Capturing Conversational Interaction for Question Answering via Global History Reasoning

Jin Qian[1], **Bowei Zou**[2], **Mengxing Dong**[1], **Xiao Li**[1], **Ai Ti Aw**[2], **Yu Hong**[1]*
[1]School of Computer Science and Technology, Soochow University, China
[2]Institute for Infocomm Research, A*STAR
{jaytsien, ayumudong, emilyxiao0512, tianxianer}@gmail.com
{zou_bowei,aaiti}@i2r.a-star.edu.sg

Abstract

Conversational Question Answering (ConvQA) is required to answer the current question, conditioned on the observable paragraph-level context and conversation history. Previous works have intensively studied history-dependent reasoning. They perceive and absorb topic-related information of prior utterances in the interactive encoding stage. It yielded significant improvement compared to history-independent reasoning. This paper further strengthens the ConvQA encoder by establishing long-distance dependency among global utterances in multi-turn conversation. We use multi-layer transformers to resolve long-distance relationships, which potentially contribute to the reweighting of attentive information in historical utterances. Experiments on QuAC show that our method obtains a substantial improvement (1%), yielding the F1 score of 73.7%. All source codes are available at https://github.com/jaytsien/GHR.

1 Introduction

ConvQA is a task of answering questions conditioned on the conversation history as well as referential contexts. It heavily relies on the traceback of conversation history. For example, the pronoun "*it*" in Q_2 in Table 1 needs to be resolved first. It is indispensable to pursue its coreference "*alchemy index*" that appeared in the first-turn historical conversation (i.e., Q_0 and A_0). Therefore, the challenge of ConvQA is to detect the relevant evidence hidden in conversation history, and use it to strengthen the current round of question answering.

Recently, utilizing global conversation history for enhancement is increasingly gaining interest, because it potentially contributes to capturing long-distance relevant evidence for answering. Both historical-answer-aware dynamic encoding of context (Qu et al., 2019b) and flow-based interaction modeling over shifting topics (Yeh and Chen, 2019) appear as successful solutions, where global conversation history is involved. However, some historical information fails to be maintained due to 1) the omission of historical questions and 2) disconnection from the earliest-stage conversation when topic frequently shifts.

Section: Thrice: The Alchemy Index (2006-2008)
Context: In September 2006, the band announced plans for a new album (later titled The Alchemy Index) on their official website. The album was conceived as a series of 4 EPs, (...) The band maintained a studio blog titled "Alchemy Index" throughout the recording process. (...) The Alchemy Index Vols. I & II was released on October 16, 2007 and sold 28,000 copies in its first week. (...)
$\mathbf{Q}_0$: What is the alchemy index? $\mathbf{A}_0$: In September 2006, the band announced plans for a new album (later titled The Alchemy Index) on their official website.
$\mathbf{Q}_1$: What is notable about the album? $\mathbf{A}_1$: The album was conceived as a series of 4 EPs (...)
$\mathbf{Q}_2$: Was it well received? $\mathbf{A}_2$: The Alchemy Index Vols. I & II was released on October 16, 2007 and sold 28,000 copies in its first week.

Table 1: An example from QuAC with the clues in conversation history (in blue) and context (in red).

In this paper, we develop a Global History Reasoning (GHR) model. GHR is not only capable of separately encoding different rounds of question-answering (QA) conversations, but sequentially fusing the encoded information of all QA pairs in visible conversations by a multi-layer attention network. It is designed to avoid the omission of available historical information and disconnection. We experiment on QuAC (Choi et al., 2018). The test results show that GHR yields substantial improvements when using BERT and ELECTRA as the baselines, and it achieves competitive performance compared to state-of-the-art methods.

2 Approach

The input of GHR comprises the referential context c, the current question in the t-th round, and all his-

* Corresponding author. Email:tianxianer@gmail.com

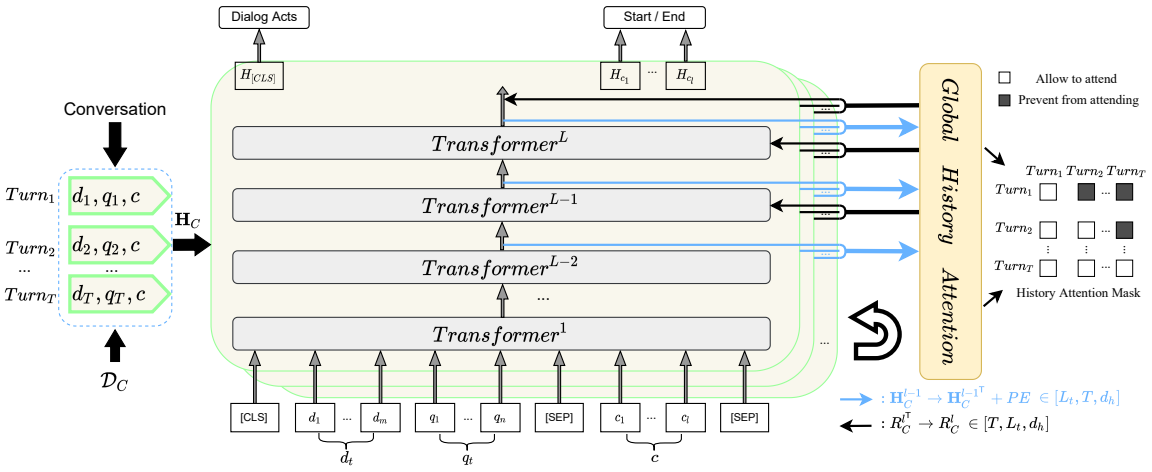

Figure 1: Architecture of global history reasoning (GHR) model.

torical QA pairs $\mathcal{H}_t = [(q_1, a_1), \ldots, (q_{t-1}, a_{t-1})]$, where q_i and a_i denote the question and answer in the i-th round in conversation, respectively. GHR models local history by learning utterance representation for the QA pair in every single round. Then, GHR fuses all visible local history and models their interaction with the referential context by global history attention. Finally, a linear layer with softmax is applied for answer prediction.

Figure 1 shows the overall architecture of GHR, where the sequentially stacked cards (drawn with light green rectangles) denote the encoding stages for the questions issued at different times. For example, the visible structure in the top card illustrates the encoding stage for the first question q_1, and at the time, both the local conversation history d_1 and the visible global history are NULL. The global attention mechanism is shown at the right side of the diagram, whose input is the representation encoded by each transformer layer (blue arrows), while the output is the refined representation by the masked global interaction (black arrows). The masking operation is used to temporally disable the subsequent QA conversations when the current question is being dealt with (as required in the task of ConvQA).

2.1 Local History Encoding

We follow the most commonly-used ConvQA scheme (Zhao et al., 2021) to form the input of our encoder. Given the current question q_t, we consider the historical QA pairs in the last two rounds $d_t = (q_{t-2}, a_{t-2}, q_{t-1}, a_{t-1})$ as the local history of q_t, and concatenate it with q_t and the context c as the input sequence. For the context whose length exceeds the maximum input length (usually 512 tokens) of the encoder, we divide them into multiple fragments and put them in a batch in order. Then we use a pre-trained language model (PLM) to encode the input sequence into contextualized representations:

$$\mathbf{h}_t = P_{LM}(d_t, q_t, c), \mathbf{h}_t \in \mathbb{R}^{L_t \times d_h} \quad (1)$$

where $P_{LM}(\cdot)$ denotes a transformer-based PLM encoder, L_t denotes the maximum length of input sequence and d_h is the hidden size.

2.2 Global History Reasoning

Most existing ConvQA studies suppose that the latest two-round conversation history has the most direct correlation to the current question. Therefore, they merely encode them into the input representation (Ohsugi et al., 2019; Ju et al., 2019) as mentioned in Eq.(1). This results in the omission of other essential information from the entire conversation, such as that signaling the long-distance reference and topic consistency. Some previous work extended the local encoding by absorbing attentive information from a larger range of QA pairs in history. However, the gradual attenuation for encoding (Yeh and Chen, 2019) causes the failure in giving more prominence to the long-distance related information. Besides, due to the limited and fixed size of the input sequence, the complete conversational interaction process actually has been divided into fragments. This makes it hard to ensure the coherence during encoding a series of conversation units (each unit is a QA pair occurring in the conversation history).

A sufficient flow of information among entire conversation units is warranted to compensate for these defects. Specifically, to enhance the interaction of multi-turn utterances, we design a global history attention mechanism to sequentially fuse contextualized representations of visible historical QA pairs $\mathcal{D}_C$. We denote the latent information representation of $\mathcal{D}_C$ as $\mathbf{H}_C = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T]$, where $\mathbf{H}_C \in \mathbb{R}^{T \times L_t \times d_h}$, T is the number of rounds of the conversation. For example, $\mathbf{h}_i$ is the representation of the i-th round QA pair.

To obtain deep interactions between different rounds, we transpose the dimension of the input matrices $\mathbf{H}_C$ as $L_t \times T \times d_h$ and obtain $\mathbf{H}_C^\intercal = [\mathbf{h}_1^\intercal, \mathbf{h}_2^\intercal, \ldots, \mathbf{h}_T^\intercal]$. Then we add absolute positional embeddings $\mathbf{p}_i$ to $\mathbf{h}_i^\intercal$ to incorporate the position information of each token. The final input embeddings is:

$$\mathbf{H}_C^\intercal = [(\mathbf{h}_1^\intercal; \mathbf{p}_1), (\mathbf{h}_2^\intercal; \mathbf{p}_2), \ldots, (\mathbf{h}_T^\intercal; \mathbf{p}_T)], \quad (2)$$

where $\mathbf{H}_C^\intercal \in \mathbb{R}^{L_t \times T \times d_h}$.

During history reasoning, we apply a Global History Attention (GHA) layer to model the whole history and learn the history-aware representations for each token in the utterance:

$$\mathbf{R}_C^\intercal = [\mathbf{r}_1^\intercal, \mathbf{r}_2^\intercal, \ldots, \mathbf{r}_T^\intercal] = GHA(\mathbf{H}_C^\intercal), \quad (3)$$

where $\mathbf{r}_i^\intercal \in \mathbb{R}^{L_t \times d_h}$ denotes the history-aware representation of the i-th round, $GHA(*)$ is a transformer layer with history attention mask. In a real dialogue scene, a speaker is able to see all happened historical utterances before the current round, while subsequent utterances are unseen. Therefore, to avoid leaking the unseen utterances, we leverage the self-attention mask mechanism (Dong et al., 2019) (as shown in the right of Figure 1) to capture the visible historical information associated with token embeddings of each position. We then re-transpose $\mathbf{R}_C^\intercal \in \mathbb{R}^{L_t \times T \times d_h}$ to obtain $\mathbf{R}_C = [\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_T] \in \mathbb{R}^{T \times L_t \times d_h}$ for easily fusing the local and global history.

To integrate the contextualized representations $\mathbf{H}_C$ with the history-aware representations $\mathbf{R}_C$, we first adopt N GHA layers to obtain $\mathbf{R}_C^l$. In particular, the l-th GHA layer (Eq.(5))[1] is connected behind the (l-1)-th Transformer layer (Eq.(4)). N is a hyper-parameter that indicates that each of the last N Transformer layers is followed by a GHA layer. Then, we perform layer normalization (Ba et al., 2016) to update the final $\mathbf{H}_C$ (Eq.(6)).

$$\mathbf{H}_C^l = Transformer(\mathbf{H}_C^{l-1}) \quad (4)$$

$$\mathbf{R}_C^l \Leftarrow GHA(\mathbf{H}_C^{l-1}) \quad (5)$$

$$\mathbf{H}_C^l = LayerNorm(\mathbf{H}_C^l + \mathbf{R}_C^l) \quad (6)$$

Finally, we get the fusion representation $\mathbf{O}_C$ by concatenating $\mathbf{H}_C^L$ and $\mathbf{R}_C^{L+1}$ for the subsequent answer prediction.

$$\mathbf{O}_C = [\mathbf{H}_C^L; \mathbf{R}_C^{L+1}]. \quad (7)$$

2.3 Answer Prediction

Given the representations $\mathbf{O}_C = [\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T]$ from the global history reasoning module, an answer span is predicted by two linear layers with softmax that calculate the probability of each token being the start and end tokens over $\mathbf{o}_t$:

$$p_t^s, p_t^e = SoftMax(Linear(\mathbf{o}_t)), t \in [1, T] \quad (8)$$

where p_t^s, p_t^e are the probabilities of the start and end positions of the answer span in the t-th round, respectively.

In the training step, we utilize cross entropy to compute the loss of the start and end predictions.

$$\mathcal{L}_{span} = -\frac{1}{T}\sum_{t=1}^{T}(y_t^s log p_t^s + y_t^e log p_t^e) \quad (9)$$

Besides optimizing by the position loss of the answer span (Eq.(9)), we also apply multi-task optimization (Zhao et al., 2021) for training. Specifically, we apply three linear layers with softmax over the "[CLS]" vector of $\mathbf{o}_t$ to determine the current question's dialog acts, including answerability, affirmation, and continuation (Choi et al., 2018).

$$p_t^{na}, p_t^{af}, p_t^{ct} = SoftMax(Linear(\mathbf{o}_t^{cls})) \quad (10)$$

where p_t^{na} denotes the fractional vector of answerable probability of the question, p_t^{af} is the probability of affirmation, and p_t^{ct} is the probability of continuing to ask subsequent questions. We use cross entropy to compute losses of the acts.

$$\mathcal{L}_{na} = -\frac{1}{T}\sum_{t=1}^{T} y_t^{na} log p_t^{na} \quad (11)$$

$$\mathcal{L}_{af} = -\frac{1}{T}\sum_{t=1}^{T} y_t^{af} log p_t^{af} \quad (12)$$

$$\mathcal{L}_{ct} = -\frac{1}{T}\sum_{t=1}^{T} y_t^{ct} log p_t^{ct} \quad (13)$$

[1]For simplicity, we use $\Leftarrow$ in Eq.(5) to indicate the calculation process with the two transpose steps mentioned above.

where y_t^{na}, y_t^{af}, and y_t^{ct} are the ground-truths of answerability, affirmation, continuation, respectively. The final optimization goal is as follow.

$$\mathcal{L} = \alpha\mathcal{L}_{span} + \beta(\mathcal{L}_{ct} + \mathcal{L}_{af} + \mathcal{L}_{na}) \quad (14)$$

where α and β are coefficients for adjusting $\mathcal{L}_{span}$ and the combination of $\{\mathcal{L}_{ct}, \mathcal{L}_{af}, \mathcal{L}_{na}\}$.

3 Experiments

3.1 Settings

We conduct experiments on QuAC (Choi et al., 2018) consisting of 100K questions obtained from 14K information-seeking dialogues, which proposes unique challenges since these questions are open-ended, descriptive, highly contextual, and probably unanswerable. In particular, many questions require sufficient co-referencing and reasoning through interactions with conversation history.

We employ BERT$_{large}$[2] (Devlin et al., 2019) and ELECTRA$_{large}$[3] (Clark et al., 2020) as local history encoders. Meanwhile, we apply the voting strategy to implement comparable baseline models. The trade-off coefficients α and β in the loss function are set to 0.7 and 0.1 respectively (Zhao et al., 2021). The max query length and the stride of sliding window of GHR is set to 128. The batch size is set to 12. The answer length is set to 50 and learning rate is 2e-5. We rely on Pytorch and HuggingFace Transformer libraries (Wolf et al., 2020) for our experiments.

Following Choi et al. (2018), we adopt word-level macro-F1 and human equivalence score (HEQ) as evaluation metrics. HEQ-Q and HEQ-D measure the percentage of answers that the model accurately predicts but the human does not by given questions and dialogues.

We compare our GHR-based ConvQA model with the state-of-the-art models that are reported for performance on the QuAC leaderboard[4], including:

BiDAF++ (Choi et al., 2018) further augments BiDAF with self-attention and contextualized embeddings.

BiDAF++ w/2-ctx (Choi et al., 2018) additionally models conversation history from the previous two turns of QA pair by encoding their positions in conversation within the question embeddings and concatenating the marker embeddings to the passage embeddings.

HAE (Qu et al., 2019a) introduces a history answer embedding to incorporate the conversation history into BERT.

FlowQA (Huang et al., 2019) feeds the model with the hidden embeddings generated by reasoning in each new round of conversation.

GraphFlow (Chen et al., 2020) encodes conversation history into context graphs for context reasoning and analysis.

FlowDelta (Yeh and Chen, 2019) passes down the information gain between different turns to ensure that the model can focus on more informative cues in context.

HAM (Qu et al., 2019b) adopts position attention embeddings for history selection and optimizes the model from both answer span prediction and dialog act prediction via a multi-task learning framework.

RoR (Zhao et al., 2021) uses chunk reader to obtain chunk answers, which are aggregated for document reader to read again, and votes for the final answer.

Models	F1	HEQ-Q	HEQ-D
BiDAF++	51.8/50.2	45.3/43.3	2.0/2.2
BiDAF++ w/2-ctx	60.6/60.1	55.7/54.8	5.3/4.0
HAE	63.9/62.4	59.7/57.8	5.9/5.1
FlowQA	64.6/64.1	-/59.6	-/5.8
GraphFlow	-/64.9	-/60.3	-/5.1
FlowDelta	66.1/65.5	-/61.0	-/6.9
HAM	66.7/65.4	63.3/61.8	9.5/6.7
RoR	75.7/74.9	73.4/72.2	17.8/16.4
BERT (ours)	67.7/-	62.9/-	7.8/-
GHR (BERT)	69.0/-	64.6/-	8.0/-
ELECTRA (ours)	73.2/72.7	69.8/68.8	12.2/11.9
GHR (ELECTRA)	74.9/73.7	71.7/69.9	14.6/13.7

Table 2: Results on the dev/test set of QuAC.

3.2 Experimental Results

Table 2 shows the comparison of the existing published models with relatively high performance on the QuAC leaderboard with our GHR models on both the dev and test set. The test results show that all improvements of GHR are statistically significant (paired t-test (Dror et al., 2018), p-value < 0.01). We obtain the following observations.

1) Compared with the baseline models, the performances of the GHR models with BERT and ELECTRA are improved by 1.3 and 1.7 absolute F1 scores on the dev set, respectively. It indicates that GHR is effective for the ConvQA task, even based on the PLM with stronger representation capability (ELECTRA). This suggests that general PLM-dependent ConvQA models may have limita-

[2]https://github.com/google-research/BERT

[3]https://github.com/google-research/electra

[4]https://quac.ai. Note that we only compare the proposed model to the methods with published papers.

Models	F1	HEQ-Q	HEQ-D
ELECTRA	73.2	69.8	12.2
w/ 1 GHA layer	74.3	71.2	14.5
w/ 2 GHA layers	74.4	71.3	14.1
w/ 3 GHA layers	**74.9**	**71.7**	14.6
w/ 4 GHA layers	74.7	71.5	14.4
w/ 5 GHA layers	73.7	70.3	**14.8**
w/ 6 GHA layers	73.3	69.8	14.0
w/ 7 GHA layers	72.6	69.3	12.7

Table 3: Effects of the number of GHA layers on the QuAC dev set.

tions without a step specifically targeting conversation history interactions.

2) GHR outperforms other models that utilize global conversation history, such as FlowQA, GraphFlow, and FlowDelta. We believe that these "flow"-based models tend to attenuate or ignore earlier conversation histories, which may prioritize recent utterances. Thus, when the current question is correlated to an earlier history or topic drifts, it is disadvantageous for the "flow"-based models. On the contrary, the GHR model still maintains a high focus on the early utterances through the GHA mechanism, so it can effectively utilize the global dialogue history.

3) Compared with the typical ConvQA model HAM, GHR outperforms it by a 2.3 absolute F1 score on the dev set with the same PLM settings (BERT). The reason might be that HAM only employs the answers of at most the first 4 rounds in conversations and pays more attention to the distribution of answer spans in the context, but does not model the question. Thus HAM is also difficult to solve the problem caused by the topic drifting. For GHR, we believe that modeling global history is the most effective factor leading to the benefits of GHR.

4) Compared with the best model on the QuAC leaderboard (RoR), our GHR is 1.2 absolute F1 scores lower than it on the test set. The reason is that RoR employs transfer learning to first finetune itself on the CoQA dataset, but such a method is meaningless for us since the target of this paper is quite different. Moreover, RoR focuses on modeling long contexts, but not modeling conversation history. It only utilizes the question in the current round. Also, GHR can be easily combined with RoR in implementation. Therefore, we suggest fusing the two models to further improve the performance of the ConvQA task.

We conduct an ablation experiment to investigate the effect of the number of the global history attention (GHA) layers on performance. Table 3 shows the results of the GHR (ELECTRA) models with from 1 to 7 of GHA layers. We can observe that the performance is improved when introducing the GHA layers, which verifies the effectiveness of our proposed approach. Moreover, GHR (ELECTRA) achieves the best performance when the number of GHA layers is 3. We also notice that when the number of layers is 1-4, the performance gaps between the models are not large, and all of them are close to the best performance. When the number of layers is greater than 4, the model performance begins to decline. We believe the reason is that the average round number of QuAC is 7, so a deeper structure may lead to overfitting.

Finally, we analyze the different results of the GHR model with its corresponding baseline, to directly observe the improvement brought by the global history attention mechanism (See Appendix A.3 for details).

4 Conclusion

In this paper, we propose a global history reasoning (GHR) approach to capture interactions between all utterances for conversational question answering. Experimental results conducted on QuAC show the effectiveness of the proposed model. In the future, we will explore how to implement interactions between conversation history and context by the position features of history answers. In addition, we will extend our conversation history modeling approach to the knowledge-grounded conversation generation task (see more related work in Appendix A.1).

Acknowledgements

The research is supported by National Key R&D Program of China (2020YFB1313601), National Science Foundation of China (62076174) and Institute of Infocomm Research of A*STAR (CR-2021-001).

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Yu Chen, Lingfei Wu, and Mohammed J. Zaki. 2020. Graphflow: Exploiting conversation flow with graph neural networks for conversational machine comprehension. In *Proceedings of the Twenty-Ninth Inter-*

national Joint Conference on Artificial Intelligence, IJCAI 2020, pages 1230–1236. ijcai.org.

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.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2019. Flowqa: Grasping flow in history for conversational machine comprehension. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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 *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Gangwoo Kim, Hyunjae Kim, Jungsoo Park, and Jaewoo Kang. 2021. Learn to resolve conversational dependency: A consistency training framework for conversational question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6130–6141, Online. Association for Computational Linguistics.

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.

Chen Qu, Liu Yang, Minghui Qiu, W. Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. BERT with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1133–1136. ACM.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W. Bruce Croft, and Mohit Iyyer. 2019b. Attentive history selection for conversational question answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1391–1400. ACM.

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.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 355–363.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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.

Mark Yatskar. 2019. A qualitative comparison of CoQA, SQuAD 2.0 and QuAC. 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 2318–2323, Minneapolis, Minnesota. Association for Computational Linguistics.

Yi-Ting Yeh and Yun-Nung Chen. 2019. FlowDelta: Modeling flow information gain in reasoning for conversational machine comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 86–90, Hong Kong, China. Association for Computational Linguistics.

Jing Zhao, Junwei Bao, Yifan Wang, Yongwei Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2021. RoR: Read-over-read for long document machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1862–1872, 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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3377–3390, Online. Association for Computational Linguistics.

Chujie Zheng, Yunbo Cao, Daxin Jiang, and Minlie Huang. 2020. Difference-aware knowledge selection for knowledge-grounded conversation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 115–125, Online. Association for Computational Linguistics.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

A Appendix

A.1 Related Work

To model the history information in conversations, a kind of study focus on how to explicitly select important history question-answer pairs at the input step (Zhu et al., 2018; Reddy et al., 2019; Ju et al., 2019). Besides, a few studies argue that historical answer spans in context are more crucial, so they mark the answers when encoding the context (Choi et al., 2018; Qu et al., 2019a; Ohsugi et al., 2019; Qu et al., 2019b).

Another mainline of studies focus on conversation history reasoning. Huang et al. (2019) proposes a flow operation, feeding models entire hidden representation obtained by reasoning process when answering previous questions. The hidden states of each turn are passed back in turn by a unidirectional GRU. To avoid the changes of the captured representations during multi-turn reasoning, Yeh and Chen (2019) propose a flowdelta mechanism to explicitly capture the information gain in the conversation flow. Moreover, Chen et al. (2020) implement history reasoning by a flow operation on context graphs.

Recently, Vakulenko et al. (2021) propose to rewrite the current question using conversation history, with the goal to seek dependable clues to recover the default contents or resolve the coreferences. On the basis, Kim et al. (2021) resolve the conversational dependency via consistency regularization, and jointly use the original and rewritten questions to lead the supervised learning of QA models.

Modeling conversation history is a hot research area. It has also been widely studied in the direction of Knowledge-Grounded Conversation Generation (Kim et al., 2020; Zheng et al., 2020; Zhao et al., 2020). Such methods tend to model the full context and knowledge to generate responses, while GHR tends to model the context that is effective for answering the current question, and simultaneously learn the global history reasoning gain brought by different turns of history and the optimization of answer extraction.

A.2 Experiments on CoQA

CoQA (Reddy et al., 2019) is another typical ConvQA dataset. Table 4 lists the results of GHR model with from 1 to 7 of the GHA layers. It shows that the GHR model only improves 0.7% F1 over the baseline ELECTRA model on CoQA's dev

Models	F1
ELECTRA	89.0
w/ 1 GHA layer	89.3
w/ 2 GHA layers	89.5
w/ 3 GHA layers	**89.7**
w/ 4 GHA layers	89.5
w/ 5 GHA layers	89.3
w/ 6 GHA layers	88.8
w/ 7 GHA layers	88.6

Table 4: Effects of the number of GHA layers on the CoQA dev set.

Case #1 **Section**: Early political career: John Sherman Cooper (id: C_5caef3e3024c4f9294e1dacda1ff09b7_1)
$\mathbf{Q}_0$: What was the first job he held? $\mathbf{A}_0$: After being urged into politics by his uncle, Judge Roscoe Tartar, Cooper ran unopposed for a seat in the Kentucky House of Representatives $\mathbf{Q}_1$: What was the first office he ran for? $\mathbf{A}_1$: Kentucky House of Representatives $\mathbf{Q}_2$: How long was he in office? $\mathbf{A}_2$: CANNOTANSWER
Current Question $\mathbf{Q}_3$: Did he run for another political office after that? **Ground Truth**: In 1929, Cooper declared his candidacy for county judge of Pulaski County. **ELECTRA**: In 1939, he sought the Republican gubernatorial nomination. **GHR (ELECTRA)**: In 1929, Cooper declared his candidacy for county judge of Pulaski County.

Table 5: An example of the results predicted by GHR (ELECTRA) and ELECTRA on the dev set.

set. This is because CoQA's problems are quite straightforward, most of which can be predicted without conversation history (Yatskar, 2019). Thus CoQA is not suitable for our global history reasoning goals. Nevertheless, we still find that the effect of GHA layers on GHR was consistent in CoQA and QuAC, which means that GHR based on the same pretraining model performed relatively stable in different ConvQA tasks. As a result, GHR can play a role in modeling global history information in a variety of ConvQA tasks.

A.3 Case Study

The example in Table 5 compares the predictions between the model with ELECTRA and GHR (ELECTRA) model. On the dev set, we observe that GHR (ELECTRA) outperforms the ELECTRA model without the global history reasoning mechanism in almost all instances that contain similar long-distance referential relations. For example, we can see in Case #1, to correctly answer the current question Q_3, the model needs to infer that *he* refers to *Cooper* in A_0.