Read, Attend, and Exclude: Multi-Choice Reading Comprehension by Mimicking Human Reasoning Process

Chenbin Zhang[1,3], Congjian Luo[1,2], Junyu Lu[1], Ao Liu[1], Bing Bai[3], Kun Bai[3] and Zenglin Xu[4,2,1*]
[1] SMILE Lab, School of Computer Science and Engineering, University of Electronic Science and Technology of China
[2] Artificial Intelligence Center, Peng Cheng Lab, Shenzhen, China
[3]Cloud and Smart Industries Group, Tencent, China
[4] School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Shenzhen, China
{aleczhang13,im.congjian,cs.junyu,zeitmond}@gmail.com,{icebai,kunbai}@tencent.com,zenglin@gmail.com

ABSTRACT

Multi-Choice Reading Comprehension (MCRC) is an essential task where a machine selects the correct answer from multiple choices given a context document and a corresponding question. Existing methods usually make predictions based on a single-round reasoning process with the attention mechanism, however, this may be insufficient for tasks that require a more complex reasoning process. To effectively comprehend the context and select the correct answer from different perspectives, we propose the Read-Attend-Exclude (RAE) model which is motivated by what human readers do for MCRC in multi-rounds reasoning process. Specifically, the RAE model includes four components: the Scan Reading Module, the Attended Intensive Reading Module, the Answer Exclusion Module, and the Gated Fusion Module that makes the final decisions collectively based on the aforementioned three modules. Extensive experiments demonstrate the strong results of the proposed model on the DREAM dataset and the effectiveness of all proposed modules.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**.

KEYWORDS

Multi-choice reading comprehension, Multi-rounds reasoning process, Answer exclusion, Gated fusion.

ACM Reference Format:
Chenbin Zhang, Congjian Luo, Junyu Lu, Ao Liu, Bing Bai, Kun Bai, Zenglin Xu. 2020. Read, Attend, and Exclude: Multi-Choice Reading Comprehension by Mimicking Human Reasoning Process. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401326

* Zenglin Xu is the corresponding author.

SIGIR '20, July 25–30, 2020, Virtual Event, China

ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401326

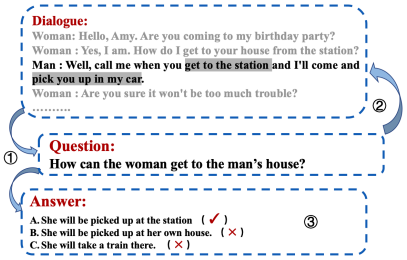

Figure 1: An illustration of the reasoning process of an experienced examine. The first step represents the first round of scan reading. At the second step of attended rereading, the grey boxes at the highlight sentences denote the matched clues to the question. Finally, the third step processes answer comparison and exclusion.

1 INTRODUCTION

As an essential task of retrieval-based question answering, Multi-Choice Reading Comprehension (MCRC) has recently become popular in the Natural Language Processing (NLP) [4, 9, 15] community but it has deep roots in Information Retrieval (IR) [2, 3, 7]. Thanks to the development of large-scale MCRC datasets like DREAM [14], neural network-based models have been proposed to solve these tasks and achieved promising results. Despite their achievements, existing models still have shortcomings in many aspects: (1) Most of the previous methods use a *single round reasoning* process (*i.e.*, encoding, representation enrichment, and inference) only , which lack the re-reading strategy that could help update and refine the information about the reading comprehension tasks. (2) Existing MCRC models [4, 16] either read each candidate answer *independently*, or use questions (and/or candidate answers) to match the document at the final stage. Thus they missed to discriminate the differences between the candidate answers, while the discrimination can be helpful to exclude unreasonable answers. These shortcomings motivate us to incorporate the reading comprehension skills of humans into the design of neural network architectures. Indeed, experienced human readers often adopt more advanced reading skills which is a systematic combination of many reasoning sub-processes [10]. They often firstly do some *scan reading* to get general concepts about the context document, the question, and the answer; and then perform *intensive reading* which looks back at the important part of the document with the awareness of the problem and candidate answers. Moreover, they discriminate the candidate answers at a multiple-granularity level (*i.e.*, words, sentences, and paragraphs),

and make comprehensive considerations for *answer exclusion*. At last, they make an overall decision by combining all the previous information. To better understand the above process, we provide an example of the reasoning process for the case in Figure 1.

Inspired from the aforementioned reasoning process of humans, we propose the Read-Attend-Exclude(RAE) network in Figure 2 which includes four components: Scan Reading, Attended Intensive Reading, Answer Exclusion, and Gated Fusion. Comprehensive experimental results on DREAM dataset demonstrate that the effectiveness of our method to boost baselines' performances. We also conduct ablation experiments to show the contribution of each module of our proposed framework. The details of each module will be explained in the next section.

2 METHODOLOGY

We first describe the problem formulation. Suppose that we have $\langle D, Q, A\rangle$, where $D = \{w_1^d, w_2^d, ..., w_{L_D}^d\}$ is the context document with L_D tokens, and $Q = \{w_1^q, w_2^q, ..., w_{L_Q}^q\}$ is the corresponding question with L_Q tokens. $A = \{A_0, ..., A_k\}$ is the candidate answer set and each candidate answer $A_k = \{w_1^{A_k}, w_2^{A_k}, ..., w_{L_{A_k}}^{A_k}\}$ has L_{A_k} tokens. The goal is to select the right answer from multiple candidates.

2.1 Scan Reading

As shown in the lower left of Figure 2, we use BERT to perform the preliminary scanning of the document along with the question and candidate answers. In detail, we follow the framework in [5] to obtain the [CLS][1] token hidden state h_{CLS} as the reasoning feature, *i.e.*,

$$(h_{CLS}, [\mathbf{D}|\mathbf{Q}|\mathbf{A_k}]) = \mathbf{BERT}([D; Q; A_k]), \tag{1}$$

where $[D; Q; A_k]$ represents a raw text concatenation of each candidate answer A_k, the question Q, and the document D, $[\cdot|\cdot]$ denotes row-wise concatenation, $h_{CLS} \in \mathbb{R}^d$, $\mathbf{D} \in \mathbb{R}^{d\times L_D}$, $\mathbf{Q} \in \mathbb{R}^{d\times L_Q}$, $\mathbf{A_k} \in \mathbb{R}^{d\times L_{A_k}}$ are the hidden states from the last layer of BERT. $\mathbf{D}$ is the representation of document, and $\mathbf{Q}$, $\mathbf{A_k}$ are the representations of question and answer respectively.

2.2 Attended Intensive Reading

After the scan-reading of the whole document, a human tends to re-read the important part of the document with the guidance of a question and its candidate answers. We mimic this process with soft attention to select the most relevant part of the document. In this vein, we first concatenate the BERT hidden states of a question and each candidate answers, then feed them into Bi-LSTM.

$$\tilde{h}_i = \text{Bi-LSTM}([\mathbf{Q}|\mathbf{A}_k], i), \tag{2}$$

where $i \in [0, L_Q + L_{A_K} - 1]$. Here L_Q and L_{A_k} denote the sequence length of the question and each candidate answer respectively. Then, we treat the last hidden state of Bi-LSTM $\tilde{h}_l$ as a query vector and apply soft attention on context document to construct attentive hints features. Let $l = L_Q + L_{A_k} - 1$, we have

$$h_a = \mathbf{D} \cdot \text{SoftMax}(\mathbf{D}^{\mathrm{T}}\tilde{h}_l), \tag{3}$$

where $h_a \in \mathbb{R}^d$, $\mathbf{D} \in \mathbb{R}^{d\times L_D}$.

[1]Delimiter [CLS] is added at the beginning of input sequence and [SEP] are added between D, Q and A_k. We omit [CLS] and [SEP] from the notation for brevity.

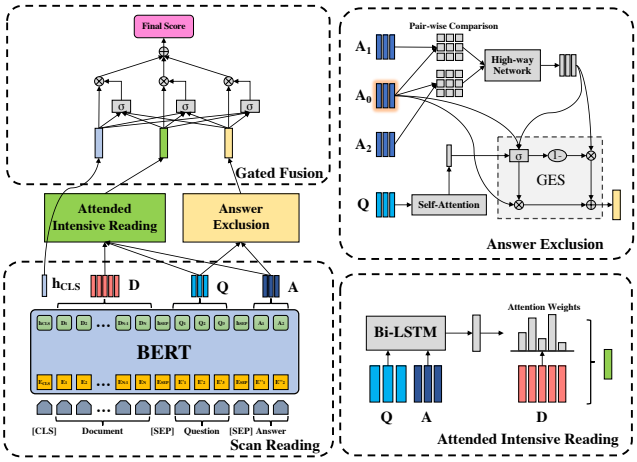

Figure 2: The RAE network for Multi-choice Reading Comprehension. It contains four parts: (1) the Scan Reading Module, (2) the Attended Intensive Reading Module, (3) the Answer Exclusion Module and (4) the Gated Fusion Module. GES is the Gated Exclusion Switching Unit in the Answer Exclusion Module. Best viewed in color.

2.3 Answer Exclusion

This module is used to compare candidate answers pair by pair at a multiple-granularity level to extract helpful correlation information.

Pair-wise Comparisons for Answers. Each candidate answer is compared with all the other candidate answers one by one to collect the pairwise correlation information. Inspired by Ran et al. [12], we use the selected candidate answer $\mathbf{A}_k \in \mathbb{R}^{d\times L_{A_k}}$ to compare with another candidate answer $\mathbf{A}_t \in \mathbb{R}^{d\times L_{A_t}}$ with soft attention mechanism. The attention function is specified by learned parameter $\mathbf{w}_a \in \mathbb{R}^{3d}$:

$$c_{ij} = \mathbf{w}_a^{\mathrm{T}}\left[\mathbf{A}_{t[:,i]}; \mathbf{A}_{k[:,j]}; \mathbf{A}_{t[:,i]} \odot \mathbf{A}_{k[:,j]}\right], \tag{4}$$

$$\mathbf{P} = \left[\frac{\exp(c_{ij})}{\sum_i \exp(c_{ij})}\right]_{i,j}, \tag{5}$$

where $[\cdot;\cdot]$ denotes column-wise concatenation and $\odot$ indicates element-wise product, $\mathbf{P} \in \mathbb{R}^{L_{A_t}\times L_{A_k}}$ is the word-wise attention matrix. Then the final comparison result $\hat{\mathbf{A}}_k^{(t)} \in \mathbb{R}^{2d\times L_{A_k}}$ is computed as follows:

$$\hat{\mathbf{A}}_k^{(t)} = [\mathbf{A}_k - \mathbf{A}_t\mathbf{P}; \mathbf{A}_k \odot \mathbf{A}_t\mathbf{P}]\,. \tag{6}$$

Integration of All Comparison Information. Afterwards, we use a three-layer Highway network [13] to combine the pair-wise comparison results from the selected answer k with all other answers together, which can retain more original information and get a multi-granularity representation. When backpropagating, more (gradient) information can be directly returned to the input without a non-linear transformation, *i.e.*,

$$\widetilde{\mathbf{A}}_k = \mathbf{Highway}(\{\hat{\mathbf{A}}_k^{(t)}\}_{t\neq k}), \tag{7}$$

where $\{\hat{A}_k^{(t)}\}_{t\neq k} \in \mathbb{R}^{((|A|-1)*2d)\times L_{A_k}}$ is the concatenation of the representations of the other $(|A| - 1)$ answers on the hidden dimension. Then we perform a non-linear transformation to reduce the dimension of $\widetilde{\mathbf{A}}_k$, which represents how $\widetilde{\mathbf{A}}_k$ can be

aligned to each hidden state in $\mathbf{A_k}$: $\widetilde{\mathbf{A}}_k^c = \tanh\left(\mathbf{W}_c^T \widetilde{\mathbf{A}}_k + \mathbf{b}_c\right)$, where $\widetilde{\mathbf{A}}_k^c \in \mathbb{R}^{d \times L_{A_k}}$ represents the "comparison information" of answer k, $\mathbf{W}_c \in \mathbb{R}^{((|A|-1)*2d) \times d}$ and $\mathbf{b}_c \in \mathbb{R}^{d \times L_{A_k}}$ are parameters to learn.

Self-Attention for Question Summarization After collecting the selected candidate answer's representation $\mathbf{A}_k$ and its comparison information $\widetilde{\mathbf{A}}_k^c$. According to the analyses in Section 1, humans will make answer exclusion based on the knowledge of the question and the candidate answers from scan reading. We apply soft-attention to get the question representation $\widetilde{\mathbf{Q}}$ as follows:

$$\widetilde{\mathbf{Q}} = \mathbf{Q} \cdot \text{SoftMax}(\mathbf{Q^T w_q}), \tag{8}$$

where $\mathbf{w}_q \in \mathbb{R}^d$ is parameter to learn, and $\widetilde{\mathbf{Q}} \in \mathbb{R}^d$.

Gated Exclusion Switching Unit for Answer Exclusion. With the knowledge of the question, the selected candidate answer, as well as the comparison result with all other candidate answers, answer exclusion can be achieved. We calculate the exclusion gate $\mathbf{g}_k$ as follows:

$$\mathbf{g}_k = \sigma\left([\mathbf{A}_k | \widetilde{\mathbf{A}}_k^c | \widetilde{\mathbf{Q}}]\mathbf{W}_g + \mathbf{b}_g\right), \tag{9}$$

where $\sigma = \frac{1}{1+e^x}$ is the sigmoid function, and $\mathbf{g}_k \in \mathbb{R}^{d \times L_{A_k}}$, $\mathbf{W}_g \in \mathbb{R}^{(2L_{A_k}+1) \times L_{A_k}}$ and $\mathbf{b}_g \in \mathbb{R}^{d \times L_{A_k}}$ are learned parameters which decide to what extent an answer k should be preserved or excluded. The final information vector I is computed as follows:

$$I = \mathbf{g}_k \odot \mathbf{A}_k + (1 - \mathbf{g}_k) \odot \widetilde{\mathbf{A}}_k^c, \tag{10}$$

and finally the output of Answer Exclusion Module, *i.e.*, h_c, is computed with MaxPooling:

$$h_c = \text{MaxPooling}(I), \tag{11}$$

where the $h_c \in \mathbb{R}^d$.

2.4 Gated Fusion

This model mimics the complicated and interpretable thinking activity of a human when he tries to integrate all the information he has gained. In this sense, the balance of these three sources of information needs to be dynamically learned. So we calculate $\hat{h}_k$ as follows, and h_a, h_{CLS}, h_c are the outputs of modules defined before:

$$\hat{h}_k = g_r \odot h_{CLS} + g_a \odot h_a + g_c \odot h_c, \tag{12}$$

and each module's information flows the gate g must consider all the information, so calculates as follow:

$$g_r = \sigma(\mathbf{W}_{r_{CLS}} h_{CLS} + \mathbf{W}_{r_a} h_a + \mathbf{W}_{r_c} h_c + b_r), \tag{13}$$

$$g_a = \sigma(\mathbf{W}_{a_{CLS}} h_{CLS} + \mathbf{W}_{a_a} h_a + \mathbf{W}_{a_c} h_c + b_a), \tag{14}$$

$$g_c = \sigma(\mathbf{W}_{c_{CLS}} h_{CLS} + \mathbf{W}_{c_a} h_a + \mathbf{W}_{c_c} h_c + b_c), \tag{15}$$

where $\sigma = \frac{1}{1+e^x}$ is the sigmoid function, and $\odot$ indicates the element-wise product.

The probability $P(k|D, Q, A_k)$ of the k-the candidate answer to be correct is computed as follows:

$$p(k|D, Q, A_k) = \frac{exp(\mathbf{W_k}\hat{h}_k + b_k)}{\sum_{k=0}^{K} exp(\mathbf{W_k}\hat{h}_k + b_k)}. \tag{16}$$

And the loss function is defined as:

$$J(\theta) = -\frac{1}{N}\sum_i log(p(\hat{k}|D, Q, A_k)) + \lambda||\theta||_2^2, \tag{17}$$

where θ denotes all trainable parameters, N is the number of training examples in dataset, and $\hat{k}$ is the ground truth answer index.

3 EXPERIMENTS

3.1 Experimental Setup

We evaluate our model on the **DREAM** [14] dataset. It contains 6,444 dialogues with 10,197 multiple choice questions collected from English as a foreign language examination designed by human experts. Each question of DREAM has three candidate answers to select. In detail, 84% of answers are non-extractive, 85% of questions require reasoning beyond a single sentence, and 34% of questions also involve commonsense knowledge.

We set training details following literature of BERT. The dimensions of embedding are 768 provided by BERT_{base} and 1024 provided by BERT_{large}, respectively. The max length L is 512 for RAE_{large} and 256 for RAE_{base}. We use the Adam optimizer to train our model. The model RAE_{base} is trained for 15 epochs with a batch size 8, max length 256 and learning rate 2×10^{-5} when BERT_{base} is used, and another model RAE_{large} trained for 10 epochs with a batch size 24, max length 512 and learning rate 3×10^{-5} when BERT_{large} is used. Max length L means the maximum length of the sum of document, the question and candidate answers. The $L2$ weight decay λ is set to 0.01 and we apply warm-up method. In order to guarantee the RAE model read the question and the candidate answer completely, the dialogue is truncated sometimes when the total length exceeds the max length.

3.2 Evaluation Results

We compare our model with various state-of-the-art methods and the results are shown in Table 1. For the DREAM dataset, the methods we compared include **SAR** (Stanford Attentive Reader) [1], **GAR** (Gated-Attention Reader) [6], **Hier-Co-Matching** [16], **FTLM** (Finetuned Transformer LM) [11], **EER**(Evidence Sentence Extraction) [15], and approaches proposed by the dataset's authors [14], including **DSW++, GBDT++, FTLM++** etc. According to Table 1, we can observe that:

(1) Firstly, we verify the effectiveness of the RAE network by comparing it with other baseline methods in Table 1. On DREAM, our model outperforms the baselines significantly, indicating the effectiveness of our model. RAE network achieves 68.5% at accuracy base on BERT_{large} which is 10.8% better than the state-of-the-art model EER_{DPL}, demonstrating the effectiveness of our model. By giving the RAE modules, our proposed method improves the performance of the BERT-large-based fine-tuned transformer model by 1.7%. Note that EER_{DPL} also uses the powerful pre-training model GPT.

(2) Note that we can directly use the other pre-training models(*e.g.* XLNet [17], RoBERTa [8], etc) as Scan Reading Module, so there might be room for further performance improvement.

3.3 Ablation Study

We also carried out model ablation to further demonstrate the effectiveness of the proposed modules of RAE network in DREAM dataset. Since the pre-training model is a powerful feature extractor that can capture rich semantics, showing a significant effect in the previous state-of-the-art model, we consider the sub-modules of our model as key factors: (1) Attended Intensive Reading Module (2) Answer Exclusion Module (3) Gated Fusion Module.

DREAM Dataset		
Model	DEV	Test
SAR [1]	40.2	39.8
GAR [6]	40.5	41.3
Hier-Co-Matching [16]	45.6	45.5
FTLM [11]	55.9	55.5
Sun et al. [14]'s Approaches:		
DSW++	51.4	50.1
GBDT++	53.3	52.8
FTLM++	57.6	57.4
Wang et al. [15]'s Approaches:		
$\text{EER}_{silver-gt}$	50.1	50.4
EER_{DS}	55.1	56.3
EER_{DPL}	57.3	57.7
BERT_{large}	66.0	66.8
BERT_{base}	63.4	63.2
Our Approaches:		
RAE_{large}	**66.8**	**68.5**
RAE_{base}	**64.2**	**64.6**

Table 1: Question answering accuracy of different methods on DREAM Dataset. Note that we run the experiments 5 times with different random seeds and report the average test set performance and the corresponding average dev set performance.

Model	DREAM
RAE Model	**64.6**
w/o Attended Intensive Reading Module	63.5
w/o Answer Exclusion Module	62.9
w/o Gated Fusion Module	63.1

Table 2: Ablations on several model module on DREAM dataset.

The results are shown in Tables 2 and verify the effectiveness of the each part in the RAE network. As shown in the Section 2, we adopt the attended intensive rereading module to enhance the ability of the model to select evidence sentences. And this module is useful in dynamically adjusting the attention values. When we remove this module, there is an absolute drop of 1.1%. This suggests that it is necessary to correlate the candidate answer between the question and the document of the text in the semantic space that can incorporate various significant information to make a decision. We also remove our answer exclusion module in the model. The result shows a significant drop in performance by 1.7 %, indicating that the proposed answer exclusion module is effective to extract correlation information between these candidate answers pairs and support reasoning. At last, we remove the Gated Fusion Module. The result shows an intuitive drop in performance by 1.5%. The result indicates that the gated fusion is effective to control the information flow in dynamical, which can combine multi-source information and make the best decision.

4 CONCLUSION AND FUTURE WORK

In this paper, we propose an integrated RAE network for Multi-choice Reading Comprehension (MCRC). Our model mimics how a human reader does reading comprehension tasks and provides good interpretability. And We showed that our model achieved strong results on DREAM MCRC dataset. In the future, we would like to combine adversarial methods to further improve the effect and experiment on other datasets.

ACCKNOWLEDGEMENT

This work was partially supported by the National Key Research and Development Program of China (No. 2018AAA0100204). We would like to thank Dr. Jian Liang and Jing Xu for their insightful suggestions.

REFERENCES

[1] Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (2016).
[2] Qin Chen, Qinmin Hu, Jimmy Xiangji Huang, Liang He, and Weijie An. 2017. Enhancing recurrent neural networks with positional attention for question answering. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 993–996.
[3] Wanyu Chen, Fei Cai, Honghui Chen, and Maarten de Rijke. 2018. Attention-based hierarchical neural query suggestion. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1093–1096.
[4] Zhipeng Chen, Yiming Cui, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. Convolutional spatial attention model for reading comprehension with multiple-choice questions. *Proceedings of the AAAI. Honolulu, HI* (2019).
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2019).
[6] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (2017).
[7] Ao Liu, Lizhen Qu, Junyu Lu, Chenbin Zhang, and Zenglin Xu. 2019. Machine Reading Comprehension: Matching and Orders. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*. 2057–2060.
[8] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[9] Junyu Lu, Chenbin Zhang, Zeying Xie, Guang Ling, Tom Chao Zhou, and Zenglin Xu. 2019. Constructing Interpretive Spatio-Temporal Features for Multi-Turn Responses Selection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 44–50.
[10] Scott G Paris, Marjorie Y Lipson, and Karen K Wixson. 1983. Becoming a strategic reader. *Contemporary educational psychology* 8, 3 (1983), 293–316.
[11] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf* (2018).
[12] Qiu Ran, Peng Li, Weiwei Hu, and Jie Zhou. 2019. Option Comparison Network for Multiple-choice Reading Comprehension. *arXiv preprint arXiv:1903.03033* (2019).
[13] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387* (2015).
[14] Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics (TACL 2019)* (2019).
[15] Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, Dan Roth, and David McAllester. 2019. Evidence Sentence Extraction for Machine Reading Comprehension. *arXiv preprint arXiv:1902.08852* (2019).
[16] Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018. A Co-Matching Model for Multi-choice Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 746–751.
[17] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*. 5754–5764.