa}_{\text{all}}$	1	58.08	74.63
$\text{RoBERTa}_{\text{ensemble}}$	-	63.13	77.31
$\text{RoBERTa}_{\text{all}}$-postproc	-	57.07(-1.01)	74.15(-0.47)
$\text{RoBERTa}_{\text{ensemble}}$-postproc	-	63.13(-0.00)	76.73(-0.58)
Test Phase			
$\text{RoBERTa}^*_{\text{ensemble}}$	-	60.74	74.95

Table 3: The results of the selected models on the testdev and test phase of subtask 1 are listed. All the results are calculated with the corresponding predictions after post-processing except those with specific notations. For the models that are trained with multiple random seeds, the average scores and the standard deviations are presented. $\text{RoBERTa}^*_{\text{ensemble}}$ denotes the results of the ensemble model on the test set.

Inference with Knowledge Evidence. During the testdev and test phase, we leverage the predictions from the KI process as the knowledge evidence components for the dialogue queries. The model generates responses based on a concatenation of the knowledge evidence and the dialogue context.

Post-processing To avoid serious information loss in the generations compared to the knowledge evidence for the OOD data samples, we compare the lengths of the knowledge evidence and the responses (denoted as L_{kn} and L_{resp}). The generated response will be replaced by the raw knowledge evidence as the final output if $L_{\text{resp}} \leq \alpha L_{\text{kn}}$, where α is set as 0.4.

4 Experiments

4.1 Training Details

Hyper-parameter Settings We apply different settings to utilize the dialogue history for the two subtasks. For subtask 1, we leverage all previous turns and build the input sequence in a reverse order to them. For subtask 2, we leverage one extra last turn in the time order and differentiate the speakers with special tokens. In Table 2, we list the selected hyper-parameters utilized in the shared task.

Model	SacreBLEU		
	val	testdev	test
BART$_{large}$ (baseline)	-	16.73	-
Gold	45.67	-	-
RoBERTa$_{ensemble}$	38.78	37.45	38.68
BART$_{gold}$	20.17	-	-
+WoW pre-traning	48.24	-	-
BART$_{pred}$	16.67	16.72	16.45
+WoW pre-traning	39.87	38.26	37.31
+WoW pre-training+postproc*	-	-	37.72

Table 4: The results of selected models on subtask 2 are listed. Gold denotes the gold knowledge evidence labels provided in the dataset. The model denoted with * is the final submission to the test phase.

Ensemble Settings In subtask 1, we make an ensemble of all the checkpoints of the models listed in Table 1 except RoBERTa$_{mrqa}$ and RoBERTa$_{mrqa_s}$. The details of the checkpoints can be found in Tabel 3.

Metrics and Model Selection In subtask 1, the Exact Match (EM) and uni-gram F1 score are utilized as the criteria, while in subtask 2, we evaluate the generation by SacreBLEU. We select the models with the best EM and SacreBLEU scores on the validation set respectively, for the two subtasks. Specifically for subtask 2, the model is selected under the gold mode.

4.2 Results and Discussion

4.2.1 Results

The results are shown in Table 3 and Table 4. For both subtasks, we observe gaps between the test-dev phase and the test phase. For some of the models in subtask 1, multiple random seeds are applied in the training process. The performance gap may result from the domain difference of the partial data samples in the test phase, where the corresponding documents are unseen in the training set. In Table 3, without post-processing on the predictions, the model performance consistently drops to a certain extent, which indicates that post-processing is suitable for the Doc2Dial scenario. Ensemble, which is a common strategy to improve performance, shows its effectiveness in this task.

For subtask 2, the pre-training on WoW dataset brings huge improvement to the model. Interestingly, by just using the knowledge evidence predicted from the subtask 1 RoBERTa$_{ensemble}$ model or the gold knowledge evidence labels, the perfor-

mance can even exceed that of the generative model on SacreBLEU scores, while the responses from BART$_{pred}$ are more fluent and natural. This may be caused by the information loss when paraphrasing the knowledge evidence to dialogue responses.

4.2.2 Discussion

In this task, we explore data augmentation methods and conduct two-stage training as auxiliary training strategy for improvement. Although resource- and time-consuming, this approach is easy to implement and effective at enabling the model to learn more general ability on the task.

4.2.3 Post-Challenge Improvements

From our findings, the hyper-parameter, the maximum answer length, is left untuned, which hurts the QA model performance to some degree. With a maximum answer length of 100, the EM and F1 score on the testdev set improve by 2.53 and 1.08, respectively, while a 64.42 EM and 77.27 F1 score are achieved on the test set. With the improved prediction from subtask 1, we achieve a 39.88 SacreBLEU score in subtask 2.

5 Related Work

Conversational QA is a type of reading comprehension task that requires understanding not only the question but also the previous conversation turns. Various datasets have been introduced in recent years, and many of them restrict answers to be extraction of a span from the reference document, while the others allow free-form responses (Choi et al., 2018; Reddy et al., 2019; Campos et al., 2020).

In addition to the works to enrich the contents of open-domain conversations by controllable generation (Lin et al., 2020; Madotto et al., 2020b), the knowledge grounded dialogue task aims to offer more informative conversation by leveraging an external knowledge source (Dinan et al., 2018; Xu et al., 2020). Relevant knowledge selection is the key to improving the whole system, and very recently, latent variable models have been attracting more attention for this purpose (Lian et al., 2019; Liu et al., 2019b; Kim et al., 2020; Chen et al., 2020; Xu et al., 2021).

6 Conclusion

In this paper, we utilize data augmentation methods and several training techniques with pre-trained language models to tackle the challenge of the

information-seeking dialogue task. The results have indicated the effectiveness of our approaches. Moreover, data augmentation methods are easy to implement, which is promising for practical use.

References

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Milan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa: accessing domain-specific faqs via conversational qa. In *Proceedings of the ACL*, pages 7302–7314.

Xiuyi Chen, Fandong Meng, Peng Li, Feilong Chen, Shuang Xu, Bo Xu, and Jie Zhou. 2020. Bridging the gap between prior and posterior knowledge selection for knowledge-grounded dialogue generation. In *Proceedings of the EMNLP*, pages 3426–3437.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the EMNLP*, pages 2174–2184.

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

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Song Feng, Kshitij Fadnis, Q Vera Liao, and Luis A Lastras. 2020. Doc2dial: a framework for dialogue composition grounded in documents. In *Proceedings of the AAAI*, volume 34, pages 13604–13605.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13.

Etsuko Ishii, Genta Indra Winata, Samuel Cahyawijaya, Divesh Lala, Tatsuya Kawahara, and Pascale Fung. 2021. Erica: An empathetic android companion for covid-19 quarantine.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the ACL*, pages 1601–1611.

Ying Ju, Fubang Zhao, Shijie Chen, Bowen Zheng, Xuefeng Yang, and Yunfeng Liu. 2019. Technical report on conversational question answering. *arXiv preprint arXiv:1909.10772*.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the ACL*, pages 7871–7880.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *IJCAI International Joint Conference on Artificial Intelligence*, page 5081.

Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2020. Xpersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zihan Liu, Jamin Shin, Yan Xu, Genta Indra Winata, Peng Xu, Andrea Madotto, and Pascale Fung. 2019b. Zero-shot cross-lingual dialogue systems with transferable latent variables. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1297–1303.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020a. Learning knowledge bases with parameters for task-oriented dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2372–2394.

Andrea Madotto, Etsuko Ishii, Zhaojiang Lin, Sumanth Dathathri, and Pascale Fung. 2020b. Plug-and-play conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2422–2433.

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.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Dan Su, Yan Xu, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu, and Pascale Fung. 2019.

Generalizing question answering system with pre-trained language model fine-tuning. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 203–211.

Genta Indra Winata, Holy Lovenia, Etsuko Ishii, Farhad Bin Siddique, Yongsheng Yang, and Pascale Fung. 2021. Nora: The well-being coach. *arXiv preprint arXiv:2106.00410*.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Animashree Anandkumar, and Bryan Catanzaro. 2020. Controllable story generation with external knowledge using large-scale language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845.

Yan Xu, Etsuko Ishii, Zihan Liu, Genta Indra Winata, Dan Su, Andrea Madotto, and Pascale Fung. 2021. Retrieval-free knowledge-grounded dialogue response generation with adapters. *arXiv preprint arXiv:2105.06232*.