

Capturing Global Structural Information in Long Document Question Answering with Compressive Graph Selector Network

Yuxiang Nie¹²³, Heyan Huang^{123*}, Wei Wei⁴, Xian-Ling Mao¹²³

¹School of Computer Science and Technology, Beijing Institute of Technology

²Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications

³Beijing Institute of Technology Southeast Academy of Information Technology

⁴Huazhong University of Science and Technology

{nieyx,hhy63,maoxl}@bit.edu.cn,weiw@hust.edu.cn

Abstract

Long document question answering is a challenging task due to its demands for complex reasoning over long text. Previous works usually take long documents as non-structured flat texts or only consider the local structure in long documents. However, these methods usually ignore the global structure of the long document, which is essential for long-range understanding. To tackle this problem, we propose Compressive Graph Selector Network (CGSN) to capture the global structure in a compressive and iterative manner. The proposed model mainly focuses on the evidence selection phase of long document question answering. Specifically, it consists of three modules: local graph network, global graph network and evidence memory network. Firstly, the local graph network builds the graph structure of the chunked segment in token, sentence, paragraph and segment levels to capture the short-term dependency of the text. Secondly, the global graph network selectively receives the information of each level from the local graph, compresses them into the global graph nodes and applies graph attention to the global graph nodes to build the long-range reasoning over the entire text in an iterative way. Thirdly, the evidence memory network is designed to alleviate the redundancy problem in the evidence selection by saving the selected result in the previous steps. Extensive experiments show that the proposed model outperforms previous methods on two datasets.¹

1 Introduction

Long document question answering (LDQA) is a task to select relevant evidence and answer questions over long text (Dasigi et al., 2021). Compared to the traditional QA tasks, whose input is often

under 512 tokens², the input of LDQA can be more than 20K tokens.

LDQA methods can be divided into two categories: end-to-end methods and select-then-read methods. The end-to-end methods usually take a question and a long text as input to select evidence and produce the answer in one step. For example, Dasigi et al. (2021) use Longformer-Encoder-Decoder (LED) model to select evidence in the encoder part and generate answers in the decoder part. The select-then-read methods firstly apply an evidence selection model to obtain evidence pieces in a long document and then use an answer generation model to generate answers given the evidence pieces and the question. These methods mainly focus on the evidence selection phase. For example, Karpukhin et al. (2020) and Zhu et al. (2021) select paragraphs in an open domain retrieving manner. Zheng et al. (2020a) and Ainslie et al. (2020) build structure on the chunked documents for evidence selection. Gong et al. (2020) model information flows among chunks to enhance the ability of the model in selecting the evidence. However, most of the two kinds of works ignore the global structure of a long document when selecting evidence pieces, which is crucial to long-range understanding. Therefore, improvement on the evidence selection phase is needed.

Motivated by the human reading process: selectively memorizing the important pieces of information and integrating them, we propose an evidence selection model in the select-then-read method, named Compressive Graph Selector Network (CGSN). It aims to capture the global structural information in a compressive and iterative manner. Specifically, the model is composed of three modules: the local graph network, the global graph network and the evidence memory

*Corresponding author

¹We have released our codes and data in <https://github.com/JerryNie/CGSN>.

²In this paper, ‘token’ means sub-tokens split from a text sequence by a specific pre-trained tokenizer.

network. Firstly, the local graph takes a segment³ of the document as input and implements graph attention among tokens, sentences, paragraphs and the segment itself. Secondly, the global graph module selectively receives the information from the local graph and compresses it with the stored information via multi-head attention. Then, the graph attention is applied to the global graph to integrate the global structural information, which is written back to the local graph nodes to enhance the expression of local nodes for evidence selection. Thirdly, the evidence memory network receives and summarizes the evidence selecting results and sends them into the global network to alleviate the evidence redundancy problem.

Extensive experiments on two datasets show that CGSN outperforms previous methods in the evidence selection phase. Using the same answer generator as the previous methods do, CGSN further reaches the best results in the answer generation phase.

Our contributions are as follows:

- To the best of our knowledge, we are the first to consider the global structure in the long document QA task.
- With the enhancement of global structural information, the proposed model, CGSN outperforms previous methods.

2 Related Works

Long Document Question Answering. Long Document question answering aims to answer the question with the comprehension of the long document and applies multi-hop reasoning among retrieved evidence paragraphs. Dasigi et al. (2021) take advantage of the pre-trained model LED (Beltagy et al., 2020) and treat the input as a long sequence to predict the evidence paragraphs and generate the answer. Zheng et al. (2020a) and Ainslie et al. (2020) model the structure on the chunked document to select the evidence paragraph. Although Ainslie et al. (2020) claims that they explicitly model the structure of long documents, the input of their model is limited in 4K tokens, which can be regarded as a relatively long chunk. Gong et al. (2020) use the recurrent mechanism to enable information flow through different chunks for evidence selection. Karpukhin et al. (2020) and Zhu et al. (2021) search for relevant evidence from

individual paragraphs in the long document. However, most of these works model the long document as a flat sequence or consider the local structure in the document segments while global structural information of the document is nearly neglected.

Graph Neural Networks. Graph neural network (GNN) is popular in various tasks (Yao et al., 2019; Schlemper et al., 2019) due to its effectiveness in modeling structural information. Among different variants of GNNs, Graph Attention Network (Velickovic et al., 2018) (GAT) can take advantage of the attention mechanism in a graph, attending neighborhood node features to the node by different attention weights. Zheng et al. (2020b) make use of a graph multi-attention network to predict traffic conditions. Abu-El-Haija et al. (2018) take advantage of the graph attention to automatically guide the random walk in graph generation. In natural language tasks, due to the limit of memory usage, GAT is often used to model short sequences. Therefore, modeling the graph structure of the long sequence is nearly unexplored.

Memory Networks. Memory network (Weston et al., 2015) is used in memorizing long-term information via learnable reading/writing components. It is first applied to the QA task for knowledge base reasoning, which also achieves much progress in summarization (Cui and Hu, 2021) and visual question answering. To memorize plenty of information, the memory network learns to read and recurrently write into an external memory via attention. Miller et al. (2016) propose Key-Value Memory Network to flexibly access knowledge for question answering. Lu et al. (2020) design a context memory for cross-passage evidence reasoning. However, these methods only consider the memory on a single level, while structural information is disregarded.

3 Compressive Graph Selector Network

In this section, we first formalize the long document question answering (LDQA) task, and then introduce the proposed evidence selection model, i.e. Compressive Graph Selector Network (CGSN). As for the answer generator, we use a vanilla LED as the answer generator and describe the implementation details in Appendix C. Finally, we discuss the advantages of the select-then-read methods over the end-to-end methods.

³A 'segment' is a series of paragraphs in a document.

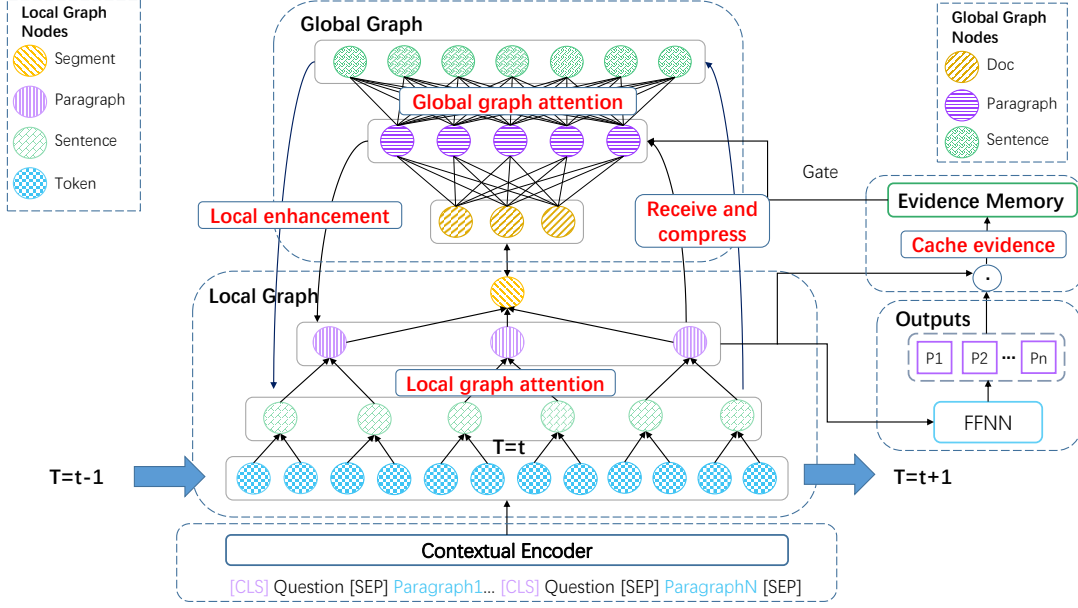


Figure 1: The architecture of CGSN.

3.1 Problem Formulation

The input to LDQA is a question $\mathbf{q} = [q_1, q_2, \dots, q_m]$ coupled with a document $\mathbf{d} = [\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n]$, $\mathbf{p}_i = [t_{i,1}, t_{i,2}, \dots, t_{i,k_i}]$ ($1 \leq i \leq n$) where m denotes the length of the question, n denotes the number of paragraphs in the document and k_i denotes the length of paragraph i . The length of the document is defined as the sum of the length of each paragraph: $c = \sum_{i=1}^n k_i$. In the LDQA setting, the length c is often unlimited, which can be larger than 20K. The goal of LDQA is to produce the evidence paragraphs $\{\mathbf{p}_{e_i}\}_{i=1}^{e^q}$ and generate the free-form answer $\mathbf{a} = [a_1, a_2, \dots, a_r]$ based on \mathbf{q} and \mathbf{d} , where \mathbf{p}_{e_i} denotes the e_i th paragraph in the document (the i th paragraph in the evidence set), e^q is the number of evidence for question \mathbf{q} .

3.2 Overview of the Model

To explore the global graph structure of long sequences, we propose Compressive Graph Selector Network (CGSN), which operates in an iterative and compressive way. CGSN is composed of three modules, local graph network, global graph network and evidence memory network. As shown in Figure 1, firstly, in time step⁴ $T = t$, the local graph network takes the t th segment of a document as the input and models the graph structure in token, sentence, paragraph and segment levels. Secondly, the global graph selectively receives the information from each granularity, compresses them into

⁴The ‘time step’ is the order of the segment to be processed.

the corresponding global graph nodes, implements the graph attention among the global graph nodes and sends the global-attended information back to the local graph to enhance the expression of the local graph nodes for evidence selection. Thirdly, the evidence memory network receives the enhanced paragraph nodes from the local graph, summarizes and caches them via the predicting logits. At the beginning of the time step $T = t + 1$, the stored memory is sent and fused with the global graph nodes in order to alleviate the redundant evidence selection problem. The detailed architecture is described in Appendix A.

3.3 Local Graph Network

Input Format Let $\mathbf{Seg}_k = [\mathbf{p}_{k,1}, \dots, \mathbf{p}_{k,N_{\text{seg}}}]$ be the k th segment in a document, which composed of N_{seg} paragraphs. To build the local graph network, firstly, we encode the N_{seg} paragraphs paired with the question. For each question-paragraph pair, the input format is “[CLS] q [SEP] p_i [SEP]”, where $1 \leq i \leq N_{\text{seg}}$. We set the embeddings of each input pair as $\mathbf{E} \in \mathbb{R}^{\ell \times d_w}$, where ℓ is the length of the input and d_w is the dimension of the embedding. The N_{seg} embedding sequences are stacked as $\mathbf{E}_k \in \mathbb{R}^{N_{\text{seg}} \times \ell \times d_w}$ and sent into the encoder f_e as follows:

$$\mathbf{H}_k = f_e(\mathbf{E}_k) \quad (1)$$

where $\mathbf{H}_k \in \mathbb{R}^{N_{\text{seg}} \times \ell \times d_h}$ is the contextual encoding, d_h denotes its dimension. In general, we use pre-trained Transformer encoders (Devlin et al.,

2019; Beltagy et al., 2019).

Local Graph Initialization Let $\mathbf{h}_{t_i}^L, \mathbf{h}_{\text{sent}_i}^L, \mathbf{h}_{p_i}^L, \mathbf{h}_{\text{seg}}^L$ be the node representations of the local graph in token, sentence, paragraph and segment level, where t_i denotes the i th token in the segment, sent_i denotes the i th sentence, p_i denotes the i th paragraph. And the question-aware token hidden states of the segment $\mathbf{H}_k^{\text{Seg}} \in \mathbb{R}^{\ell_k \times d_h}$ are used to initialize $\mathbf{h}_{t_i}^L (1 \leq i \leq \ell_k)$, where ℓ_k is the number of tokens in a segment. The sentence node $\mathbf{h}_{\text{sent}_i}^L$ is initialized by the mean-pooling of $\mathbf{h}_{t_i}^L (i \in \mathcal{S}_i)$, where \mathcal{S}_i denotes the set of token indexes of sentence i . The paragraph node $\mathbf{h}_{p_i}^L$ is initialized by the encodings of “[CLS]” token and the segment node $\mathbf{h}_{\text{seg}}^L$ is initialized by the mean-pooling of all the paragraph nodes.⁵

To enhance the interaction among the nodes at the same level, we apply the BiLSTM in the token-level, sentence-level and paragraph-level nodes, and fuse the encoded states with the node representation by a feed-forward neural network (FFNN) to generate the ultimate node representation:

$$\mathbf{h}_{\text{LSTM}}^L = \text{BiLSTM}(\mathbf{h}^L) \quad (2)$$

$$\mathbf{h}_{\text{init}}^L = \text{FFNN}(\mathbf{h}_{\text{LSTM}}^L, \mathbf{h}^L) \quad (3)$$

where $\mathbf{h}^L = [\mathbf{h}_1^L, \dots, \mathbf{h}_m^L]$ denotes the set of nodes at the same level ordered by the text sequence.

Local Graph Attention After initialization, the multi-hop graph attention is applied among the local graph nodes. The aim is to capture the local structural information for understanding the current segment of paragraphs. Let the edge set of the local graph \mathcal{G}^L be $\mathcal{E} = \{(i, j)\}_{(i, j) \in \mathcal{G}^L}$, where the (i, j) denotes an edge from node i to node j . In the proposed model, the edges of the local graph are predefined as unidirectional, which can only send information from the lower level into the neighboring higher one, including token to sentence, sentence to paragraph and paragraph to the segment node. The hidden output of the graph attention is then fused with the graph node representation and then used in the graph attention operation in the next hop. Let \mathbf{h}_j^o be a node representation in the local graph of hop o , followed Velickovic et al. (2018), its representation in hop $o + 1$ is:

$$e_{ij} = \frac{(\mathbf{h}_j^o \mathbf{W}^Q)(\mathbf{h}_i^o \mathbf{W}^K)^T}{\sqrt{d_z}} \quad (4)$$

⁵We also experiment CGSN with “LED Encoder”. The input is “<s> q </s> </s> p₁ </s>”. We use the last “</s>” token to initialize the paragraph node in the local graph.

where e_{ij} denotes the attention coefficients from node i to node j . To normalize them, there is:

$$\alpha_{ij} = \text{softmax}_i(e_{ij}) = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_j} \exp(e_{kj})} \quad (5)$$

Therefore, the output of the attention operation is:

$$\mathbf{z}_j^{\text{head}_m} = \sum_{i \in \mathcal{N}_j} \alpha_{ij} \mathbf{h}_i \mathbf{W}^V \quad (6)$$

$$\mathbf{h}_j^{o+1} = \text{Cat}[\mathbf{z}_j^{\text{head}_1}, \dots, \mathbf{z}_j^{\text{head}_k}] \quad (7)$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V$ are the parameters of the attention query, key and value of head $_m$, \mathcal{N}_j is the neighboring nodes of node j , $\mathbf{z}_j^{\text{head}_m}$ is the output of head $_m$ in the multi-head attention, and all these outputs are finally concatenated into the vector \mathbf{h}_j^{o+1} , which serves as the representation of node j in the time step $o + 1$.

3.4 Global Graph Network

The global graph is composed of three types of nodes: global sentence nodes $\mathbf{h}_{\text{sent},i}^G (1 \leq i \leq N_{\text{sent}}^G)$, global paragraph nodes $\mathbf{h}_{p,i}^G (1 \leq i \leq N_p^G)$ and global document nodes $\mathbf{h}_{d,i}^G (1 \leq i \leq N_d^G)$, where the token-level nodes are neglected because it’s hard to memorize a large number of the tokens in the finite nodes. N_{sent}^G, N_p^G and N_d^G are predefined node numbers in each level.

Local Information Reception and Compression

Global structural reasoning is based on the interaction among local information. Therefore, to store and process local information efficiently, each level of global graph nodes only receives information from one level in the local graph. The sentence-level global graph nodes receive the information from the sentence-level local nodes, and the same thing happens between the two kinds of paragraph nodes. In particular, the document-level nodes in the global graph receive information from the segment-level nodes in the local graph. All of the receptions are done with the multi-head attention mechanism. This procedure is similar to that in Section 3.3. The only difference is that the direction of the attention is from all of the local graph nodes to the global graph node at the corresponding level. For each global graph node i , let $\mathbf{h}_{i,\text{local}}^G$ be its local-aware global graph node representation. Based on the information from the local graph nodes, a gated network is used to fuse the representations of the global graph node \mathbf{h}_i^G and the local-aware global node $\mathbf{h}_{i,\text{local}}^G$:

Models	Answer F1		Overall
	Extractive	Abstractive	
LED (doc)	28.22	14.00	23.58
LED (gold)	52.41	23.44	42.97

Table 1: The answer F1 scores on the test set of Qasper. “doc” and “gold” denote the input types: the entire long document/gold evidence.

$$\mathbf{z}_{i,\text{fuse}}^G = \text{FFNN}(\mathbf{h}_i^G, \mathbf{h}_{i,\text{local}}^G) \quad (8)$$

$$\gamma = \text{Gate}(\mathbf{z}_{i,\text{fuse}}^G) \quad (9)$$

$$\mathbf{h}_{i,\text{fuse}}^G = (1 - \gamma)\mathbf{h}_i^G + \gamma\mathbf{z}_{i,\text{fuse}}^G \quad (10)$$

where $\mathbf{h}_{i,\text{fuse}}^G$ is the representation of the updated global graph node i .

Global Graph Attention To accomplish the global structural reasoning process, we implement the *global* graph attention. As shown in Figure 1, for each pair of nodes between different levels of the Global Graph, cross attention is used to extract information from one level of nodes to the other, including sentence-paragraph, paragraph-document and document-sentence (not displayed). This process is executed m times to fully extract global structural information. The applied attention is similar to Section 3.3.

Enhance the Local Nodes with Global Representation To enhance the local expression, the local paragraph nodes attend information from the global paragraph nodes by multi-head attention and combine the attended representations with the local paragraph nodes via a feed-forward network. Let the global-enhanced local representation of paragraph j be $\mathbf{h}_j^{L \leftarrow G}$. The evidence selection loss is:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{n_i} [\log P(b|\mathbf{h}_j^{L \leftarrow G})] \quad (11)$$

where N is the number of training instances, n_i is the number of paragraphs in the instance i and b is the binary label (“1” denotes the paragraph is an evidence paragraph, “0” otherwise).

3.5 Evidence Memory Network

The evidence memory network aims to integrate the evidence-selecting decision and enhance the results into the global network for the next time step. It is composed of a cache and a writing network. To properly update the parameters of the network, all of the manipulations are done in the next time step,

where we assume that the current time step is t and the next time step is $t + 1$. In the time step t , the evidence predicting logits of time t are converted via the softmax function into weights, which are multiplied with the corresponding local paragraph nodes, summed and cached in the time t :

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^n \exp(e_j)} \quad (12)$$

$$\mathbf{h}_{\text{sum}}^L = \sum_{i=1}^n \alpha_i \mathbf{h}_i^{L \leftarrow G} \quad (13)$$

where e_i is the logit of paragraph i in the evidence selection step, n is the number of paragraphs and $\mathbf{h}_i^{L \leftarrow G}$ is the global-enhanced representation of paragraph node i in the local graph.

At time $t + 1$, $\mathbf{h}_{\text{sum}}^L$ is fused with the global graph paragraph nodes $\mathbf{h}_{p,i}^G$ via the writing network:

$$\mathbf{h}_{i,\text{cat}}^G = \text{Cat}[\mathbf{h}_{p,i}^G, \mathbf{h}_{\text{sum}}^L] \quad (14)$$

$$\mathbf{h}_{i,\text{merge}}^G = \text{FFNN}(\mathbf{h}_{i,\text{cat}}^G) \quad (15)$$

$$\gamma = \text{Gate}(\mathbf{h}_{i,\text{cat}}^G) \quad (16)$$

$$\mathbf{h}_{i,\text{updated}}^G = (1 - \gamma)\mathbf{h}_{p,i}^G + \gamma\mathbf{h}_{i,\text{merge}}^G \quad (17)$$

where $\mathbf{h}_{i,\text{updated}}^G$ is the updated paragraph node i of the global graph in time $t + 1$.

3.6 Select-then-read Methods versus End-to-end Methods

Intuitively, end-to-end methods can learn to produce answers and usually regard evidence selection as an auxiliary task. However, there are two disadvantages to these methods. Firstly, evidence selection is essential. According to Table 1, given the gold evidence, the performance of the LED on answer generation improves by a large margin. Secondly, the long input is memory-consuming. In the long input setting, with sparse attention models (Beltagy et al., 2020), due to the relatively large local window size, the demands for GPU memory still become high. Meanwhile, the average token length of gold evidence in Qasper (Dasigi et al., 2021) is only 149.81, which shows that it’s probably memory-efficient only to regard the selected evidence as the context to generate answers. Therefore, a well-designed evidence selection model in the select-then-read method can be beneficial.

4 Experimental Setup

4.1 Datasets

Qasper Qasper (Dasigi et al., 2021) is a QA dataset for long scientific papers. It aims to select

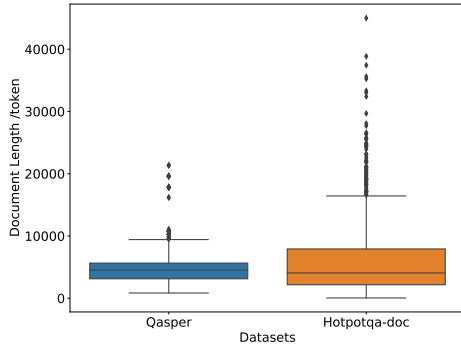


Figure 2: The box plots of document length distribution on Qasper and HotpotQA-Doc.

the evidence and make multi-hop reasoning to produce answers given the long text and the question. It contains extractive, abstractive, yes/no and unanswerable answers. Followed Dasigi et al. (2021), we treat all the answers as free-form answers.

HotpotQA-Doc HotpotQA (Yang et al., 2018) is a dataset to answer multi-hop questions based on the supporting fact sentences.

We create a long document version of the HotpotQA dataset, called *HotpotQA-Doc*. To build the dataset, we select all the hard-level questions where multi-hop reasoning is required, crawl the corresponding Wikipedia⁶ articles containing these sentences and concatenate these articles as a long document. Each paragraph with the supporting fact sentence is regarded as a gold evidence paragraph. We remove the QA pairs whose supporting facts cannot be found in any article. We split the original dev set into a new dev set and test set. We finally collect 15,658/2,468/4,937 QA pairs in train/dev/test set. Due to the change of the content in the Wikipedia pages, some answers are not guaranteed to be the span in the content but are still similar. Therefore, we regard all the answers as free-form answers.

The boxplot of Qasper and HotpotQA-Doc are shown in Figure 2.

4.2 Implementation Details

We experiment on two backbones: the SCIBERT-SCIVOCAB-uncased version of SciBERT (Beltagy et al., 2019) and the encoder of LED-base-16384 (Beltagy et al., 2020). In the SciBERT setting, the local hop number is 4 while in the LED Encoder, the local hop number is 2. For both the two backbones, global hop is 1, paragraph number in

⁶<https://www.wikipedia.org>

a segment is 16, global graph sentence nodes are 64, paragraph nodes are 32, and document nodes are 4. We searched over {1e-6, 5e-6, 1e-5, 5e-5} learning rates, {5%, 10%, 20%, 30%} warmup proportions, {1, 0.1, 0.01, 0.001} weight decays, {2, 4, 8} batch sizes and {1, 3, 5, 7} epochs. And the final learning rate is 1e-5, the warmup proportion is 10%, the weight decay is 0.01, the batch size is 4 and the epoch number is 5. We train an LED model (called *LED (evidence-trained)*) with the QA pairs and gold evidence in the training set as the answer generator (Appendix C). It receives selected evidence pieces in the testing phase from evidence-selecting methods (including CGSN and other evidence selection baselines) to further evaluate the performance of these evidence-selection methods on answer generation. It took 3.33 hours to train the model on four 11GB GPUs.

5 Experiments

In this section, firstly, we discuss the main result of the CGSN model on Qasper and HotpotQA-Doc, and then further explore the proposed model.

5.1 Main Results

We evaluate the performance of the end-to-end LED model (Dasigi et al., 2021) and the evidence selection methods⁷, including the proposed CGSN, the encoder of LED model⁸, the open domain retrieving method DPR (Karpukhin et al., 2020) and AISO (Zhu et al., 2021), evidence selection part of RCM (Gong et al., 2020), BERT-DM (Zheng et al., 2020a) and ETC (Ainslie et al., 2020). We use evidence F1 and answer F1 (Dasigi et al., 2021) as the metrics. All the retrieving methods only retrieve candidate paragraphs from the specific document. For AISO, we remove the link retrieval and stop evidence selection when the answer action is taken. In RCM, BERT-DM and ETC, we keep the evidence selection loss.

The main result of the CGSN on Qasper is shown in Table 2. It can be found that CGSN (SciBERT) outperforms all the baseline methods in evidence selection and answer generation. Although the pre-trained model ETC can deal with relatively long sequences (4K) and construct the structure via global-local attention, CGSN still outperforms

⁷These evidence selection methods are coupled with LED(evidence-trained) to evaluate the performance on answer generation.

⁸We only report its performance on HotpotQA-Doc because it is hard to converge on the training set of Qasper.

Models	Extractive	Abstractive	Yes/No	Unanswerable	Evidence	Overall
LED (Dasigi et al., 2021)*	28.22	14.00	65.88	51.92	29.20	33.12
ETC (Ainslie et al., 2020)	32.88	13.49	58.00	62.28	51.17	35.37
DPR (Karpukhin et al., 2020)	22.69	9.73	56.10	69.02	36.98	30.22
AISO (Zhu et al., 2021)	26.66	10.79	55.56	70.72	42.74	32.52
RCM (Gong et al., 2020)	30.69	11.95	64.32	70.39	47.28	35.99
BERT-DM (Zheng et al., 2020a)	31.84	14.38	63.86	66.86	48.09	36.52
AISO (SciBERT)	28.86	11.90	60.30	71.51	45.92	34.30
RCM (SciBERT)	30.85	13.92	63.45	68.53	47.69	35.88
BERT-DM (SciBERT)	32.30	12.98	62.56	71.69	48.52	36.63
CGSN (LED Encoder)	31.11	12.52	63.00	77.90	49.55	37.21
CGSN (SciBERT)	34.75	14.39	68.14	71.84	53.98	39.44
LED (Gold Evidence)	52.41	23.44	76.96	77.91	-	52.87
Human	58.92	39.71	78.98	69.44	71.62	60.92

Table 2: The performance of F1 scores on the test set of Qasper. In the first row, “Extractive, Abstractive, Yes/No, Unanswerable” are four types of answers. “Evidence” is the F1 score of evidence selection. “Overall” is the F1 score of all the answers. “*” denotes a reimplementation, which is slightly different from the reported result in the original paper. “SciBERT” and “LED Encoder” are backbone models for evidence selection methods in some of our implementations. Other implementations use the backbone models in their original papers. The last two rows are the upperbounds of the task, where the gold evidence is used or the human expert is engaged.

Models	E-F1	A-F1
LED (Dasigi et al., 2021)	68.36	51.50
LED Encoder (Beltagy et al., 2020)	76.02	53.99
ETC (Ainslie et al., 2020)	91.01	57.01
DPR (Karpukhin et al., 2020)	87.81	56.62
AISO (Zhu et al., 2021)	88.28	55.61
RCM (Gong et al., 2020)	90.03	57.20
BERT-DM (Zheng et al., 2020a)	91.45	57.63
AISO (BERT)	88.97	56.04
CGSN (LED Encoder)	92.02	57.80
CGSN (SciBERT)	92.02	57.79
CGSN (BERT)	91.88	57.92
LED (Gold Evidence)	-	58.94

Table 3: The performance of Evidence F1 (E-F1) and Answer F1 (A-F1) on the test set of HotpotQA-Doc. “BERT” denotes the model “BERT-base-uncased”.

it. For a fair comparison, we also use SciBERT as the backbone in AISO, RCM and BERT-DM. The result indicates the advantages of CGSN. Besides, to fairly compare CGSN with the LED, we also take the encoder of LED as the backbone of CGSN. Although LED Encoder is not designed as a discriminative model, CGSN (LED Encoder) also obtains strong performance.

Additionally, we also evaluate the proposed model on the HotpotQA-Doc dataset. As shown in Table 3, we notice that in evidence F1, most of the methods can reach high scores. This is because, in HotpotQA, sentence-level understanding is enough to select evidence without considering the entire Wikipedia article. Therefore, the method without

long-range modeling can get a good performance. However, the result shows that with the enhancement of the global graph, the proposed CGSN can further improve its performance.

5.2 Ablation Study

We conduct an ablation study on components of the CGSN model. As shown in Table 4, firstly, we can see the performance drops when removing the local graph, which suggests its value on segment-level reasoning. Secondly, when the global graph edges are ablated, the performance of the model is lower. Furthermore, when removing the global graph, the performance of the model is even lower, showing the importance of the global graph. Finally, the result in the last row shows that evidence memory also plays a part in the CGSN model.

5.3 Effects of Paragraph Number and Global Node Number

We conduct the experiment to explore the effectiveness of the paragraph number in a segment as well as the global graph nodes at each level. As shown in Table 5, firstly, compared between line 1 and line 2, it can be found that with the increase of paragraph number in a segment, the performance of the model gains. It is intuitive because when the number of paragraphs increases, the “perception field” of the local graph becomes larger, which leads to better performance. Secondly, compared to lines 1,3,4 and 5, it can be observed that line 1 is a proper setting for the global graph. Thirdly,

Models	SciBERT		LED Encoder	
	Evidence F1	Answer F1	Evidence F1	Answer F1
CGSN	49.72	33.20	41.55	29.76
w/o Local Graph	47.86	32.80	38.85	28.33
w/o Global Edges	48.13	31.14	39.55	27.86
w/o Global Graph	48.09	31.70	39.54	28.10
w/o Evidence Memory	49.21	32.61	40.78	29.30

Table 4: Ablation study of CGSN with SciBERT/LED Encoder, evaluating on the dev set of Qasper.

Combinations				SciBERT	
N_{seg}	N_{sent}^G	N_p^G	N_d^G	E-F1	A-F1
16	64	32	4	49.72	33.20
8	64	32	4	48.33	32.85
16	48	32	4	48.18	31.95
16	64	16	4	48.66	31.72
16	64	32	2	48.57	33.01
16	64	80	4	48.86	32.13
16	64	32	48	49.08	32.45

Table 5: The Evidence F1 (E-F1)/Answer F1 (A-F1) under different combinations of paragraph number in a segment (N_{seg}), the node number in sentence (N_{sent}^G), paragraph (N_p^G) and document (N_d^G) levels of the global graph, evaluated on the dev set of Qasper.

compared among lines 1,6 and 7, it suggests that it is proper to set the number of sentence nodes larger than that of paragraph nodes, and the same relation maintains between paragraph nodes and document nodes.

5.4 Effects of Graph Hop Number

We conduct the experiment on the number of attention hops on the local graph and on the global graph. The result is shown in Figure 3. We can see an improvement in the performance from 0 to 4 (in the local graph network) and from 0 to 1 (in the global graph network). The experimental results demonstrate the importance of multi-hop attention among local and global nodes in the graph.

5.5 Effects of the Global Graph Network

To better understand the functionality of the global graph, we conduct an experiment as shown in Figure 4. First of all, it illustrates that CGSN outperforms any other ablated versions in each range of “document length” or “maximum evidence distance”. Besides, when the document length and the maximum evidence distance are large, the gap

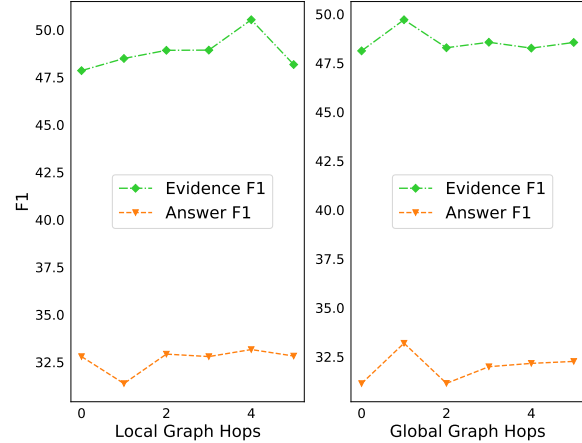


Figure 3: Evidence/Answer F1 scores of the proposed CGSN model with varied local/global GAT hops, evaluated on the dev set of Qasper.

between CGSN and other versions becomes bigger, which demonstrates the ability of the global graph to capture long-range dependency. Furthermore, comparing the two ablated versions, it suggests that the edges of the global graph can enhance the ability of the model on long text understanding.

5.6 Effects of Evidence Memory

Redundancy is an essential problem in LDQA since in the answer generation phase, repetitious paragraphs could make the QA model confused (Appendix D). Therefore, it’s crucial to select important and diverse evidence pieces for the QA model. To this end, we explore the effects of the evidence history module on redundancy reduction.

As shown in Table 6, we evaluate the redundancy of the proposed model CGSN and the model “w/o Mem”. **REP** (Zhong et al., 2019) is used to calculate the repetitions in the generated summary. Based on it, we introduce **REP**_{inter} to evaluate the redundancy among different paragraphs selected:

$$\text{REP}_{\text{inter}} = \frac{\text{CountUniq}(\{\text{ngram}_i\}_{i=1}^N)}{\sum_{i=1}^N \text{Count}(\text{ngram}_i)} \quad (18)$$

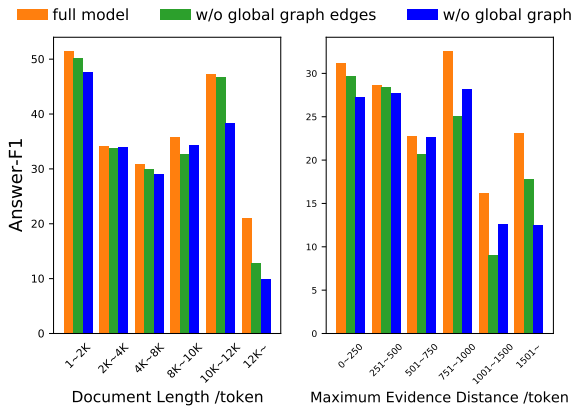


Figure 4: Answer F1 of the proposed CGSN model and two ablated models on the dev set of Qasper. In the left subfigure, we evaluate each model by the context document length in each data instance. On the right side, the dev set is partitioned via the number of tokens between the first and the last gold evidence paragraphs in the context document of each QA pair, called “Maximum Evidence Distance”.

Models	$\mathbf{REP}_{inter} \uparrow$	$\mathbf{REP} \uparrow$	$\mathbf{Score}_{red} \downarrow$
CGSN	0.8751	0.5708	-0.3166
w/o Mem	0.8688	0.5661	-0.3133
Gold	0.9345	0.6301	-0.5092

Table 6: Comparison between the full CGSN model and the model without evidence memory (w/o Mem). \mathbf{REP} (Zhong et al., 2019), \mathbf{Score}_{red} (Chen et al., 2021) and the proposed \mathbf{REP}_{inter} are used to evaluate the redundancy among the selected paragraphs. “ \uparrow ” denotes higher is better while “ \downarrow ” means lower is better. “Mem” denotes the evidence memory. “Gold” denotes the gold evidence.

where N is the number of selected paragraphs, $\text{Count}(\text{ngram}_i)$ is the number of n -grams in the paragraph i , $\text{CountUniq}(\{\text{ngram}_i\}_{i=1}^N)$ is the number of grams in all the selected paragraphs.

According to Table 6, gold evidence obtains the least redundancy in them. Besides, it is found that without the evidence memory module, the redundancy of the proposed model gains. It demonstrates that the evidence memory can alleviate the redundancy problem in the evidence selection phase.

6 Conclusion

To solve the problem of lacking global structure in LDQA methods, we propose Compress Graph Selector Network to capture the global structure over the long document when selecting evidence pieces. Extensive experiments demonstrate the strong per-

formance of the model.

Limitations

Since the proposed model is applied in an iterative way to read the long document, the limitation is the time cost. To alleviate the problem, we pack a series of paragraphs into a segment and encode them in parallel. However, methods better at parallelized computing could further save the required time for training and inference.

Acknowledgements

The work is supported by National Key R&D Plan (No. 2020AAA0106600), National Natural Science Foundation of China (No. U21B2009, 62172039 and L1924068). We thank the EMNLP reviewers for their helpful feedback. We would like to acknowledge Yuming Shang for the helpful discussions.

References

- Sami Abu-El-Haija, Bryan Perozzi, Rami Al-Rfou, and Alexander A Alemi. 2018. [Watch your step: Learning node embeddings via graph attention](#). *Advances in neural information processing systems*, 31.
- Joshua Ainslie, Santiago Ontanon, Chris Alberti, Václav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. [Etc: Encoding long and structured inputs in transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [Scibert: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#). *arXiv preprint arXiv:2004.05150*.
- Wang Chen, Piji Li, and Irwin King. 2021. [A training-free and reference-free summarization evaluation metric via centrality-weighted relevance and self-referenced redundancy](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 404–414.
- Peng Cui and Le Hu. 2021. [Sliding selector network with dynamic memory for extractive summarization](#)

- of long documents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5881–5891.
- Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. [A dataset of information-seeking questions and answers anchored in research papers](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.
- 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.
- Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu. 2020. [Recurrent chunking mechanisms for long-text machine reading comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6751–6761.
- 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.
- Junru Lu, Gabriele Pergola, Lin Gui, Binyang Li, and Yulan He. 2020. [Chime: Cross-passage hierarchical memory network for generative review question answering](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2547–2560.
- Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. [Key-value memory networks for directly reading documents](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409.
- Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. 2019. [Attention gated networks: Learning to leverage salient regions in medical images](#). *Medical image analysis*, 53:197–207.
- Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. [Graph attention networks](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. [Memory networks](#). *CoRR*, abs/1410.3916.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. [Hotpotqa: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. [Graph convolutional networks for text classification](#). In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377.
- Bo Zheng, Haoyang Wen, Yaobo Liang, Nan Duan, Wanxiang Che, Daxin Jiang, Ming Zhou, and Ting Liu. 2020a. [Document modeling with graph attention networks for multi-grained machine reading comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6708–6718.
- Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. 2020b. [Gman: A graph multi-attention network for traffic prediction](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1234–1241.
- Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2019. [Searching for effective neural extractive summarization: What works and what’s next](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1049–1058.
- Yunchang Zhu, Liang Pang, Yanyan Lan, Huawei Shen, and Xueqi Cheng. 2021. [Adaptive information seeking for open-domain question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3615–3626.

Appendix

A Algorithm

We describe the detailed architecture of CGSN in Algorithm 1.

B Analysis of Memory Usage

In this section, we analyze the memory usage of the end-to-end method LED and the proposed CGSN. Assuming the length of a document is L ($L \geq 4K$), the local window size⁹ is W ($W \geq 512$), the number of global tokens is G_t , the memory usage of LED method is $O(L(W + G_t))$. When $G_t \ll W$, the memory usage is $O(LW)$. For CGSN, set the paragraph number in a segment as B . For a fair comparison, the maximum length of a paragraph is $W_{\text{half}} = \lfloor \frac{W}{2} \rfloor$. Therefore, the memory usage in CGSN is $O(BW_{\text{half}}^2 + f_{\text{local}}(B, W) + M_{\text{global}})$,

⁹Attention to the $\frac{W}{2}$ tokens ahead and $\frac{W}{2}$ tokens behind.

Algorithm 1: Compressive Graph Selector Network

Data: LDQA training dataset \mathcal{S} , CGSN model \mathcal{M} , composed of a contextual **Encoder**, a Local Graph Network **LGN**, a Global Graph Network **GGN**

Result: A fine-tuned model \mathcal{M}' .

```
1 Initialize the parameters of  $\mathcal{M}$ ;  
2 Set the evidence memory Mem  $\leftarrow$  None;  
3 Shuffle the dataset  
    $\mathcal{S} \leftarrow \{(d^{(l)}, q^{(l)}, E^{(l)}, a^{(l)})\}_{l=1}^S$ ;  
4 foreach document  $d^{(l)}$  in  $\mathcal{S}$  do  
5   Split  $d^{(l)}$  into segments  $[P_1, P_2, \dots, P_N]$ ,  
   where  $P_l = [p_{l_1}, \dots, p_{l_N}]$ ;  
6   foreach segment  $P_l$  in  $d^{(l)}$  do  
7     Time step  $T \leftarrow 1$ ;  
8     if  $T \neq 0$  then  
9       | write_evidence(Mem, GGN);  
10    end  
11    Segment  $\mathbf{B} \leftarrow \{(q^{(l)}, p_{l_r})\}_{m=1}^{l_N}$ ;  
12     $\mathbf{H} \leftarrow$  Encoder( $\mathbf{B}$ );  
13    local_graph_init(H, LGN);  
14    for  $l \leftarrow 1$  to Local_hops do  
15      | unidirect_hops(LGN);  
16    end  
17    multihead_attention(LGN, GGN);  
18    for  $l \leftarrow 1$  to Global_hops do  
19      | bidirect_hops(GGN);  
20    end  
21    multihead_attention(GGN, LGN);  
22    p_logits, loss  $\leftarrow$  cal_loss(LGN,  
    $E_{P_i}$ );  
23    Mem  $\leftarrow$  weighted_sum(p_logits,  
   LGN);  
24    update_params( $\mathcal{M}$ , loss);  
25  end  
26 end
```

where f_{local} is the function of memory usage in local graph and M_{global} is the memory usage of global graph, which is pre-defined as a constant. It means that memory usage of CGSN is constant to the document length while LED is nearly linear to the length (when W becomes larger, the linearity will convert into quadratic).

C Answer Generator

Following the ‘led_base_smaller_context’ setting¹⁰ in Dasigi et al. (2021), a vanilla LED is used as the answer generator, which is trained with gold evidence coupled with QA pairs in the training set. To

¹⁰<https://github.com/allenai/qasper-led-baseline>

evaluate the performance of an evidence selection method in perspective of answer generation, for each question and its long document. The evidence selection method is firstly used to provide evidence pieces given the question and the long document. And then, the corresponding selected evidence and the question are concatenated and fed into the LED model. After that, the LED model generates answer tokens for the question. The generated answer tokens are then used to calculate the ‘Answer-F1’ score given the gold answer.

D Case Study

we conduct the case study in this section to investigate the effectiveness of the global graph and the evidence memory qualitatively. As shown in Table 7, the proposed model can correctly select all the evidence pieces across different positions of the text and integrate them into the complete evidence.¹¹ However, for the model without the global graph, a piece of evidence is lost. It indicates that in a long document, the understanding of long-range is crucial for evidence selection and answer generation.

Table 8 compares the proposed model and the one without evidence memory, where the ablated model chooses “MEDDOCAN 2019 shared task” twice as well as adds some other noise information (“BERT-based sequence labeling approach”) which makes the model confused in the answer generation phase.

¹¹The original case in the dataset has some errors and we correct them in Table 7.

The proposed full model
<p>Selected Paragraphs: LiLi uses a Knowledge Store (KS) for knowledge retention. KS has four components: (i) Knowledge Graph (INLINEFORM0) (ii) Relation-Entity Matrix (INLINEFORM2) (iii) Task Experience Store (INLINEFORM15) (iv) Incomplete Feature DB (INLINEFORM29) LiLi also uses a stack, called Inference Stack (INLINEFORM0) to hold query</p> <p>Generated Answer: Knowledge Graph, Relation-Entity Matrix, Task Experience Store, Incomplete Feature DB, Stack</p>
Ablated Model (w/o global graph)
<p>Selected Paragraphs:LiLi uses a Knowledge Store (KS) for knowledge retention. KS has four components: (i) Knowledge Graph (INLINEFORM0) (ii) Relation-Entity Matrix (INLINEFORM2) (iii) Task Experience Store (INLINEFORM15) (iv) Incomplete Feature DB (INLINEFORM29)</p> <p>Generated Answer: Knowledge Graph, Relation-Entity Matrix, Task Experience Store, Incomplete Feature DB</p>

Table 7: Comparison between the proposed full model and the model ablated global graph. In this case, the length of the document is 8768, which cannot be processed by most of the pre-trained models in one step. **Matched content** is colored in blue while **mismatch content** is colored in red.

The proposed full model
<p>Selected Paragraphs: Two datasets are exploited in this article NUBes BIBREF4 is a corpus of around 7,000 real medical reports the second set of experiments uses the MEDDOCAN 2019 shared task competition dataset</p> <p>Generated Answer: NUBes BIBREF4, MEDDOCAN 2019 shared task competition dataset</p>
Ablated Model (w/o evidence memory)
<p>Selected Paragraphs: Two datasets are exploited in this article NUBes BIBREF4 is a corpus of around 7,000 real medical reports the second set of experiments uses the MEDDOCAN 2019 shared task competition dataset We have compared this BERT-based sequence labelling against other methods and systems. One of the experiments uses the MEDDOCAN 2019 shared task dataset.....</p> <p>Generated Answer: NUBes BIBREF4, MEDDOCAN 2019 shared task competition dataset</p> <p>Generated Answer: NUBes BIBREF4, BERT-based sequence labelling approach</p>

Table 8: Comparison between the proposed full model and the model without evidence memory. **Matched content** is colored in blue while **redundant content** is colored in red.