A QUESTION-ORIENTED PROPAGATION NETWORK FOR NEWS READING COMPREHENSION

Liang Wen[1,2,*] *Houfeng Wang*[1] *Dehong Ma*[3] *Jun Fan*[3] *Yingwei Luo*[1,2]
Xiaolin Wang[1,2] *Daiting Shi*[3] *Zhicong Cheng*[3] *Dawei Yin*[3]

[1]School of Computer Science, Peking University, China
[2]Peng Cheng Laboratory, Shenzhen, China [3]Baidu Inc., Beijing, China

ABSTRACT

Machine reading comprehension of news articles remains to be a challenging task since the lengths of its context documents are long. Such reading comprehension task usually requires document-level language understanding while state-of-the-art, pretrained question answering models can only encode sequences with a predefined length limit. In this paper, we propose a novel Question-Oriented Propagation Network (QOPN) model for such task. Specifically, our proposed QOPN first uses a context encoding module to find local question-related clues. Then, it employs a multi-step reasoning module to aggregate question-focused information for iterative reasoning. The novel design put emphasis on capturing question-related information and allow long-range information integration, which is especially beneficial for long-context reading comprehension task. Experiments on two challenging machine comprehension datasets show that the proposed QOPN significantly outperforms previous state-of-the-art models.

Index Terms— Long-Context Question Answering, Machine Reading Comprehension, Question-Oriented Propagation Mechanism

1. INTRODUCTION

Machine reading comprehension (MRC) aims to teach machines to answer questions based on a given context. As a key technology of natural language understanding, it has recently received increasing attention from both academic and industry field. Over the last few years, with the availability of large-scale, high-quality datasets such as SQuAD [1] and pretrained language models like BERT [2], remarkable improvements have been made. However, reading comprehension of news articles still remains to be a challenging real-world task mainly due to the reasons that 1) news articles usually are long while the maximum input length of state of the art question answering (QA) models such as BERT [2] and RoBERTa [3] is limited to 512 owing to their memory and computational requirements. 2) To answer a question related to a given news article, one need to synthesize information across different parts of an article [4, 5].

A straightforward way for long-context MRC tasks is to use a sliding window mechanism [5, 6, 7, 8]. This approach first chunks a long document into small ones with a sliding window and then processes each one individually. However, limited by window size, the method can not model long-range attention interaction, which is especially beneficial for long-context question answering (QA) tasks that require document-level language understanding [5, 9]. Another line of research [9, 10] adopts a coarse-to-fine paradigm. For example, Ding et al. [9] first adopt a bert-based module to extract key sentences from long context, and then use another module to reason over the concatenation of key sentences. Though their approach could gather sentences across different parts of long document for reasoning, they still suffer from the lack of long-range attention when judging which sentence is important due to the length limit of BERT (usually 512 tokens). Besides, such method is not suitable for QA tasks that contain long answers that span multiple sentences, like NLQuAD [5]. In addition to the above methods, several studies [11, 12, 13, 14] apply sparse attention mechanism to make each token attend to partial input tokens which are specified by hand-designed attention patterns. However, due to their dependency on pre-defined hand-designed patterns, such models are designed to build attention based on pre-selected positions. Thus, this type of method is still not sufficient to fully capture long-range dependencies.

Intuitively, when humans want to find an answer from a given long document, i.e., a news article, they tend to 1) read the long article segment by segment, and only focus on question aspects and measures to which extent question aspects are covered by each segment. 2) based on the question-focused impression from all the segments, re-considerate the question and the question-related context to decide the answer. Inspired by human's thinking process, we propose a Question-Oriented Propagation Network (QOPN) model. Our proposed method marries both coarse-to-fine and sparse attention-based methods to achieve effective long-context

*Contribution during internship at Baidu.

question answering. Specifically, QOPN mainly consists of two types of components. The first one (noted as context encoding module) applies RoBERTa [3] to discover local question-related clues segment by segment. The other one (noted as multi-step reasoning module) adopts a novel question-oriented propagation mechanism to simulate human reasoning, which targets at synthesizing question-focused information across different parts of an article and performing reasoning on them. Compared with previous coarse-to-fine methods, our method could take the whole article into consideration and jointly learn to find question-related clues and make inference over them implicitly. Compared with previously discussed sparse attention-based methods, our method does not rely on hand-designed patterns and directly aims at question-focused information. Experimental results on two challenging MRC datasets, NewsQA [4] and NLQuAD [5], show that our proposed method significantly outperforms all the previous state-of-the-art methods.

2. MODEL

As shown in Fig. 1, the QOPN consists of 3 modules: Context Encoding, Multi-step Reasoning and Answer Prediction.

2.1. Context Encoding

The Context Encoding Module is designed to capture question-related information from each segment, which is similar to human reading through the whole article segment by segment to measure how well all question aspects are covered by each segment. Specifically, given a question $Q = \{q_0, q_1, \ldots, q_{M-1}\}$ and a article $P = \{p_0, p_1, \ldots, p_{N-1}\}$, we first split the article text into small non-overlapping K segments so that the length of the concatenation of question and each segment is less than or equal to 512. Then, to collect question-focused information from the k-th segment $S^k = \{s_0^k, s_1^k, \ldots, s_{L-1}^k\}$, we concatenate the question Q and segment S^k and feed it into pre-trained RoBERTa as:

$$\mathbf{H}^k = \text{RoBERTa}\left([CLS], Q, [SEP], S^k, [SEP]\right) \quad (1)$$

where $\mathbf{H}^k \in \mathbb{R}^{d\times(M+L+3)}$ is the last layer's output of RoBERTa. To keep all question-related clues from segment S^k for next-step reasoning, we use the first $M+1$ columns of $\mathbf{H}^k$, namely $\mathbf{H}^k_{0:M} \in \mathbb{R}^{d\times(M+1)}$ to represent the matching of question aspects by segment S^k.

2.2. Multi-step Reasoning

After finding out all question-related clues, humans tend to synthesize all question-focused information distributed across multiple segments and re-considerate the question and the supporting evidences to decide the answer. If unsure they may repeat the above process. Inspired by such human

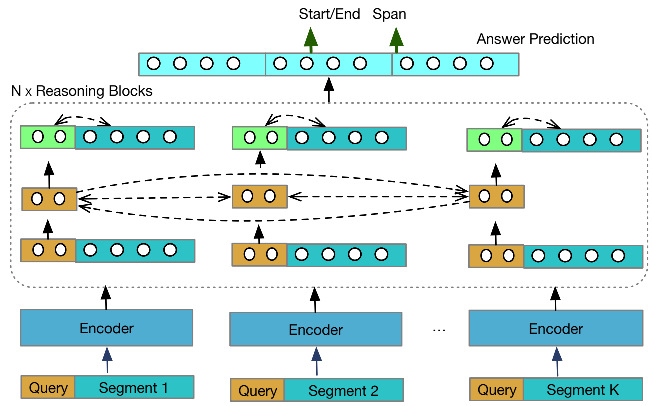

Fig. 1. The architecture of the proposed QOPN model.

experience, we propose a novel question-oriented propagation mechanism to simulate this procedure. In particular, the question-oriented propagation mechanism is implemented by a multi-step reasoning module, which consists of a stack of reasoning block. And the reasoning block consists of the following three units: Question-Oriented Information Interaction (QOII), Gate-Based Information Fusion (GBIF) and Question-Guided Information Propagation (QGIP).

QOII The QOII unit is used to take a comprehensive consideration of all question-focused information. And we adopt a token-wise multi-head self-attention mechanism [15] to achieve it. To be specific, for each question token q_i, we take all its hidden representations as inputs and update them as:

$$\begin{pmatrix} \mathbf{Q_i} \\ \mathbf{K_i} \\ \mathbf{V_i} \end{pmatrix} = \begin{pmatrix} \mathbf{W_q} \\ \mathbf{W_k} \\ \mathbf{W_v} \end{pmatrix} \mathbf{R_i} + \begin{pmatrix} \mathbf{b_q} \\ \mathbf{b_k} \\ \mathbf{b_v} \end{pmatrix} \quad (2)$$

$$\hat{\mathbf{R}}_\mathbf{i} = \text{softmax}\left(\frac{\mathbf{Q_i}\mathbf{K_i}^T}{\sqrt{\lambda}}\right)\mathbf{V_i} \quad (3)$$

where $\mathbf{R_i} = \left[\mathbf{H}_i^0, \mathbf{H}_i^1, \ldots, \mathbf{H}_i^{K-1}\right] \in \mathbb{R}^{d\times K}$, $0 \le i \le M$, $\mathbf{H}_i^k$ is the corresponding representation of the i-th query token from segment S^k and $[\cdot,\cdot]$ denotes the concatenation operation along the row, λ is the scaling factor, and $\mathbf{W_q}$, $\mathbf{W_k}$, $\mathbf{W_v}$, $\mathbf{b_q}$, $\mathbf{b_k}$ and $\mathbf{b_v}$ are learnable parameters.

GBIF The GBIF unit is built upon the outputs of the QOII unit. It is devised to combine the local question-focused information representations and the corresponding global attention vectors. In particular, for each question token q_i, we first adopt a non-linear transformation to fuse the local representations $\mathbf{R_i}$ and the corresponding global attention vectors $\hat{\mathbf{R}}_\mathbf{i} \in \mathbb{R}^{d\times K}$ as follow:

$$\mathbf{F} = \tanh\left(\mathbf{W_f}[\mathbf{R_i}; \hat{\mathbf{R}}_\mathbf{i}; \mathbf{R_i} \circ \hat{\mathbf{R}}_\mathbf{i}; \mathbf{R_i} - \hat{\mathbf{R}}_\mathbf{i}] + \mathbf{b_f}\right) \quad (4)$$

where $\circ$ denotes the element-wise product, $[\cdot;\cdot]$ denotes the concatenation operation along the column, $\mathbf{W_f}$, $\mathbf{b_f}$ are trainable parameters. And the output dimension is projected back to the same size as the original representation $\mathbf{R}$ via the projected matrix $\mathbf{W_f}$.

Then, we use a gating mechanism to selectively incorporate the fusion representations $\mathbf{F} \in \mathbb{R}^{d \times K}$ with the original question-focused information representations $\mathbf{R_i}$ as:

$$\mathbf{G} = \sigma\left(\mathbf{W_g}[\mathbf{R_i}; \hat{\mathbf{R}}_\mathbf{i}; \mathbf{R_i} \circ \hat{\mathbf{R}}_\mathbf{i}; \mathbf{R_i} - \hat{\mathbf{R}}_\mathbf{i}] + \mathbf{b_g}\right) \quad (5)$$

$$\tilde{\mathbf{R}}_\mathbf{i} = \mathbf{G} \circ \mathbf{F} + (1 - \mathbf{G}) \circ \mathbf{R_i} \quad (6)$$

where σ is sigmoid function, $\mathbf{W_g}$, $\mathbf{b_g}$ are trainable parameters, and $\tilde{\mathbf{R}}_\mathbf{i} = \left[\tilde{\mathbf{H}}_i^0, \tilde{\mathbf{H}}_i^1, \ldots, \tilde{\mathbf{H}}_i^{K-1}\right] \in \mathbb{R}^{d \times K}$ denotes the gated fusion representations of question token q_i.

QGIP The QGIP unit aims at spreading the gated fusion representations to the corresponding question-aware segment representations. Specially, for segment S^k, we first concatenate the gated fusion representations of the whole question $\mathbf{X}^k = \left[\tilde{\mathbf{H}}_0^k, \tilde{\mathbf{H}}_1^k, \ldots, \tilde{\mathbf{H}}_M^k\right] \in \mathbb{R}^{d \times (M+1)}$ with the corresponding context encoding $\mathbf{H}^k_{M+1:M+L+2}$. Then, we employ a multi-head self-attention operation over it (as in Eq. (2) and (3)) to obtain the representations $\mathbf{Y}^k = \left[\overline{\mathbf{H}}_0^k, \overline{\mathbf{H}}_1^k, \ldots, \overline{\mathbf{H}}_{M+L+2}^k\right]$ of all tokens from segment S^k.

2.3. Answer Prediction

In this module, we first gain the representations of the article, $\mathbf{Z} = \left[\mathbf{Y}^0_{M+1:M+L+2}, \mathbf{Y}^1_{M+1:M+L+2}, \ldots, \mathbf{Y}^{K-1}_{M+1:M+L+2}\right]$. Then, we adopt the strategy of [16] to decompose the answer span prediction into predicting the start and end positions of the answer span:

$$\mathbf{p}^s = \mathrm{softmax}\left(\mathbf{w}_s \mathbf{Z}\right), \quad \mathbf{p}^e = \mathrm{softmax}\left(\mathbf{w}_e \mathbf{Z}\right) \quad (7)$$

Finally, the training loss function is defined as the negative sum of the log probabilities of the predicted distributions indexed by true start and end indices:

$$\mathcal{L}(\theta) = -\frac{1}{n} \sum_i^n \log\left(\mathbf{p}^s_{y_i^s}\right) + \log\left(\mathbf{p}^e_{y_i^e}\right) \quad (8)$$

where y_i^s and y_i^e are respectively the gold starting and ending position of example i, n is the number of examples, and θ contains all the trainable weights.

3. EXPERIMENTS

3.1. Experimental Settings

Datasets & Evaluation Metrics We evaluate our method on two news MRC dataset, namely NewsQA [4] and NLQuAD [5]. Both of them require document-level language understanding and a significant proportion of questions of them cannot be solved without reasoning. The statistics of the two datasets are summarized in Table 1. From it, we can see that

Datasets	#Samples	#Doc.	Avg #Words
NewsQA	100k	13k	616
NLQuAD	31k	13k	877

Table 1. Statistics of MRC datasets. #Samples and #Doc. are the number of samples and documents respectively. Avg #Words denotes the average number of words per document.

Model	EM(%)	F1(%)
FastQAExt [18]	42.8	56.1
AMANDA [19]	48.4	63.7
MINIMAL [20]	50.1	63.2
DECAPROP [21]	53.1	66.3
RoBERTa-large [3] (sliding window)	49.6	66.3
CogLTX [9]	55.2	70.1
QOPN (RoBERTa-base)	**61.2**	**75.1**
QOPN	**65.5**	**79.8**

Table 2. Performance on the NewsQA test set.

both datasets have a long average document length by words. Note that the original word is usually split into several smaller subwords (tokens) by WordPiece tokenizer. Hence, the average sample length of NewsQA and NLQuAD make them challenging for the state of the art QA models such as BERT [2] and RoBERTa [3] which can only encode sequences with a maximum length of 512 tokens (subwords).

We use EM [1], F1 [1] and IoU [5] as evaluation metrics. Here, EM determines if the prediction exactly matches the target. F1 measures the overlap between the words in the prediction and the target. IoU measures position-sensitive overlap between the predicted and the target answer spans.

Implementation details Due to the limited computational resources, we use the same hyperparameter settings across all variants of QOPN and datasets if not specified. We use Adam [17] optimizer for training. And we set the number of epochs to 6 and warm-up proportion to 10%, the learning rate to 3e-5. The training batch size for NewsQA is set to 32 and the training batch size for NLQuAD is set to 12. We utilize RoBERTa-large as the backbone encoder. The maximum number of segments is 6. The number of reasoning blocks is tuned amongst {1, 2, 3, 4, 5}.

Baselines On NewsQA, we compare our model QOPN with FastQAExt [18], AMANDA [19], MINIMAL [20], DECAPROP [21], RoBERTa-large [3], CogLTX [9]. Since previous state-of-the-art model (CogLTX) adopts RoBERTa-base as its building block, we also report our model that uses the base version of RoBERTa as its backbone. On NLQuAD, we compare our model with several strong baselines, including: BERT [2], RoBERTa [3], Longformer [12], where Longformer is the previous state-of-the-art model.

Model	EM(%)	F1(%)	IoU(%)
BERT-base [2]	25.0	64.0	53.8
BERT-large [2]	30.3	67.9	58.4
RoBERTa-base [3]	29.1	67.2	57.7
RoBERTa-large [3]	33.4	71.1	62.4
Longformer [12]	50.3	81.4	73.6
QOPN	**54.0**	**82.9**	**75.8**

Table 3. Performance on the NLQuAD test set.

Model	EM	F1	IoU
QOPN(full model)	**54.0**	**82.9**	**75.8**
- Question-oriented interaction	47.0	79.4	71.2
- Gate-based fusion	52.8	82.3	75.1
- Question-guided propagation	51.2	80.9	73.3

Table 4. Ablation studies of QOPN on the NLQuAD dataset.

3.2. Experimental Results

Experimental results for NewsQA and NLQuAD are reported in Table 2 and Table 3 respectively. Our model QOPN consistently outperforms the previous models on the two datasets by a large margin. On NewsQA, our model QOPN (RoBERTa-base) outperforms previous state-of-the-art model CogLTX that also adopts RoBERTa-base as its backbone model by 6.0% EM. When using RoBERTa-large as its backbone, our model QOPN further pushes the state of the art to 65.5% EM and 79.8% F1. On NLQuAD, our method also makes remarkable performance improvements. Our model QOPN not only significantly outperforms RoBERTa-large that adopts a sliding window approach [5], but also surpasses Longformer [12] that requires additional expensive pretraining process by 3.7% EM. We infer that our method could focus on question-related contexts and aggregate them for reasoning in a multi-step mode while others cannot.

3.3. Analysis

Ablation studies To isolate individual components' contributions, we run ablation studies on NLQuAD dataset. Table 4 shows the performance of QOPN and several variants of it. From Table 4, we can see that, after only using the [CLS] representations to replace the corresponding question representations, we could see a significant performance drop, which demonstrates the importance of relying on all representations of question tokens for reasoning. The gate-based fusion module accounts for about 1.2% of the performance degradation (in EM), which clearly shows its effectiveness. To evaluate the contribution of question-guided information propagation module, we remove it but retain the overall architecture and the global normalization. The result shows that the proposed multi-step, question-guided propagation mechanism helps to improve model's performance by nearly 3% in EM, 2% in F1 and 2.5% in IoU.

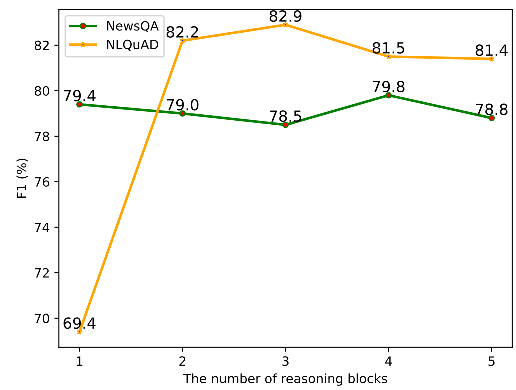

Fig. 2. The effect of number of reasoning blocks.

Effect of Number of Reasoning Blocks We also test the effect of different numbers of reasoning blocks on model performance. The experimental results are shown in Fig. 2. From it, we can observe that the model with 3 reasoning blocks performs best on the NLQuAD dataset. However, on the NewsQA dataset, the model with 4 reasoning blocks works best, though only outperforms the model with 1 reasoning blocks by 0.4% F1. This indicates that the required number of reasoning blocks changes with different datasets. We hypothesize this is due to that different tasks may necessitate different level reasoning ability.

4. CONCLUSION

In this paper, we propose QOPN, a novel model for news machine reading comprehension. It first adopts a context encoding module to capture question-related hints, then apply question-oriented propagation mechanism to perform multi-step reasoning. Experiments demonstrate that the proposed QOPN model achieves new state-of-the-art performance on two challenging machine comprehension datasets. For future work, we will explore the way to apply the idea to other document-level task like query-focused multi-document summarization.

5. ACKNOWLEDGEMENT

We thank all the reviewers for their valuable comments to improve this paper. The work is supported by National Natural Science Foundation of China (Grant No.62036001, No.62032001) and PKU-Baidu Fund (No. 2020BD021). The corresponding author of this paper is Houfeng Wang. Contact emails: {yuco, wanghf}@pku.edu.cn

6. REFERENCES

[1] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "Squad: 100,000+ questions for machine comprehension of text," in *EMNLP*, 2016, pp. 2383–2392.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[4] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman, "Newsqa: A machine comprehension dataset," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, 2017, pp. 191–200.

[5] Amir Soleimani, Christof Monz, and Marcel Worring, "Nlquad: A non-factoid long question answering data set," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 1245–1255.

[6] Christopher Clark and Matt Gardner, "Simple and effective multi-paragraph reading comprehension," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 845–855.

[7] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin, "End-to-end open-domain question answering with bertserini," *NAACL HLT 2019*, p. 72, 2019.

[8] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang, "Multi-passage bert: A globally normalized bert model for open-domain question answering," in *EMNLP-IJCNLP*, 2019, pp. 5878–5882.

[9] Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang, "Cogltx: Applying bert to long texts," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12792–12804, 2020.

[10] Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant, "Coarse-to-fine question answering for long documents," in *ACL*, 2017, pp. 209–220.

[11] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever, "Generating long sequences with sparse transformers," *arXiv preprint arXiv:1904.10509*, 2019.

[12] Iz Beltagy, Matthew E. Peters, and Arman Cohan, "Longformer: The long-document transformer," *arXiv:2004.05150*, 2020.

[13] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al., "Big bird: Transformers for longer sequences.," in *NeurIPS*, 2020.

[14] Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang, "Etc: Encoding long and structured inputs in transformers," in *EMNLP*, 2020, pp. 268–284.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[16] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi, "Bidirectional attention flow for machine comprehension," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017, OpenReview.net.

[17] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.

[18] Dirk Weissenborn, Georg Wiese, and Laura Seiffe, "Making neural qa as simple as possible but not simpler," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 2017, pp. 271–280.

[19] Souvik Kundu and Hwee Tou Ng, "A question-focused multi-factor attention network for question answering," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 5828–5835.

[20] Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong, "Efficient and robust question answering from minimal context over documents," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1725–1735.

[21] Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su, "Densely connected attention propagation for reading comprehension," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 4911–4922.