A Knowledge Storage and Semantic Space Alignment Method for Multi-documents Dialogue Generation

Minjun Zhu[1,2], Bin Li[3], Fei Xia[1,2], Yixuan Weng[1]*
[1] National Laboratory of Pattern Recognition,, Institute of Automation, CAS
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences
[3] College of Electrical and Information Engineering, Hunan University
{zhuminjun2020,xiafei2020}@ia.ac.cn, libincn@hnu.edu.cn, wengsyx@gmail.com

Abstract

Question Answering (QA) is a Natural Language Processing (NLP) task that can measure language and semantics understanding ability, it requires a system not only to retrieve relevant documents from a large number of articles but also to answer corresponding questions according to documents. However, various language styles and sources of human questions and evidence documents form the different embedding semantic spaces, which may bring some errors to the downstream QA task. To alleviate these problems, we propose a framework for enhancing downstream evidence retrieval by generating evidence, aiming at improving the performance of response generation. Specifically, we take the pre-training language model as a knowledge base, storing documents' information and knowledge into model parameters. With the Child-Tuning approach being designed, the knowledge storage and evidence generation avoid catastrophic forgetting for response generation. Extensive experiments carried out on the multi-documents dataset show that the proposed method can improve the final performance, which demonstrates the effectiveness of the proposed framework.

1 Introduction

With the rapid and vigorous development of the field of artificial intelligence and language intelligence, Question Answering (QA) systems has received more and more extensive attention. Specifically, the QA system aims to provide precise answers in response to the user's questions in natural language. An essential task in the QA system is conversational question answering and document-grounded dialogue modeling. The conversational question answering dialogue-like interface that enables interaction between human users and the documentation provides sufficient information. Prior work typically formulates the task as a machine reading comprehension task assuming the associated document or text snippet is given, such as QuAC (Choi et al., 2018), ShARC (Saeidi et al., 2018), CoQA (Reddy et al., 2019), OR-QuAC (Qu et al., 2020) and Doc2Dial (Feng et al., 2020).

One of the difficulties of conversational QA tasks is to model the historical information in the process of system retrieval and generation. The recently released conversational question answering datasets like CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018) aim to lead a reader to answer the latest question by comprehending the given context passage and the conversation history. As they provide context passages in their task setting, they omit the stage of document retrieval. While on the MultiDoc2Dial (Feng et al., 2021) dataset, retrieval is necessary. Recently, Qu et al. (2020) extend the QuAC dataset to a new OR-QuAC dataset by adapting to an open retrieval conversational question answering system (OpenConvQA), it can retrieve relevant passages from a large collection before inferring the answer, taking into account the conversation QA pairs, which is similar with the MultiDoc2Dial dataset.

To enhance the modeling of historical sessions and avoid the problem of weak semantic relatedness between problems and evidence in the retrieval stage. we propose a novel three-stage framework, which stores knowledge and makes alignment in semantic space. Specifically, we find that it is inconsistent to search for most question-related evidence only by the inner product of the question and long text of dialogue history. As stated by Feng et al. (2021) about task 2: Agent Response Generation is more difficult than task 1: Grounding Span Prediction, because agent utterance varies in style and is not directly extracted from document content. Different language styles and sources lead to different semantic spaces of question and evidence document embedding. As a result, it inspires us

*Corresponding author.

Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 130 - 135

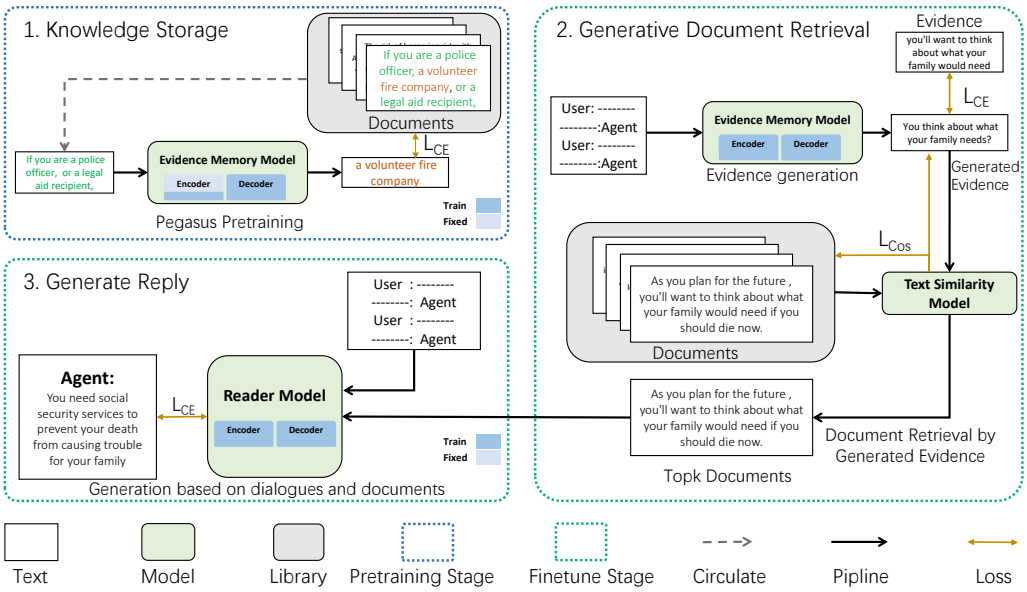

Figure 1: The overview of the proposed three-stage method, where the evidence memory model in the Knowledge Storage and the Generation Document retrieval module is the same T5.

to propose a framework that uses model-generated evidence to enhance question-related evidence. We summarize our contributions as follows:

- To address the inconsistency between the semantic space of questions and evidence documents, we propose a framework for enhancing downstream evidence retrieval by generating evidence and enhancing the performance of response generation.
- We take the pre-training language model as a knowledge base, and store documents' information and knowledge into model parameters through the Pegasus pre-training method (Zhang et al., 2019), which effectively improves the memory of the pre-trained language model for documents. This constitutes our knowledge storage stage.
- We applied the Child-Tuning approach in Xu et al. (2021) to knowledge storage and evidence generation to avoid catastrophic forgetting caused by two-stage training.

2 Main method

In this section, the overall framework is illustrated in Figure 1, where we will elaborate on the main method for the MultiDoc2Dial task. Based on the pre-trained language model, we design a three-stage semantic alignment method including the knowledge storage stage, generative document retrieval, and reply generation modules, which are described in turn as follows.

2.1 Knowledge storage

In this stage, we trained the pre-training language model for knowledge storage. Because the semantic space of the question embedding and the documents' embedding is inconsistent (Feng et al., 2021), we generate additional possible evidence as auxiliary features to increase the semantic alignment of embedding in downstream tasks. The traditional retrieval method (Qu et al., 2020) is designed to search related documents based on question embedding and documents embeddings. However, in this scenario, the genre, style, and size of questions and documents are different, which will lead to question and documents embeddings in different semantic spaces. To improve the accuracy of document retrieval, they should be searched in the same semantic space. we believe that the maximum inner product search of relevant evidence-based evidence can match with stronger semantic relevance. Before generating evidence, we use the pre-training method to make the model memory document knowledge more deeply. We pre-trained T5 (Raffel et al., 2019; Tay et al., 2021) with Pegasus (Zhang et al., 2019) method, randomly sampling 3/4 of the sentences of the document and training the model to generate the other 1/4 of the sentences. We think this way can enable the model to learn complete document information. In addition,

Model	Method	F1_U	sacreBLEU_U	Meteor_U	Rouge_U	All
T5 Model	Finetune with Utterance	28.090	12.386	25.627	26.199	92.302
Pegasus Pre-trained Model	Zeroshot	10.485	1.144	8.723	10.267	41.104
	Finetune with Utterance	28.556	13.062	26.429	26.434	94.481
	Finetune with Evidence	**35.672**	**16.171**	**34.318**	**34.013**	**130.174**

Table 1: Comparison results between different methods without using retrieval.

Following Xu et al. (2021), we use the child-tune method to perform Pegasus (Zhang et al., 2019) pre-training, only 25% of the parameters of the encoder and 100% of the parameters of the decoder are detected as the most important child network for the target task. Fisher information for the i-th parameter is as follows:

$$F_i = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{\partial \log p\left(y_j \mid x_j;\theta\right)}{\partial \theta_i}\right)^2 \quad (1)$$

After this phase, we believe it will benefit the later evidence generation task.

2.2 Generate document retrieval

The goal of generating document retrieval is to obtain the most relevant evidence-based on the question and dialogue history. We formulate this problem in two steps. First, we use the evidence memory model to generate relevant evidence. The model trained from the knowledge storage can be considered evidence modeling. Second, the generated evidence is used to retrieve a shred of evidence from the document collections. We use SimCSE Model[1] (Gao et al., 2021) to obtain the embedding of the question and generated shreds of evidence, and use MIPS (Maximum Inner Product Search) to get the most relevant evidence from the evidence base. For top-k searching, we use the loss function based on the Cos function for training.

$$L_{\text{Cos}} = \log \frac{e^{\text{sim}\left(h_i, h_i^+\right)/\ell}}{\sum_{j=1}^{N} e^{\text{sim}\left(h_i, h_j^+\right)/\ell}} \quad (2)$$

2.3 Reply generation

In the reply generation module, the T5 model is used to generate the next sentence reply answer with the input obtained in the previous generate document retrieval module.

[1] https://huggingface.co/princeton-nlp/sup-simcse-roberta-large

Domain	# Doc	# Dial	Two-seg	>Two-seg	Single
Ssa	109	1191	701	188	302
Va	1398	1337	648	491	198
Dmv	149	1328	781	257	290
Student	92	940	508	274	158
Total	488	4796	2638	1210	948

Table 2: Statistics of the MultiDoc2Dial task dataset.

3 Experimental

3.1 Data description

MultiDoc2Dial (Feng et al., 2021) is a Multi-Document-Grounded Dialogue dataset, which is derived from Doc2Dial dataset (Feng et al., 2020) with changing a single document to multiple documents. The task is to generate grounded agent responses given dialogue queries and domain documents. Specifically, the system gets the latest user turn, dialogue history, and all domain documents as inputs, and requires the system to return agent responses in natural language. The specific distribution of the MultiDoc2Dial task data set is shown in Table 2.

3.2 Evaluation metrics

We follow the previous settings in Feng et al. (2020, 2021). In the retrieval task, we calculate recall (@1), which measures the number of correct documents found in the first prediction. We report token-level F1 scores, Exact Match (EM) (Rajpurkar et al., 2016) scores, and sacreBLEU (Post, 2018) scores for the generated text.

3.3 Implementation details

In these tasks, we are mainly based on the huggingface framework[2] (Wolf et al., 2020). We use the AdamW (Loshchilov and Hutter, 2018) optimizer. Linear decay of learning rate and gradient clipping of 1e-4. Dropout (Srivastava et al., 2014) of 0.1 is applied to prevent overfitting. We implemented the code of training and reasoning based on PyTorch[3] (Paszke et al., 2019) in one NVIDIA A100 GPU.

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

[3] https://pytorch.org

Method	D^{token}-bm25	D^{struct}-bm25	D^{token}-nq	D^{struct}-nq	D^{token}-ft	D^{token}-ft	$GDR^{w/o}$-ques.	GDR^{with}-ques.
Top-1 Acc	20.5	18.0	27.7	28.6	36.4	39.1	24.5	**42.5**

Table 3: Result of the TopK accuracy in the retrieval task between different baseline methods, where GDR means generate document retrieval, and with and w/o-ques. mean whether adding input question.

Model	Method	F1	Exact Match	sacreBLEU	All
Baseline	D^{struct}-bm25	27.9	2.0	12.5	42.4
	D^{struct}-nq	33.0	3.6	17.6	54.2
	D^{struct}-ft	<u>36.0</u>	<u>4.1</u>	<u>21.9</u>	<u>62.0</u>
Pegasus Pretrained Model	without retrieval	35.7	3.9	16.2	55.8
	with retrieval	34.4	3.0	20.6	58.0
T5 Model	without retrieval	28.1	2.9	15.6	46.6
	with retrieval	**43.4**	**5.1**	**24.8**	**73.3**

Table 4: Comparison with different methods for the final results on the Validation set. In the baseline, we follow the previous settings: Struct means the corresponding document index is based on structure-segmented passages, nq means using the original pre-trained bi-encoder from DPR, ft means fine-tune. We adopt <u>underline</u> to show the score of second place.

All experiments select the best parameters on the valid set and then report the score of the best model (valid set) on the test set.

Knowledge storage We use Google's open-source T5 large model[4] for pre-training. We use the AdamW (Loshchilov and Hutter, 2018; Xu et al., 2021) optimizer and the learning rate is set to $1e-4$ with the warm-up (He et al., 2016). We also fixed some parameters in the T5 model whose gradient change was less than 75% of all parameters in the first round of training. The batch size is 6. We set the maximum length of 350. We intercepted according to the document fragments, randomly selected $1/4$ of the subfragments as labels, and repeated 50 rounds as knowledge storage.

Generate document retrieval We fine-tune the Knowledge Storage Model with "context -> evidence", and then we use this model to generate the evidence of the dev set. After that, we use the Text Similarity Model[5] (Gao et al., 2021) to retrieve the top K documents from the document library. Here, we set $K = 1$. In detail, we input the final problem into the model together with the evidence generated by the previous model. Then use the same model to obtain the semantic vectors of all documents, and use cosine similarity to calculate the most similar documents.

Reply generation We re-use a new T5 model, which uses "the last question of the dialogue </s> dialogue history information </s> related documents" to fine-tune. We set the maximum length of 700 and batch size is set at 6. If the document content exceeds the limit, it will be deleted.

3.4 Experimental results

We conducted three comparative experiments as shown in Table 1, Table 3 , Table 4and Table 5 respectively, where the first is the non-retrieval experiment. When training with utterance as the label, compared with T5 original model, the model trained with Pegasus can obtain better performance. Even without training samples, it can achieve good results. It is worth noting that better performance can be achieved if the evidence is used as a training label. The reason may be that in this training scenario the output is relatively consistent with Pegasus training, which can stimulate the potential knowledge base features of the model. In the retrieval task shown in Table 3, we first use the context to generate possible evidence, then fine-tune it in Simcse, then find the most likely documents based on MIPS. We tested two cases, one in which the generated evidence is embedded into the semantic vector for retrieval, and the other in which the question and the generated evidence are co-embedded into the semantic vector for retrieval. The experimental results show that although the retrieval performance of single evidence is not good, it can achieve better results if it is used as input together with the problem as an additional auxiliary feature. After the retrieval performance is improved, we use the T5 model to take evidence and context for training. About all the evaluation metrics, on the validation set, we conduct an exhaustive comparison experiment among our Pega-

[4] google/t5-efficient-large-nl36

[5] https://huggingface.co/princeton-nlp/sup-simcse-roberta-large

Model	F1	sacreBLEU	METEOR	RougeL	Total
Baseline	35.85	22.26	34.28	33.82	126.21
Ours	**36.69**	**22.78**	**35.46**	**34.52**	**129.44**

Table 5: Comparison with different methods for the final results on the Test set.

sus Pre-trained Model, T5, and baselines in Table 4. And it can also be significantly improved compared with the baseline methods on the Test set, which is shown in Table 5.

4 Conclusion

In this paper, we propose a generative evidence retrieval method, which transforms the context and problems into possible evidence for further retrieval. Specifically, we first use Pegasus to completely save the knowledge base into the language model and use Child-tune to avoid the catastrophic forgetting problem for response generation. More precisely, it avoids the problem of weak semantic relatedness between the "question text" to be retrieved and the retrieved "answer text", and can effectively increase the accuracy of retrieval. In the future, we will study how to combine the evidence generation model with the utterance generation model to further improve the generation quality.

References

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 *Empirical Methods in Natural Language Processing*.

Song Feng, Kshitij P. Fadnis, Q. Vera Liao, and Luis A. Lastras. 2020. Doc2dial: A framework for dialogue composition grounded in documents. In *National Conference on Artificial Intelligence*.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Matt Post. 2018. A call for clarity in reporting bleu scores.

Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W. Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing*.

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 S. H. Lewis, Sameer Singh, Tim Rocktäschel, Michael Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Empirical Methods in Natural Language Processing*.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2021. Scale efficiently: Insights from pre-training and fine-tuning transformers. *CoRR*, abs/2109.10686.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz,

Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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 (EMNLP)*. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.