DialDoc 2022 Shared Task: Open-Book Document-grounded Dialogue Modeling

Song Feng*
Amazon AWS AI
sofeng@amazon.com

Siva Sankalp Patel
IBM Research
siva.sankalp.patel@ibm.com

Hui Wan
IBM Research
hwan@us.ibm.com

Abstract

The paper presents the results of the Shared Task hosted by the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering co-located at ACL 2022. The primary goal of this Shared Task is to build goal-oriented information-seeking conversation systems that are grounded in the domain documents, where each dialogue could correspond to multiple sub-goals that are based on different documents. The task is to generate agent responses in natural language given the dialogue and document contexts. There are two task settings and leaderboards based on (1) the same sets of domains (*SEEN*) and (2) one unseen domain (*UNSEEN*). There are over 20 teams participating in Dev Phase and 8 teams participating in both Dev and Test Phases. There are multiple submissions that significantly outperform the baseline. The best-performing system achieves 52.06 F1 and the total of 191.30 on the *SEEN* task; and 34.65 F1 and the total of 130.79 on the *UNSEEN* task.

1 Introduction

Goal-oriented document-grounded dialogue systems enable end users to interactively query about domain-specific information based on the given documents. The tasks of querying document knowledge via conversational systems continue to attract a lot of attention from both research and industrial communities for various applications such as ORConvQA (Qu et al., 2020), MultiDoc2Dial (Feng et al., 2021), QReCC (Anantha et al., 2021), TopiOCQA (Adlakha et al., 2022) and Abg-CoQA (Guo et al., 2021). The previous Shared Task (Feng, 2021) by the First DialDoc Workshop addressed the task of goal-oriented information-seeking dialogue systems in the machine reading comprehension setting, where the dialogue is aiming at querying about the information provided in a given document (Feng et al., 2020). However, in real-life scenarios, for conversation in a given domain, the grounding document is often unknown, a dialogue turn could arbitrarily correspond to any document, hence each dialogue could be grounded in multiple documents. Thus, we propose to explore the open-book closed-domain setting for goal-oriented information-seeking dialogue systems that are grounded in the given domain documents.

We introduce the Shared Task at the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2022 Shared Task). The Shared Task aims to deal with the information-seeking goal-oriented dialogues that have multiple sub-goals corresponding to different documents. The input includes the dialogue history, the current user turn, and a set of domain documents, the output is the agent's utterance in natural language. It comprises two tasks that address two different evaluation settings: (1) the *SEEN* task where the test data shares the same sets of domains as the training data; and (2) the *UNSEEN* task where the test data is all in one unseen domain different from the training data. We host the leaderboards for Dev and Test Phases for the *SEEN* and *UNSEEN* tasks respectively on eval.ai[1].

There are over 20 teams participating in Dev Phase and 8 teams participating in both Dev and Test Phases. Multiple submissions significantly outperform the baseline. The best-performing system achieves 52.06 F1 and the total of 191.30 on the *SEEN* task comparing to 35.85 and 126.21 by the baseline; and 34.65 F1 and the total of 130.79 on the *UNSEEN* task comparing to 19.26 and 59.52 by the baseline.

In this report, we first describe the dataset and the two task settings. Then, we summarize the approaches and evaluation results of several top participating teams.

* Work done while at IBM Research

[1] https://eval.ai/

Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 155 - 160
May 26, 2022 ©2022 Association for Computational Linguistics

domain	#doc	#dial	two-seg	>two-seg	single
ssa	109	1191	701	188	302
va	138	1337	648	491	198
dmv	149	1328	781	257	290
student	92	940	508	274	158
total	488	4796	2638	1210	948

Table 1: MultiDoc2Dial data statistics (Feng et al., 2021)

2 Dataset

In this Shared Task, the dataset is based on MultiDoc2Dial introduced by (Feng et al., 2021). It contains 4796 conversations with an average of 14 turns grounded in 488 documents from four domains including `va.org` and `studentaid.org`. For document data, each document includes a title, the body content with the span/section information as well as the HTML mark-ups such as `list` and `title`. For dialogue data, each turn in a dialogue contains: (1) the speaker role, (2) the dialogue act, (3) the grounding text span along with the title of the document, and (4) human generated utterance in natural language. Each dialogue contains one or multiple segments where each indicates that all turns within one segment are grounded in the same document. Table 1 shows the statistics of the dataset by domain, including the number of dialogues with two segments (two-seg), more than two segments (>two-seg), and no segmentations (single).

For model development, we provide the original split of training and validation data. For the leaderboard setup, we use a small portion (30%) of the test split based on the number of dialogues for Dev Phase and entire test split for the final Test Phase. For the *UNSEEN* task setting, the final test set includes the dialogue and document data all from an unseen domain *cdccovid* that is not in the original MultiDoc2Dial dataset. The dialogues from the unseen domain were collected in the same data collection process as MultiDoc2Dial dataset.

3 Task Description

Our Shares Task centers on building open-book goal-oriented dialogue systems, where an agent could provide an answer or ask follow-up questions for clarification or verification. The main goal is to generate grounded agent responses in natural language based on the dialogue context and domain knowledge in the documents. The provided training data is mainly based on MultiDoc2Dial dataset but the participants could utilize any public dataset without any additional human annotations on the MultiDoc2Dial dataset. It includes two task settings depending on whether the cases are from unseen domains (*SEEN* task) or one unseen domain (*UNSEEN* task) from training data. Here we only consider the cases where user queries are answerable. For test split, there is only one turn to predict per dialogue. Table 2 presents the number of dialogues ('dials') as well as the total turns for prediction ('predicts') in each data split, where the last column contains the numbers of examples for Test Phase evaluation for *SEEN* and *UNSEEN*, respectively.

#	train	val	t-*SEEN*/*UNSEEN*
dials	3474	661	661 / —
predicts	21453	4201	661 / 126

Table 2: Statistics of dialogue data in train, dev and test splits for *SEEN* and *UNSEEN* task settings.

4 Evaluation

The evaluation is focused on the groundedness and naturalness of the generated agent response. We consider the automatic metrics as intrinsic evaluation metrics, and human annotations for extrinsic evaluations.

4.1 Intrinsic Evaluation

We use the following metrics: F1 (Rajpurkar et al., 2016), SacreBLEU (Post, 2018), METEOR (Banerjee and Lavie, 2005) and RougeL (Lin, 2004). The rankings on the leaderboards are based on the sum of all four scores. For each leaderboard, we select the three top-ranked teams for further human evaluation.

4.2 Extrinsic Evaluation

We ask human annotators to rank three generated utterances, each from a different team, based on the relevance and fluency given the dialogue history and the grounding document passages as reference. *relevance* is used to measure how well the generated utterance is relevant to the grounding span as a response to the previous dialogue turn(s). *fluency* indicates whether the generated utterance is grammatically correct and generally fluent in English.

Rank	Participant Team	F1	SacreBLEU	METEOR	RougeL	Total
1	CPII-NLP	**52.06**	**37.41**	**51.64**	**50.19**	**191.30**
2	zsw_dyy_lgz	48.56	33.27	48.73	46.75	177.31
3	UGent-T2K	46.90	32.23	47.96	44.89	171.98
4	CMU_QA	46.22	31.82	46.02	44.19	168.24
5	JLP	37.78	22.94	36.97	35.46	133.15
6	Docalog	36.07	23.70	35.67	34.44	129.87
7	LingJing	36.69	22.78	35.46	34.52	129.44
-	Baseline	35.85	22.26	34.28	33.82	126.21

Table 3: The participating teams and the scores for Test Phase of *SEEN* leaderboard.

Rank	Participant Team	F1	SacreBLEU	METEOR	RougeL	Total
1	CPII-NLP	**34.65**	**27.57**	**34.08**	**34.49**	**130.79**
2	CMU_QA	33.01	25.04	32.92	31.95	122.91
3	UGent-T2K	33.36	21.20	33.57	31.47	119.60
4	zsw_dyy_lgz	32.78	21.32	32.74	31.44	118.28
5	Docalog	28.44	20.52	27.54	26.57	103.07
-	Baseline	19.26	6.32	16.77	17.16	59.52

Table 4: The participating teams and the scores for Test Phase of *UNSEEN* leaderboard.

For the *SEEN* task setting, we randomly select 100 generated turns where the normalized utterances are not all the same; for *UNSEEN*, we randomly select 80. We have three experts as annotators, with 10% overlap for the annotations.

5 Shared Task Submissions

We hosted the leaderboards[2] for Dev and Test Phases for the two task settings *SEEN* and *UNSEEN* on `eval.ai`. The Dev Phase lasted for three and a half months and the Test Phase lasted for a week. There are over 500 submissions by over 20 teams that participated in Dev Phase. For the final Test Phase, 8 teams submitted to the *SEEN* leaderboard, and 6 teams submitted to the *UNSEEN* leaderboard. Next, we summarize the approaches adopted by the top teams who submitted their technical papers.

The baseline approach for the Shared Task is based on RAG (Lewis et al., 2020b), where the DPR (Karpukhin et al., 2020) passage retriever is fine-tuned on MultiDoc2Dial dataset, as described in (Feng et al., 2021). Several teams significantly improved the results over the baseline as shown in Table 3 and 4. Team CPII-NLP achieved the highest scores on both *SEEN* and *UNSEEN* leaderboard.

[2] `https://eval.ai/web/challenges/challenge-page/1437/overview`

5.1 CPII-NLP

The team presents a pipeline system of retriever, re-ranker, and generator. The retriever adopts DPR (Karpukhin et al., 2020). The re-ranker is an ensemble of three cross-encoder models using BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020), respectively. The generator leverages the pre-trained sequence-to-sequence model BART$_{\texttt{large}}$ (Lewis et al., 2020a) jointly trained with a grounding span predictor. The three components are individually optimized, while passage dropout and regularization techniques are adopted to improve the response generation performance. CPII-NLP ranked 1st on both *SEEN* and *UNSEEN* leaderboards on F1, SacreBLEU, METEOR and RougeL scores.

5.2 zsw_dyy_lgz

The team presents their system named Grounding-Guided Goal-oriented dialogues Generation(G4), a three-stage approach composed of a retriever adopting ANCE(Xiong et al., 2021), a reader predicting grounding spans restricted to whole phrases, and a generator adopting FiD (Izacard and Grave, 2021) which leverages explicitly markings of grounding spans together with the original passages. Experiment results show that this approach effectively generates responses better grounded to text spans and closer to correct responses. To alleviate the issue of the reader accuracy being lower at inference than during training, they also present a data augmentation approach as regularization to account for more diverse groundings and improve the robustness.

5.3 CMU_QA

The team also follows the retriever-reader architecture and presents their system called Refined Retriever-Reader (R3). R3 includes several improvements over the baseline approach, including adopting a sparse retriever based on DistilSplade (Formal et al., 2021) instead of dense retriever, adding a RoBERTa-based cross-encoder passage reranker, using FiD (Izacard and Grave, 2021) as the generator, and a curriculum learning training paradigm. The experiment results show significant improvement over the baseline performance.

5.4 UGent-T2K

The team presents a cascade pipeline dialogue system for the task. The system consists of three modules: a document retriever, a passage retriever, and a response generator. The system uses DPR for the passage retrieval and FiD (Izacard and Grave, 2021) for the response generation. Then they use LambdaMART (Burges, 2010) for reranking. The experiment results show that document ranking could be helpful for passage retrieval and the multi-passage-fusing generator outperforms the RAG model.

5.5 Docalog

The team presents a three-stage pipeline consisting of (a) Document Retriever with Title Embedding and IDF on Texts (DR.TEIT); (2) a grounding span predictor; (3) an ultimate span picker. Their experiment results indicate that incorporating contextualized embedding information along with semantic similarity on the character level between the answer and question history can further improve the prediction of the ultimate answer.

5.6 JLP

The team explores various strategies for the dialogue task, including multi-task learning, tuning the generator BART (Lewis et al., 2020a) on additional QA datasets, data augmentation via synonym augmenter [3], and contrastive learning based on extra-hard negative examples. The experiment results indicate that all techniques help further improve the performance comparing to the baseline approach.

[3]https://github.com/makcedward/nlpaug

Team	Affiliation
CMU_QA	Carnegie Mellon University
CPII-NLP	The Chinese University of Hong Kong (CUHK) & Centre for Perceptual and Interactive Intelligence (CPII) Limited
Docalog	Sharif University of Technology & Volkswagen AG
JLP	Seoul National University
UGent-T2K	Ghent University
zsw_dyy_lgz	Tencent Cloud Xiaowei & Beihang University & Tianjin University

Table 5: Teams and their affiliations.

5.7 LingJing

The team presents a framework that most different than the baseline among the teams. It proposes to enhance downstream evidence retrieval by generating evidence into model parameters through pre-training. More specifically, it uses Pegasus (Zhang et al., 2020) to store document knowledge into a language model and then Child-Tuning (Xu et al., 2021) approach for evidence generation. The results are marginally better the baseline performance.

6 Conclusion

We present the results of DialDoc 2022 Shared Task. World-wide researchers and practitioners brought their individual perspectives on the task through this data competition. We received over 500 submissions during the Dev Phase by over 20 teams for both *SEEN* and *UNSEEN* leaderboards. For the final Test Phase, there were officially 8 teams submitted to the *SEEN* leaderboard and 6 teams submitted to the *UNSEEN* leaderboard. Most of the submissions during Test Phase beat the baseline performance by large margins.

Acknowledgements

We would like to thank IBM Research AI for sponsoring the Shared Task and competition prizes. We thank `eval.ai` for their help and support on hosting the leaderboards on their platform.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

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, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. 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.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coqa: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. 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.

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. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

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.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. *Open-Retrieval Conversational Question Answering*, page 539–548. Association for Computing Machinery, New York, NY, USA.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9514–9528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.