Docalog: Multi-document Dialogue System using Transformer-based Span Retrieval

Sayed Hesam Alavian[†,*], **Ali Satvaty** [†*], **Sadra Sabouri** [◇*], **Ehsaneddin Asgari**[+§] **and Hossein Sameti** [†§]
[†] AI Group, Computer Engineering Department, Sharif University of Technology, Tehran, Iran
[◇] Electrical Engineering Department, Sharif University of Technology, Tehran, Iran
+ NLP Expert Center, Data:Lab, Volkswagen AG, Munich, Germany
{alavian, stvty}@ce.sharif.edu, sadra@ee.sharif.edu, [§]**asgari@berkeley.edu**, [§]**sameti@sharif.edu**

Abstract

Information-seeking dialogue systems, including knowledge identification and response generation, aim to respond to users with fluent, coherent, and informative answers based on users' needs. This paper discusses our proposed approach, *Docalog*, for the DialDoc-22 (MultiDoc2Dial) shared task. *Docalog* identifies the most relevant knowledge in the associated document, in a multi-document setting. *Docalog*, is a three-stage pipeline consisting of *(1) a document retriever model (DR. TEIT)*, *(2) an answer span prediction model*, and *(3) an ultimate span picker* deciding on the most likely answer span, out of all predicted spans. In the test phase of MultiDoc2Dial 2022, *Docalog* achieved f1-scores of 36.07% and 28.44% and SacreBLEU scores of 23.70% and 20.52%, respectively on the *MDD-SEEN* and *MDD-UNSEEN* folds.

1 Introduction

Introducing a machine-generated dialogue with a human level of intelligence has been consistently among dreams of artificial intelligence with a vast number of applications in different domains, ranging from entertainment (Baena-Perez et al., 2020) to healthcare systems (Montenegro et al., 2019; Bharti et al., 2020). In such a system, the machine has to (i) understand the flow of conversation, (ii) raise informative questions, and (iii) answer problems in different domains of interest, and in some cases it has to act as an all-knowing agent (Dazeley et al., 2021). Recent advances in NLP have made this dream closer to reality. In the last decade, the success of the neural language model in language understanding and generation has encouraged more and more contributions from both academia and industry in the area of conversational artificial intelligence (Fu et al., 2020).

The major efforts in conversational artificial intelligence can be categorized into three subareas (Zaib et al., 2021): **(i) chat-oriented systems**, where the aim is to engage the users through a natural and fluent conversation (Nio et al., 2014), the examples are Alexa[1], Siri[2], or Cortana[3]; **(ii) task-oriented systems**, which are designed for a particular action, such as reserving a restaurant or planning an event by understanding the conversation (Yan et al., 2017); and **(iii) QA dialog systems** attempting to answer the user exploiting information deducted from a collection of seen documents or a knowledge base, for instance CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018). Our work in this paper also falls in the third category.

In this system paper, we present our work on the DialDoc Shared Task 2022 centered on developing a QA dialogue system. A common approach to this problem comprises two subtasks of **(i) knowledge identification (KI)** to retrieve the knowledge from the documents and **(ii) response generation (RG)** to generate an answer based on the retrieved knowledge (Feng et al., 2020b; Kim et al., 2021). The multi-document scenario, meaning that the related documents have to be retrieved before the answer generation, is the main distinction between the DialDoc Shared Tasks in 2021 and 2022. To tackle this problem, we propose a three-stage pipeline, called *Docalog*, consisting of *(1) document retriever model (DR. TEIT)*, *(2) an answer span prediction model*, a state-of-the-art transformer-based model taking single documents (DR. TEIT results) as input and outputting the answer span for every input document, and *(3) an ultimate span picker* deciding on the most likely answer span, out of all predicted spans in the

[*] Equal contribution
[§] Corresponding authors

[1] https://developer.amazon.com/en-US/alexa
[2] https://www.apple.com/uk/siri/
[3] https://www.microsoft.com/en-us/cortana

Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 142 - 147

step (2). In Multidoc2dial 2022 challenge, during the test phase, *DocAlog* achieved an f1-score of 36.07% and a SacreBLEU of 23.70% on the *MDD-SEEN*, and an f1-score of 28.44% and a SacreBLEU of 20.52% on the *MDD-UNSEEN*.

2 Related Work

The main focus of DialDoc shared tasks has been on developing task-oriented information-seeking dialogue systems, an important setting in the domain of conversational AI (Feng et al., 2021). Some of the performing models in this domain have been CAiRE (Xu et al., 2021), SCIRDT (Li et al., 2021), and RWTH (Daheim et al., 2021). The proposed approaches of CAiRE and SCIRDT utilize additional data for the augmentation of pre-trained language models in span detection, and RWTH (Daheim et al., 2021) model uses neural retrievers for obtaining the most relevant document passages.

In a broader context, the major work in document-grounded dialogue modeling can be divided into the following categories: (i) QA in an unstructured content, e.g., CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), ShARC (Saeidi et al., 2018), DoQA (Campos et al., 2020), and Doc2Dial (Feng et al., 2020b) (ii) QA in a semi-structured content, such as tables or lists, e.g., SQA (Iyyer et al., 2017), and HybridQA (Chen et al., 2020) and thirdly (iii) QA in a multimedia content (images and videos with associated textual descriptions), e.g., RecipeQA (Yagcioglu et al., 2018), PsTuts-VQA (Colas et al., 2020), and MIMOQA (Singh et al., 2021).

3 Materials and Models

3.1 MultiDoc2Dial Shared Task Dataset

Training material used in this shared task is derived from the MultiDoc2Dial, a new dataset constructed based on Doc2Dial dataset V1.0.1 (Feng et al., 2020b). It contains a collection of documents and conversations exchanged between the user(s) and an agent grounded in the associated documents.

3.2 Model

The three-stage workflow of *Docalog* is depicted in Figure 1. Firstly, *DR. TEIT* predicts the N best documents based on the user input (q_t), and a query history of the respective user ($q_{1:(t-1)}$). Afterward, the span prediction model finds matching spans for a given query for each of the N best documents in the step before. Eventually, the ultimate span picker selects the most related span among predicted spans using a combination of the cosine similarity between the query and the span embeddings, as well as char-level *TF-IDF*-based cosine similarity between the query and the span vectors.

3.2.1 Document Retriever

In our retrieval model to encode the texts, we use a pre-trained language-agnostic BERT sentence embedding (LaBSE) (Feng et al., 2020a). One of our contributions here is to include the dialogue history in our document retriever model. We also found that the title tokens and their synonyms are extremely useful in document-changing dialogues, i.e., questions changing the context document during the conversation.

Our document retriever model, Document Retriever with Title Embedding and IDF on Texts (DR. TEIT), uses two scoring measures and aggregates them through a hyper-parameter in a convex combination (Eq. 1).

$$\lambda S_{TE} + (1 - \lambda) S_{TI}, \tag{1}$$

where S_{TE} is the title embedding based on the similarity between the sequence of query and the history ($q_{1..t}$) and the document titles. S_{TI} is a character n-gram ($2 \leq n \leq 8$) similarity score calculated between the aggregation of the query and the history ($q_{1..t}$) and the document texts using TF-IDF-based cosine similarity (Figure 1-c).

3.2.2 Span Predictor

Our span predictor is a RoBERTa language model (Zhuang et al., 2021) fine-tuned to predict the start and the end positions of the answer span, similar to CAiRE (Xu et al., 2021), one of the best performing models in DialDoc-2021. To model the history of questions, we append the last two history turns to the current question, as also proposed in (Ohsugi et al., 2019), and feed it to the model as part of the current question.

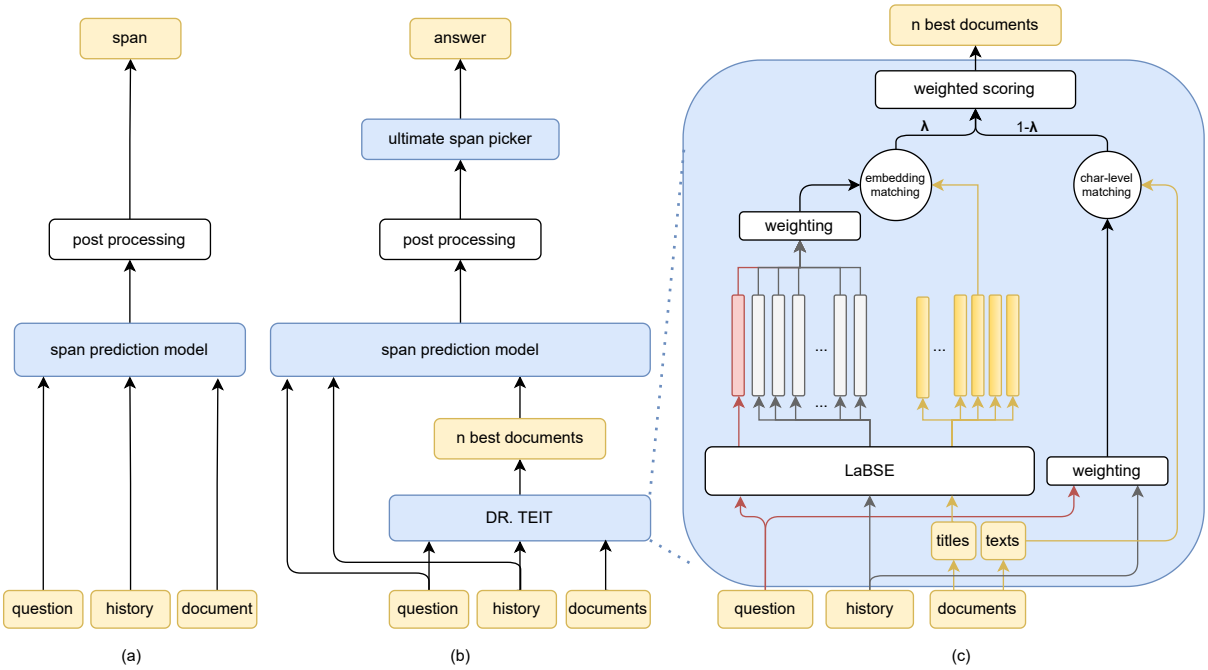

Figure 1: *Docalog* **model architecture and the overview diagram**: a) a standalone answer span prediction model. b) our three-stage model consists of (i) Dr. TEIT retriever model connected to the (ii) the span prediction model, and (iii) an aggregator which works as an ultimate span-picker deciding on the most likely span of the answer, out of all predicted spans. c) A detailed view of Dr. TEIT, the retriever architecture.

Prior to training our model on the DialDoc 2022 dataset, to gain more global knowledge in question answering, the span predictor of *Docalog* undergoes a pre-training phase on several CQA datasets such as CoQA (Reddy et al., 2019), QuAC (Choi et al., 2018), DoQA (Campos et al., 2020), and Doc2Dial (Feng et al., 2020b). Next, we fine-tune this model on the MultiDoc2Dial dataset using the grounding documents for each question. In this fine-tuning stage, we consider the task as a single-document question answering task. Therefore, at each training step, we only feed the model with the grounding document. The reason behind having a standalone span prediction model is to prevent the propagation of the retrieval error in the training phase.

3.2.3 Ultimate Span Picker

As discussed, the span detector provides the most-likely spans for each of the N best documents by the retriever. Since the answer-span probabilities are not comparable across documents, we need to rank the top-N identified spans searching for the ultimate answer. Therefore, similar to our document retriever, we use a convex combination between the embedding-based and character-level-based cosine similarities of the query and the detected spans through a hyper-parameter α that can be tuned on a validation set:

$$\alpha S_{SE} + (1 - \alpha) S_{SI}, \tag{2}$$

where S_{SE} is the span embedding similarity and S_{SI} is character-level *TF-IDF* similarity.

To summarize the workflow of *Docalog*, (1) a document retriever model using both embedding and character-level information retrieves the N most relevant documents to the current question. Based on the validation data we choose the hyper-parameter N in a way that we ensure selecting the answer document. (2) Using a trained span detector model, for each N document we detect the answer spans. (3) We use another document retriever model, this time to select the best-detected span, and the ultimate answer to the question is the post-processed version of this final span.

3.2.4 Experimental Settings

For the span prediction, we use a large RoBERTa language model [4] (Liu et al., 2019). During the training and the prediction phase, we feed the

[4] https://github.com/huggingface/transformers

Phase	Model	$F1_U$	SacreBLEU	METEOR	RougeL	$F1_G$	EM_G
MDD-SEEN (Dev)	baseline	36.23%	21.41%	34.16%	34.01%	44.90%	28.64%
	Docalog@1	**36.84%**	21.80%	**36.67%**	**34.44%**	**49.18%**	**36.18%**
	Docalog@2	34.99%	**23.30%**	33.81%	32.89%	46.62%	35.1%
	Docalog@3	35.19%	22.73 %	35.20%	33.56%	48.39%	35.67%
MDD-UNSEEN (Dev)	baseline	18.66%	5.99%	16.40%	16.95%	-	-
	Docalog@1	**26.12%**	**17.72%**	**25.52%**	**24.47%**	**33.36%**	**13.42%**
	Docalog@2	24.75%	15.07%	24.59%	22.76%	29.64%	9.59%
	Docalog@3	22.37%	14.21%	23.68%	21.02%	25.31%	7.75%
MDD-SEEN (Test)	baseline	35.85%	22.26%	34.28%	33.82%	-	-
	Docalog@1	**36.07%**	**23.70%**	**35.67%**	**34.44%**	**48.11%**	**34.19%**
	Docalog@2	33.41%	20.30%	33.52%	31.74%	44.11%	29.34%
	Docalog@3	29.90%	16.81%	30.25%	28.13%	39.33%	24.50%
MDD-UNSEEN (Test)	baseline	19.26%	6.32%	16.77%	17.16%	-	-
	Docalog@1	**28.44%**	**20.52%**	**27.54%**	**26.57%**	**35.41%**	**15.87%**
	Docalog@2	28.43%	20.51%	27.54%	26.57%	35.41%	15.87%
	Docalog@3	28.40%	20.51%	27.54%	26.57%	35.41%	15.87%

Table 1: *Docalog* **results on Multidoc2dial 2022 challenge.** Docalog@k indicates our method when working on the best k documents retrieved by the document retriever for the span detection and providing the final answer.

documents to the model with a stride size of 128 tokens. We pre-train our span-prediction model for 1 epoch on the CQA datasets and then fine-tuning was done on the MultiDoc2Dial dataset for 3 epochs. Our pre-training lasted around 13 hours and our fine-tuning step 15 hours, both of which were processed on a GeForce RTX 3070 GPU with 12GB memory.

Availablity: Our implementation of *Docalog* is available at github [5].

4 Results

Document Retriever: in our experiments, Dr. TEIT achieved a Precision@5 of 86% and a Mean Reciprocal Rank (MRR) of 0.72 indicating that on average, the hit is among the first two retrieved documents and it would be more than sufficient to take top-5 documents to the next step, i.e., span detection.

Docalog Results: In our final model, we combine DR. TEIT, as the retriever with our span predictor model. The comprehensive report of *Docalog* is provided in Table 1. We obtained the best F1 score of 36.07% with Docalog@1, suggesting that the ultimate span picker needs further improvements.

[5]https://github.com/Sharif-SLPL-NLP/Docalog-2022

5 Conclusions

We proposed Docalog, a solution for the DialDoc-22 challenge. *Docalog* is a three-stage pipeline consisting of *(1) a document retriever model (DR. TEIT)*, *(2) an answer span prediction model*, and *(3) an ultimate span picker* deciding on the most likely answer span, out of all predicted spans. Our experiments show that combining contextualized embedding information with character-level similarities between the answer and the question history can effectively help in the prediction of the ultimate answer. In the test phase of MultiDoc2Dial 2022, *Docalog* achieved f1-scores of 36.07% and 28.44% and SacreBLEU scores of 23.70% and 20.52%, respectively on the *MDD-SEEN* and *MDD-UNSEEN* folds.

References

Rubén Baena-Perez, Iván Ruiz-Rube, Juan Manuel Dodero, and Miguel Angel Bolivar. 2020. A framework to create conversational agents for the development of video games by end-users. In *International Conference on Optimization and Learning*, pages 216–226. Springer.

Urmil Bharti, Deepali Bajaj, Hunar Batra, Shreya Lalit, Shweta Lalit, and Aayushi Gangwani. 2020. Medbot: Conversational artificial intelligence powered chatbot for delivering tele-health after covid-19. In *2020 5th international conference on communication and electronics systems (ICCES)*, pages 870–875. IEEE.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA

- accessing domain-specific FAQs via conversational QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7302–7314, Online. Association for Computational Linguistics.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Zhe Wang, and Doo Soon Kim. 2020. TutorialVQA: Question answering dataset for tutorial videos. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5450–5455, Marseille, France. European Language Resources Association.

Nico Daheim, David Thulke, Christian Dugast, and Hermann Ney. 2021. Cascaded span extraction and response generation for document-grounded dialog. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 57–62, Online. Association for Computational Linguistics.

Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. 2021. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence*, 299:103525.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020a. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

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. 2020b. 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.

Zuohui Fu, Yikun Xian, Yongfeng Zhang, and Yi Zhang. 2020. Tutorial on conversational recommendation systems. In *Fourteenth ACM Conference on Recommender Systems*, pages 751–753.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Boeun Kim, Dohaeng Lee, Sihyung Kim, Yejin Lee, Jin-Xia Huang, Oh-Woog Kwon, and Harksoo Kim. 2021. Document-grounded goal-oriented dialogue systems on pre-trained language model with diverse input representation. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 98–102, Online. Association for Computational Linguistics.

Jiapeng Li, Mingda Li, Longxuan Ma, Wei-Nan Zhang, and Ting Liu. 2021. Technical report on shared task in DialDoc21. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 52–56, Online. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications*, 129:56–67.

Lasguido Nio, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2014. Utilizing human-to-human conversation examples for a multi domain chat-oriented dialog system. *IEICE TRANSACTIONS on Information and Systems*, 97(6):1497–1505.

Yasuhito Ohsugi, Itsumi Saito, Kyosuke Nishida, Hisako Asano, and Junji Tomita. 2019. A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 11–17, Florence, Italy. Association for Computational Linguistics.

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.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine

reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. MIMOQA: Multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, Online. Association for Computational Linguistics.

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

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Thirty-first AAAI conference on artificial intelligence*.

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2021. Conversational question answering: A survey. *arXiv preprint arXiv:2106.00874*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.