Enhanced CGSN System for Machine Reading Comprehension

Liwen Zheng, Haoran Jia, Hongyan Xie, Xi Zhang, and Yuming Shang(✉)

Beijing University of Posts and Telecommunications, Beijing, China
{zhenglw,jiahaoran,zhangx,shangym}@bupt.edu.cn

Abstract. This paper introduces the system proposed by the "Guess Right or Not (Ours)" team for NLPCC 2023 Shared Task 2 (https://github.com/Yottaxx/NLPCC23_SciMRC)--Multi-perspective Scientific Machine Reading Comprehension. This task requires participants to develop a reading comprehension model based on state-of-the-art Natural Language Processing (NLP) and deep learning techniques to extract word sequences or sentences from the given scientific texts as answers to relevant questions. In response to this task, we use a fine-grained contextual encoder to highlight key contextual information in scientific texts that is highly relevant to the question. Besides, based on existing advanced model CGSN [7], we utilize a local graph network and a global graph network to capture global structural information in scientific texts, as well as the evidence memory network to further alleviate the redundancy issues by saving the selected result in the previous steps. Experiments show that our proposed model performs well on datasets released by NLPCC 2023, and our approach ranks 1st for SMRC Task 2 according to the official results.

Keywords: SMRC · Fine-grained Contextual Information · Global Structural Information

1 Introduction

Machine Reading Comprehension (MRC) is a task that involves answering questions about a given context paragraph, which enables machines to read and understand unstructured text. MRC is a rapidly growing field of research due to its potential for various enterprise applications, and it holds the potential to revolutionize the way humans interact with machines. For example, as shown in Fig. 1, search engines equipped with MRC techniques can directly output correct answers to questions rather than a series of related web pages, which can significantly enhance the efficiency of information retrieval.

Early MRC systems primarily relied on rule-based or machine learning techniques, which depended on manually crafted rules or features [1–3]. The drawbacks of these methods lie in their limited ability to comprehend contextual information and their

Supported by the Natural Science Foundation of China (No.61976026), the Fundamental Research Funds for the Central Universities.

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14304, pp. 83–92, 2023.
https://doi.org/10.1007/978-3-031-44699-3_8

reduced generalization capabilities. Therefore, researchers have been dedicated to studying deep learning-based approaches in recent years, and as deep learning continues to evolve, researchers primarily focus on two research paradigms: attention mechanisms-based methods and fine-tuning or optimization performed on pre-trained language models. For example, BiDAF (Bi-Directional Attention Flow) [4] utilizes a bidirectional attention mechanism to capture contextual information at different granularity levels. Gong H et al. [5] employ a pre-trained transformer model, such as BERT [6], to encode the joint contextual information of texts and questions.

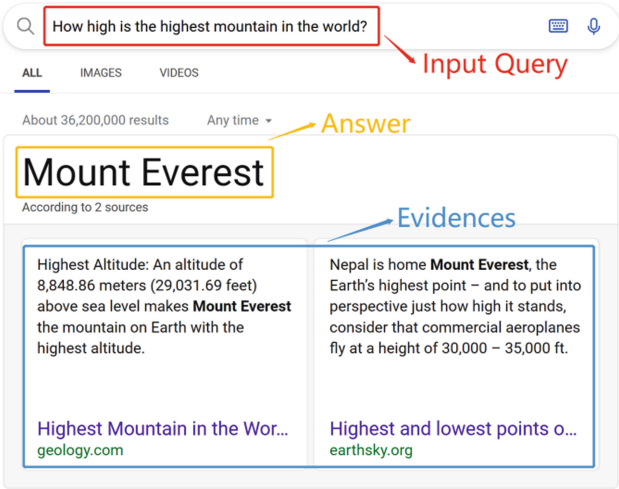

Fig. 1. An example of search engine Bing[1] with MRC.

Most of the current deep learning-based MRC methods still encounter challenges such as information loss during text vectorization and the long-distance semantic dependency. To tackle these issues, we initially focus on the embedding layer and conduct fine-grained context encoding on the input texts to prevent premature loss of essential feature information. Additionally, we leverage gating mechanisms to set up redundant filter, thereby further enhancing the refinement of vector representation. To capture long-distance semantic dependency information in scientific texts and problems effectively, we use the CGSN model [7], which employ local graph network and global graph network to capture both local semantic information and global contextual information separately, and establish long-range reasoning through iterative processes. Furthermore, the evidence memory network is utilized to store the selection results from previous steps and mitigate redundancy issues effectively.

In summary, our contributions are as follows:

- Following CGSN, we utilize local graph network and global graph network to enhance global structural contextual information.

[1] https://cn.bing.com/

- Based on CGSN, we propose a Fine-grained Contextual Encoder to highlight the most relevant features and eliminate redundant information.
- Our model achieves the first place in NLPCC 2023 shared task 2 on Scientific Machine Reading Comprehension (SMRC).

2 Related Work

Machine reading comprehension can be roughly categorized into four types: cloze tests, multiple choice, span extraction, and free answering [8]. This paper primarily focuses on the span extraction tasks. Due to limitations in dataset size, Traditional rule-based and machine learning-based methods exhibit poor performance and are impractical for deployment in practical applications. In recent years, researchers have discovered that methods based on deep neural networks excel in extracting contextual information, which results in significantly improved model performance compared to traditional methods [9].

MRC models based on deep learning techniques typically contain four steps [8]: embedding layer, feature extraction, question-text interaction, and answer prediction. Firstly, it is essential to convert the input natural language into vector representation. Match LSTM model [10] utilized word vectors to encode the input text. DocQA [11] and MPMRC [12] employed a combination of word embedding and character embedding techniques to extract the text representation vector. To capture contextual information, Zhang W et al. [13] integrated the dynamic text representation model ELMO to derive more precise text vectors. The encoding layer is primarily employed to extract key features that are highly relevant to questions from the input texts. Recurrent neural networks (RNNs) and variants are extensively utilized in machine reading comprehension tasks. To comprehensively incorporate both forward and backward information, bi-directional RNN networks are commonly employed in MRC [14, 16]. KAR model [15] utilized Bidirectional Long Short-Term Memory (BiLSTM) to extract contextual features. The interaction between questions and texts primarily relies on attention mechanisms to capture correlations and key features. To address the challenges posed by long-distance semantic dependencies, current methods predominantly employ self-attention mechanisms [19, 20]. R-Net model [17] introduced a gated self-attention mechanism to capture internal connections within texts, and Fastform model [18] utilized a multi-head self-attention mechanism for interaction. And finally, we aggregate information from all modules, make predictions, and output the final answer.

The adventure of pre-trained language models, such as BERT and XLnet [26], has revolutionized the intricate architecture of MRC models. It is now possible to achieve excellent results by solely fine-tuning these pre-trained language models. RoBERTa [21], ALBERT [22], and other models are all enhancements built upon BERT, which demonstrate remarkable performance in natural language processing tasks. Furthermore, methods based on pre-trained models have emerged as mainstream solutions for MRC tasks [23]. The length limitation of input texts in pre-trained models poses challenges for addressing long document reading comprehension tasks. Gong H et al. [5] employed reinforcement learning to enable the model to learn and determine the input length. They also utilized a loop mechanism to capture dynamic semantic information. Ding M et al. [24] drew inspiration from human cognitive processes and employed similar

mechanisms to process long texts. Zhao J et al. [25] proposed a read-over-read method to alleviate the challenges associated with length limitations.

3 The Proposed Approach

3.1 Task Definition

Following CGSN, we take a question $\boldsymbol{Q} = \left[q_1, q_2, ..., q_m\right]$ along with scientific texts $\boldsymbol{P} = \left[\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_n\right]$ as input of MRC model, and the goal of our task to extract a free-form answer $\boldsymbol{A} = [a_1, a_2, ..., a_l]$ for the input question from texts $\boldsymbol{P}$, where m and n denotes the length of question and the number of paragraphs separately, $\boldsymbol{p}_i = \left[w_i^1, w_i^2, ..., w_i^{k_i}\right](1 \leq i \leq n)$ denotes paragraph i with the word length of k_i.

3.2 Model Structure

As shown in Fig. 2, our model is composed of four modules: fine-grained contextual encoder, local semantic extractor, global semantic extractor and memory network. Firstly, pre-trained Transformer encoder SciBERT [27] are utilized to encode the question-paragraph pair, and fine-grained contextual encoding is conducted to further refine and capture detailed information. Subsequently, local graph network is constructed at the token, sentence, paragraph, and texts levels. The information obtained at each granularity is then fed into the subsequent module to form the global graph network. The global information will feedback to enhance the local representation. Finally, at each time step, the memory network receives the enhanced local representation and the predicted logits, and updates the global graph paragraph nodes for the next step.

Fine-grained Contextual Encoder. We set the initial embedding of each question and paragraph as $\boldsymbol{E}_q$ and $\boldsymbol{E}_p$, and utilize 2 layers of BiLSTM to obtain the fine-grained contextual information $\boldsymbol{H}_q$ and $\boldsymbol{H}_p$.

$$\boldsymbol{H}_q = BiLSTM\left(\boldsymbol{E}_{\boldsymbol{q}}\right) \tag{1}$$

$$\boldsymbol{H}_p = BiLSTM\left(\boldsymbol{E}_{\boldsymbol{p}}\right) \tag{2}$$

The gating mechanism is employed to eliminate redundant information and capture semantic information that is highly relevant to the question.

$$\boldsymbol{H}_p^q = Gate[\boldsymbol{H}_q, \boldsymbol{H}_p] \tag{3}$$

where $\boldsymbol{H}_p^q$ denotes the question-aware paragraph embedding.

Local Semantic Extractor. The local graph network is constructed by sub-graphs at four different granularity: token, sentence, paragraph, and texts. To initial the node representation, the question-aware paragraph embedding $\boldsymbol{H}_p^q$ will be fed into the Local Semantic Extractor. Token-level nodes $\boldsymbol{h}_t^{Local}$ are initialized by the corresponding token representation of $\boldsymbol{H}_p^q$, and vector of sentence-level node $\boldsymbol{h}_s^{Local}$ can be calculated by the

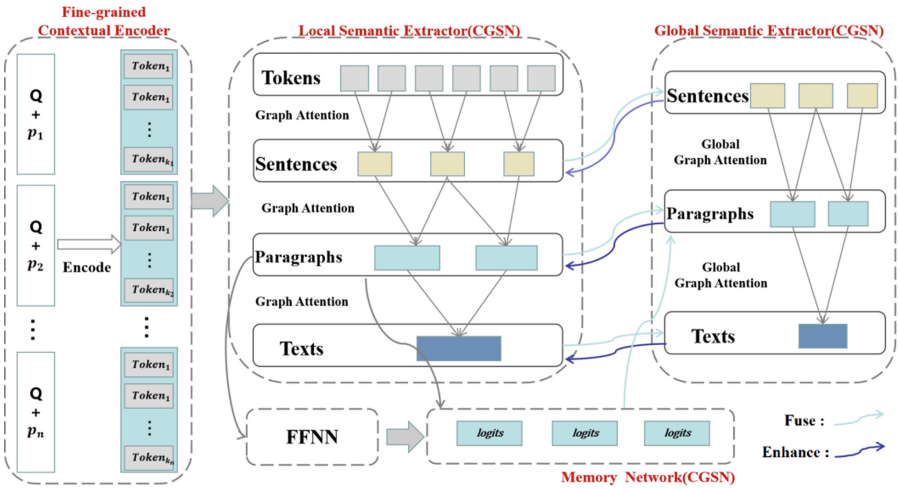

Fig. 2. The architecture of our proposed model. The Fine-grained Contextual Encoder is designed to highlight the most relevant features and eliminate redundant information, and the Local Semantic Extractor, Global Semantic Extractor and Memory Network proposed by CGSN are utilized to enhance global structural contextual information.

mean-pooling of $\boldsymbol{h}_t^{Local}$. [CLS] of $\boldsymbol{H}_p^q$ can represent paragraph-level node $\boldsymbol{h}_p^{Local}$, and texts-level nodes $\boldsymbol{h}_{text}^{Local}$ are initialized by the mean-pooling of $\boldsymbol{h}_p^{Local}$.

The information propagation between sub-graphs in the unidirectional local graph network can only be implemented from low-level to high-level. By executing graph attention [28] sequentially between adjacent sub-graphs, it can finally capture fine-grained semantic information and local structural information in the input scientific texts. Taking sentence-level and paragraph-level sub-graphs as an example, the mathematical process of updating paragraph-level node $\boldsymbol{h}_p^o$ with sentence-level node $\boldsymbol{h}_s^o$ at time step o is as follows:

$$e_{sp} = \frac{\left(\boldsymbol{h}_p^o \boldsymbol{W}^Q\right)\left(\boldsymbol{h}_s^o \boldsymbol{W}^K\right)^T}{\sqrt{d_z}} \tag{4}$$

$$\alpha_{sp} = softmax_s\left(e_{sp}\right) = \frac{exp\left(e_{sp}\right)}{\sum_{i \in N_S} exp\left(e_{ip}\right)} \tag{5}$$

$$\boldsymbol{z}_p^{head_x} = \sum_{s \in N_S} \alpha_{sp} \boldsymbol{h}_s^o \boldsymbol{W}^V \tag{6}$$

$$\boldsymbol{h}_p^{o+1} = Cat\left[\boldsymbol{z}_p^{head_1}, \boldsymbol{z}_p^{head_2}, ..., \boldsymbol{z}_p^{head_k}\right] \tag{7}$$

where e_{sp} denotes the attention coefficients between $\boldsymbol{h}_p^o$ and $\boldsymbol{h}_s^o$, α_{sp} is the normalization of e_{sp}, $\boldsymbol{z}_p^{head_x}$ denotes the output of the multi-head attention, $\boldsymbol{W}^Q$, $\boldsymbol{W}^K$, and $\boldsymbol{W}^V$ are the parameters of the query, key and value of attention mechanism, and $\boldsymbol{h}_p^{o+1}$ is the representation of paragraph-level node at time step $o + 1$, which is composed of the concatenation of the multi-head outputs.

Representation of sentence, paragraph and texts-level nodes will be updated through the method mentioned above.

Global Semantic Extractor. Local sentence, paragraph and texts-level nodes are delivered into the global graph network through the similar multi-head attention mechanism in **Local Semantic Extractor**, and form the local-aware global nodes h_{local}^{global}. To fuse features from local and global nodes, a feed-forward neural network (FFNN) and a gated network are employed.

$$z_f^{global} = FFNN\left(\boldsymbol{h}^{global}, h_{local}^{global}\right) \tag{8}$$

$$\gamma = Gate\left(z_f^{global}\right) \tag{9}$$

$$h_f^{global} = (1-\gamma)\boldsymbol{h}^{global} + \gamma z_f^{global} \tag{10}$$

where $\boldsymbol{h}^{global}$ denotes the original global node representation, h_f^{global} denotes the updated global node representation.

To further extract global structure information and interaction information sufficiently, we employ cross attention mechanism between adjacent sub-graphs for m times.

Besides, to integrate the local and global information for more precise prediction, the global nodes are fed back to the local graph network, and multi-head attention mechanism is employed to obtain enhanced local graph nodes. We define L as the extraction loss:

$$L = -\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{N_i}[logP(y|\boldsymbol{h}_p^{L\leftarrow G})] \tag{11}$$

where $\boldsymbol{h}_p^{L\leftarrow G}$ represents the enhanced local node representation, and y denotes the predicted label of paragraph.

Memory Network. The enhanced local nodes and the prediction logits at time step o will be cached and utilized at time step $o+1$ to mitigating the adverse effects of redundant information. Specifically, using prediction logits as the importance weights for paragraphs and performing feature fusion based on these weights can highlight important information while concealing redundant information.

4 Experiments

4.1 Dataset and Metric

We use the dataset released by NLPCC 2023 Shared task 2 to train and evaluate our MRC model, and we extract a portion of the training dataset to form a validation dataset. Table 1 shows the number of QA(question-answer) pairs in the given training and testing datasets. We use the "Free_form_answer" field as final answer and take F1 value as the evaluation metric following the official task guidelines.

Table 1. Size of QA pairs for the official datasets.

	# of Texts	# of QA pairs
Train	372	3278
Dev	120	1595
Test	147	1169

4.2 Experiment Settings

Our model is implemented based on PyTorch[2] and the hugging-face[3] framework. We use several pre-trained models, such as SciBERT[4], LED Encoder and BERT[5], and selected the pre-trained Transformer encoder SciBERT as the final initial encoder. To determine the optimal parameter settings, we conducted multiple sets of experiments with different batch size, epoch, learning rate, weight decay and warm-up proportion.

Table 2. Hyper-parameter setting.

Hyper-parameter	Value
Epoch num	5
Batch size	4
Optimizer	AdamW[31]
Learning rate	1e-5
Weight decay	0.01
Warm-up proportion	0.1
Max_token_len	256
Max_sentence_len	32
Max_paragraph_len	1
Local hop	4
Global hop	1

To limit the number of nodes in token-level, sentence-level, and paragraph-level subgraphs, we set the maximum number of nodes to 256, 32, and 1 respectively based on experimental results. Besides, following CGSN, we implied multi-hop graph attention in local and global graph network to capture global structural information and contextual information, and conducted experiments to determine the number of local hop and global hop. Some final training hyper-parameters are shown in Table 2.

[2] https://pytorch.org
[3] https://github.com/huggingface/transformers
[4] https://huggingface.co/allenai/scibert_scivocab_uncased
[5] https://huggingface.co/bert-base-uncased.

4.3 Baselines

To demonstrate the effectiveness of our model, we compare it with several existing methods, the specific description is listed as follows:

- LED [29]: LED is a method designed for long document question answering tasks, which utilize the pre-trained LED model [30] as answer generator. With the question and corresponding pre-selected evidences as input, the model trained with gold evidences and question-answer pairs will output the predicted answer for the input question.
- LED Encoder [30]: The large-scale pre-trained language models has achieved tremendous success in the field of natural language processing. we choose LED Encoder as backbone, which serves as the pre-trained encoder for the above LED model, and its involvement contributes to the final outstanding performance of LED on MRC.

4.4 Results and Analysis

Table 3 shows the top-4 official result of NLPCC 2023 Task 2. Our team "Guess Right or Not (Ours)" obtained a final evaluation score of 0.5459, and get the 1st place among 16 participating teams.

Table 3. Top-4 official result of NLPCC 2023 Task 2.

Team ID	System Name	Final Evaluation Score
1	Ours	0.5459
2	IMUNLP	0.4519
3	PIE	0.4181
4	OUC_NLP	0.3574

Table 4 shows the experimental results conducted on the Qasper [29] dataset. Comparison result with several typical baselines can validate the superiority of our method. As is shown in Table 4, our method achieves the best performance, which demonstrates the effectiveness of our solution to MRC task. Specifically, the fine-grained contextual information is highly important for understanding the semantic information of long texts, and the involvement of global structural information is also crucial in machine reading comprehension tasks. Moreover, evidence memory network also contributes to the improvement of the model performance.

Table 4. The experimental results of different models on Qasper dataset.

Model	F1 value
LED	51.50
LED Encoder	53.99
Our Method	54.37

5 Conclusion

In this paper, we design an MRC model based on the existing method CGSN, which employ a local graph network and a global graph network to capture local and global structural information in scientific texts. Besides, we propose a Fine-grained Contextual Encoder to highlight features relevant to questions and eliminate redundant information. Furthermore, we utilize various optimization strategies to optimize and improve the base model CGSN to achieve optimal performance on the NLPCC dataset. And according to the Official results of NLPCC 2023 shared task s on Scientific Machine Reading Comprehension make known that our solution takes the first place among all participants, which demonstrates the effectiveness of our solution to MRC task. However, there is still a long way for machine reading comprehension tasks, and how to extract more precise and interpretable answers remains an ongoing challenge that requires continuous and in-depth exploration.

References

1. Poon, H., Christensen, J., Domingos, P., et al.: Machine reading at the University of Washington. In: Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading, pp. 87–95 (2010)
2. Hirschman, L., Light, M., Breck, E., et al.: Deep read: a reading comprehension system. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 325–332 (1999)
3. Riloff E, Thelen M.: A rule-based question answering system for reading comprehension testsIn. In: ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-based Language Understanding Systems (2000)
4. Seo, M., Kembhavi, A., Farhadi, A., et al.: Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 (2016)
5. Gong, H., Shen, Y., Yu, D., et al.: Recurrent chunking mechanisms for long-text machine reading comprehension. arXiv preprint arXiv:2005.08056 (2020)
6. Devlin, J., Chang, M.W., Lee, K., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Nie, Y., Huang, H., Wei, W., et al.: Capturing global structural information in long document question answering with compressive graph selector network. arXiv preprint arXiv:2210.05499 (2022)
8. Liu, S., Zhang, X., Zhang, S., et al.: Neural machine reading comprehension: methods and trends. Appl. Sci. **9**(18), 3698 (2019)
9. Gu, Y., Gui, X., Li, D., Shen, Y., Liao, D.: A review of machine reading comprehension based on neural networks. J. Softw. **31**(07), 2095–2126 (2020)

10. Wang S, Jiang J.: Machine comprehension using match-LSTM and answer pointer. arXiv preprint arXiv:1608.07905 (2016)
11. Clark C, Gardner M.: Simple and effective multi-paragraph reading comprehension. arXiv preprint arXiv:1710.10723 (2017)
12. Wang, Y., Liu, K., Liu, J., et al.: Multi-passage machine reading comprehension with cross-passage answer verification. arXiv preprint arXiv:1805.02220 (2018)
13. Zhang, W., Ren, F.: ELMo+ gated self-attention network based on BiDAF for machine reading comprehension. In: 2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS), pp. 1–6 (2020)
14. Lee, H., Kim, H.: GF-Net: Improving machine reading comprehension with feature gates. Pattern Recogn. Lett. **129**, 8–15 (2020)
15. Wang, C., Jiang, H.: Explicit utilization of general knowledge in machine reading comprehension. arXiv preprint arXiv:1809.03449 (2018)
16. Ma, X., Zhang, J.: GSA-Net: gated scaled dot-product attention based neural network for reading comprehension. Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije, **61**(4), 643–650 (2020)
17. Wang, W., Yang, N., Wei, F., et al.: Gated self-matching networks for reading comprehension and question answering. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 189–198 (2017)
18. Wu, C., Wu, F., Qi, T., et al.: Fastformer: additive attention can be all you need. arXiv preprint arXiv:2108.09084 (2021)
19. Shen, T., Zhou, T., Long, G., et al.: DiSAN: directional self-attention network for RNN/CNN-free language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, issue 1 (2018)
20. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
21. Liu, Y., Ott, M., Goyal, N., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
22. Lan, Z., Chen, M., Goodman, S., et al.: ALBERT: a lite BERT for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
23. Su, C., Fukumoto, F., Huang, X., et al.: DeepMet: a reading comprehension paradigm for token-level metaphor detection. In: Proceedings of the Second Workshop on Figurative Language Processing, pp. 30–39 (2020)
24. Zhao J, Bao J, Wang Y, et al.: RoR: Read-over-read for long document machine reading comprehension. arXiv preprint arXiv:2109.04780 (2021)
25. Ding, M., Zhou, C., Yang, H., et al.: Cogltx: applying BERT to long texts. Adv. Neural. Inf. Process. Syst. **33**, 12792–12804 (2020)
26. Yang, Z., Dai, Z., Yang, Y., et al.: XLNet: generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems (2019)
27. Beltagy, I., Lo, K., Cohan, A.: SciBERT: a pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019)
28. Veličković, P., Cucurull, G., Casanova, A., et al.: Graph attention networks. arXiv preprint arXiv:1710.10903 (2017)
29. Dasigi, P., Lo, K., Beltagy, I., et al.: A dataset of information-seeking questions and answers anchored in research papers. arXiv preprint arXiv:2105.03011 (2021)
30. Ainslie, J., Ontanon, S., Alberti, C., et al.: ETC: encoding long and structured inputs in transformers. arXiv preprint arXiv:2004.08483 (2020)
31. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)