CONTEXT MODELING WITH EVIDENCE FILTER FOR MULTIPLE CHOICE QUESTION ANSWERING

Sicheng Yu[1], *Hao Zhang*[2], *Wei Jing*[3] , *Jing Jiang*[1]

[1]Singapore Management University, Singapore
[2]Nanyang Technological University, Singapore
[3]Alibaba Group, China

scyu.2018@phdcs.smu.edu.sg, hao007@e.ntu.edu.sg, 21wjing@gmail.com, jingjiang@smu.edu.sg

ABSTRACT

Multiple-Choice Question Answering (MCQA) is one of the challenging tasks in machine reading comprehension. The main challenge in MCQA is to extract "evidence" from the given context that supports the correct answer. In OpenbookQA dataset [1], the requirement of extracting "evidence" is particularly important due to the mutual independence of sentences in the context. Existing work tackles this problem by annotated evidence or distant supervision with rules which overly rely on human efforts. To address the challenge, we propose a simple yet effective approach termed **evidence filtering** to model the relationships between the encoded contexts with respect to different options collectively, and to potentially highlight the evidence sentences and filter out unrelated sentences. In addition to the effective reduction of human efforts of our approach compared, through extensive experiments on OpenbookQA, we show that the proposed approach outperforms the models that use the same backbone and more training data; and our parameter analysis also demonstrates the interpretability of our approach.

Index Terms— Natural Language Processing, Machine Reading Comprehension, Question Answering, Evidence Extraction

1. INTRODUCTION

Multiple-Choice Question Answering (MCQA) is a natural language processing task that has been attracting much attention recently due to its wide range of applications [2–4]. One of the key challenges for MCQA is to retrieve the evidence sentences, to support answer prediction. Unfortunately, those evidence sentences are usually overwhelmed by a large number of unrelated sentences in the context [5, 6]. This is more serious in OpenbookQA dataset [1] as the context of OpenbookQA consists of several independent facts and distribution of evidence sentences is random, which is difficult to locate the evidence sentence precisely. To address this issue, one solution is to train the model to learn how to extract evidence from context [6–10]. However, it requires a large amount of annotated data, which is not practical in real-world scenarios. Another stream of work [11] utilizes distant supervision with a series of human-designed rules, which still requires much human effort.

In this paper, we present a simple yet effective approach, termed evidence filter, to potentially filter out irrelevant context and highlight the evidence without any human intervention. The key idea is to model the relationships between the encoded contexts with respect to different options. Specifically, we observe that most of methods for multiple-choice question answering typically encode each option with the context independently [12–14].

Question: What is the best way to guess a baby's eye color?

A. The surroundings they are born in.
B. Their parents' usual diet.
C. Just take a random guess.
D. The genealogy records of their family.

Context: It is an academic guess too. Hypothesis means scientific guess about the cause and effect of an event. **Eye color is an inherited characteristic.** Some monkey babies can be raised with two parents. Ancestors are family members. The color of an object is the result of the way the object reflects or emits light. **Having offspring produces a family.** Adults have babies. Seals are born on waterfronts. Sugars are important for a plant's diet. Climate is the usual kind of weather in a location. Babies need milk to live. Frog babies in sacs are in eggs. Animals take in oxygen. The crust is just above the mantle. The vision organ is the eye. A person's diet determines nutrient levels. An omnivore includes animals in its diet. Drought is a period of less than usual precipitation.

Fig. 1. An example in OpenbookQA. Sentences in blue and bold are the evidence while the correct option is underlined.

Our method is based on the following observations and assumptions: (1) If a sentence in the context has a similar level of relevance on all of the given options, then it is highly likely that this sentence is not useful for answering the question. For instance, the sentence *"The color of an object is the result of the way the object reflects or emits light."* is not relevant to any of the options in Figure 1. We therefore believe that this sentence is unlikely to be an evidence sentence. (2) An evidence sentence in the context is likely to be closely related to the correct option but irrelevant to the incorrect options. For instance, the evidence sentences shown in blue are indeed more related to the ground-truth option *D* than to the other options. Motivated by the aforementioned assumptions, we propose a method to capture the differences among context sentences with respect to the options via a carefully designed evidence filter, which can be treated as a denoising process or a high-pass filter.

In a nutshell, we propose a simple yet effective evidence filter to potentially extract evidence from context with *interpretable* parameters of evidence filter. Experimental results on the OpenbookQA dataset demonstrate the effectiveness of our approach, which outperforms BERT Multi-task by 1.8% with the same backbone and less training data.

2. METHODOLOGY

Our proposed framework consists of two components: 1) an *evidence extraction* module that uses pre-trained BERT to process the context C, the question Q and one of the options O_i in order to implicitly extract evidence from C that justifies O_i as the answer to Q; 2) an *evidence filter* module that adjusts the extracted evidence representation by considering the relationship between the evidence extracted from the previous module with respect to different options. Figure 2 depicts the overall architecture of the proposed framework.

2.1. Evidence Extraction

The goal of the evidence extraction module is to process the context with respect to the question and a particular option in order to obtain a vector that represents the evidence from the context justifying the option. Suppose we have the context C (which consists of a list of sentences extracted from a given corpus by an information retrieval method), question Q and four answer options (O_1, O_2, O_3, O_4). The evidence extraction module uses BERT [13] to process the sequence $[C; (Q; O_i)]$ for $i = 1, 2, 3, 4$, where C is treated as the first segment and the concatenation of Q and O_i is treated as the second segment for BERT to process. Since BERT has multiple blocks (*i.e.*, layers), we use $\text{BERT}_k([C; (Q; O_i)]) \in \mathbb{R}^d$ to denote the output vector (after pooling) from the k-th block of BERT after processing context question and i-th option. With 4 options, we use $\text{BERT}_k([C; (Q; O_i)]^4) \in \mathbb{R}^{4\times d}$ to denote all 4 output vectors. We adopt BERT-large with $K = 24$ blocks due to its better performance and generalization ability compared to BERT-base.

2.2. Evidence Filter

Recall that in Section 1, we pointed out that whether or not a sentence serves as evidence to support an option also depends on whether this sentence is relevant to other options. Our assumptions are that (1) sentences related to the four options in similar ways are unlikely to be useful evidence; (2) sentences related to one option but not others are likely to be useful evidence. Based on these assumptions, we introduce an evidence filter matrix to adjust the extracted evidence representations from the *evidence extraction* module; specifically, the evidence filter matrix will reduce the importance of evidence that is equally relevant to the 4 options, and subsequently place more emphasis on evidence relevant to a particular option.

Specifically, this is a 4×4 matrix inspired by [15], denoted as $\mathbf{A}$, that will be applied to $\text{BERT}_k([C; (Q; O_i)]^4)$. Its diagonal values represent how much we want to maintain the originally extracted evidence representation for each option, whereas off-diagonal values represent how much reduction we would like to incur based on relations across different options. We thus expect the signs of the diagonal values to be the opposite of the signs of the off-diagonal values in $\mathbf{A}$. After the adjustment, the new evidence representations will be $\mathbf{A} \cdot \text{BERT}_k([C; (Q; O_i)]^4)$.

In the preliminary experiments, we observe that randomly and individually initializing each entry in the evidence filter may lead to inconsistent predictions when the 4 answer options are shuffled, *i.e.*, when the order of the options changes. To alleviate such drawback, we apply the following constraints to the evidence filter. Specifically, we constrain the evidence filter matrix such that all its diagonal entries are the same, denoted as α, and all its off-diagonal entries are also the same, denoted as β. They are represented as the blue cells and grey cells in the evidence filter matrix shown in Figure 2.

Dataset	OpenbookQA		
subset	Train	Dev	Test
# questions	4957	500	500
Avg. question sentences	1.08		
Avg. question tokens	11.46		
Avg. choice tokens	2.89		
Avg. science fact tokens	9.38		
Vocabulary size	12839		

Table 1. Statistics on OpenbookQA dataset from [1]

***Derivation for Evidence Filter*:** Firstly, we explain why evidence filter without constrains may cause different prediction when shuffling the options. For clear illustration, we use $\mathbf{R} \in \mathbb{R}^{4\times4}$ to represent a row (column) exchange matrix which has the following properties:

$$\sum_j \mathbf{R}_{ij} = 1, \quad \sum_i \mathbf{R}_{ij} = 1, \quad \mathbf{R}_{ij} = 0 \text{ or } 1. \tag{1}$$

The output of block k in BERT corresponding to a same sample with shuffled options is equivalent to the row exchange of $\text{BERT}_k[C; (Q; O_i)]^4$, which can be written as $\mathbf{R}\cdot\text{BERT}_k[C; (Q; O_i)]^4$. Shuffling the order of inputs to the evidence filter should be equivalent to first entering the evidence filter and then shuffling the output. The expected property of the evidence filter can be formulated as:

$$\begin{aligned} &\mathbf{A} \cdot (\mathbf{R} \cdot \text{BERT}_k[C; (Q; O_i)]^4) \\ &= \mathbf{R} \cdot (\mathbf{A} \cdot \text{BERT}_k[C; (Q; O_i)]^4) \end{aligned} \tag{2}$$

which is not satisfied if evidence filter is no-constrains attached.

Then we prove that evidence filter with constrains is about to completely solve this unexpected phenomenon. We need to demonstrate that $\mathbf{A} \cdot \mathbf{R} = \mathbf{R} \cdot \mathbf{A}$. Recall the definition of $\mathbf{A}$ and $\mathbf{R}$. The i, j entry of $\mathbf{A} \cdot \mathbf{R}$ and $\mathbf{R} \cdot \mathbf{A}$ can be easily derived as following:

$$\begin{aligned} (\mathbf{A} \cdot \mathbf{R})_{ij} &= \sum_k \mathbf{A}_{ik} \cdot \mathbf{R}_{kj} \\ &= \alpha\mathbf{R}_{ij} + \beta \sum_{k\neq i} \mathbf{R}_{kj} = (\alpha - \beta)\mathbf{R}_{ij} + \beta, \end{aligned} \tag{3}$$

$$\begin{aligned} (\mathbf{R} \cdot \mathbf{A})_{ij} &= \sum_k \mathbf{R}_{ik} \cdot \mathbf{A}_{kj} \\ &= \alpha\mathbf{R}_{ij} + \beta \sum_{k\neq j} \mathbf{R}_{ik} = (\alpha - \beta)\mathbf{R}_{ij} + \beta, \end{aligned} \tag{4}$$

where α represents the diagonal values and β denotes the rest parameters of evidence filter.

With this constrained evidence filter matrix, the representation of evidence for the 1st option after this filter, for example, is $\alpha \cdot \text{BERT}_k([C; (Q; O_1)]) + \sum_{i\neq1} \beta \cdot \text{BERT}_k([C; (Q; O_i)])$, which exactly extracts the difference when the values of α and β have opposite signs. It helps the model to highlight the evidence through taking the differences of the extracted contexts with respect to different options.

We also adopt different evidence filters $\mathbf{A}_k$ for each block in BERT due to its better expression than sharing one evidence filter among all blocks. Meanwhile, it also makes the block fusion layer more explicit.[1] A residual connection is adopted after the evidence filter followed by layer normalization [16]. The intermediate output,

[1]The same pattern of evidence filter in Figure 2 is for simplicity.

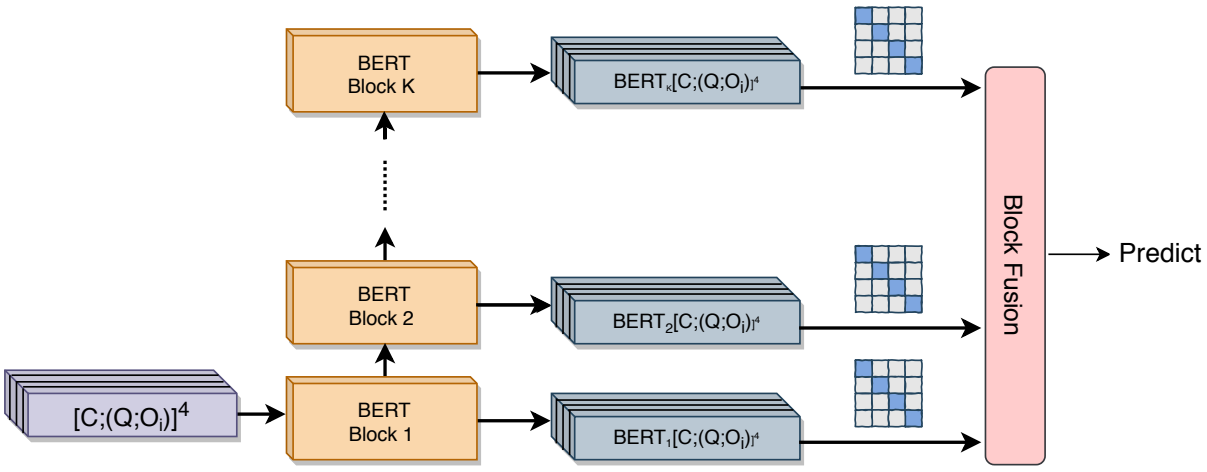

Fig. 2. An overview of our proposed model.

Methods	OpenbookQA Dev (%)	OpenbookQA Test (%)
Question Match + ELMo [1]	54.6	50.2
Odd-one-out Solver [1]	56.9	50.2
ESIM + ELMo [1]	53.9	48.9
OFT [12]	-	52.0
OFT (ensemble) [12]	-	52.8
Reading Strategies+GPT [12]	-	55.2
Reading Strategies+GPT (ensemble) [12]	-	55.8
BERT-large (leaderboard)	-	60.4
BERT(large) Multi-task (leaderboard)	-	63.8
Ours Model	**66.8**	**65.6**

Table 2. Accuracy comparison between ours and other methods, where "-" means not available.

$\text{BERT}'_k([C;(Q;O_i)]^4)$, for block k before block fusion is computed by:

$$\text{BERT}'_k([C;(Q;O_i)]^4) = \text{LayerNorm}\left(\text{BERT}_k([C;(Q;O_i)]^4) + \mathbf{A}_k \cdot \text{BERT}_k([C;(Q;O_i)]^4)\right). \quad (5)$$

Then a block fusion layer is adopted to integrate the intermediate output from each block by a single linear layer, and the output of block fusion layer $\mathbf{M}$ is defined as:

$$\mathbf{M} = \mathbf{W}_{\text{bf}}[\text{BERT}'_{1:K}([C;(Q;O_i)]^4)], \quad (6)$$

where "$1:K$" denotes the concatenation from 1 to K and $\mathbf{W}_{\text{bf}} \in \mathbb{R}^{1\times K}$. Output is $\mathbf{M} \in \mathbb{R}^{4\times d}$.

Finally, the model makes prediction after a linear layer and the standard cross-entropy loss is utilized as loss function. Compared to the BERT model, our method only requires a few more parameters, *i.e.*, those for the evidence filters and those at the fusion layer.

3. EXPERIMENTS

3.1. Datasets and Implementation

3.1.1. Datasets

To evaluate our model, we conduct experiments on a challenging MCQA dataset, OpenbookQA [1]. The statistics of OpenbookQA is shown in Table 1. which contains around 6000 4-way multiple-choice science questions and a corpus including roughly 7300 facts. Compared to other MCQA datasets, OpenbookQA is more complicated since the context in this dataset may not contain the answers explicitly [17].

Modification	Accuracy (%)
(1) w/o block fusion; w/o evidence filter	60.0
(2) w/o block fusion; evidence filter w/o constraints	63.8
(3) w/o block fusion; evidence filter	65.0
(4) block fusion with same evidence filter	64.0
block fusion with different evidence filter (ours)	65.6

Table 3. The performance of our model and its ablations on OpenbookQA test set.

3.1.2. Experimental Settings

Following [12], we utilize PyLucene [18] toolkit to retrieve the top-30 ranked sentences from the corpus containing the facts[2]. Then, we fine-tune BERT-large provided by Transformers [19] on RACE [20] with the naïve model as in [13]. Finally, we train the model on OpenbookQA dataset following the architecture in Figure 2. The batch size is set to 32 and learning rate is $1e-5$. Adam [21] is adopted as optimizer. The linear warm up strategy is used with the first 10% of the whole training steps. We also add another group of input, $[Q;O_i]$ sharing the parameters of BERT but evidence filter are not used for this group of input. The reason is that BERT contains commonsense knowledge which is helpful in scientific questions in OpenbookQA.

3.2. Results and Analysis

3.2.1. Main Results

We compared our model with several previous state-of-the-art methods, which uses the same backbone as ours. The results are summarized in Table 2. Observed that our model outperforms all the

[2]We mainly focus on the method after the retrieval step. Thus, the information retrieval module is not included in this section. In other words, our model can be easily adapted to any information retrieval methods.

Index	1	2	3	4	5	6	7	8	9	10	11	12
α	1.3418	1.3418	1.3418	1.3418	1.3418	1.3418	1.3408	1.3389	1.3447	1.3457	1.3447	1.3389
β	-1.0693	-1.0693	-1.0693	-1.0693	-1.0693	-1.0693	-1.0703	-1.0723	-1.0664	-1.0664	-1.0664	-1.0723
Index	13	14	15	16	17	18	19	20	21	22	23	24
α	1.3379	1.3457	1.3369	1.3477	1.3467	1.3457	1.3398	1.3467	1.3477	1.3477	1.3486	1.3408
β	-1.0723	-1.0654	-1.0742	-1.0635	-1.0645	-1.0654	-1.0713	-1.0654	-1.0635	-1.0635	-1.0625	-1.0713

Table 4. Values of α and β in evidence filter for each block of model after training.

Error category	I	II	III
# of error instances	11	7	2

Table 5. The statistics of 20 error instances.

compared methods[3]. The BERT-based model also shows stable superiority over the models based on other pre-trained language model such as ELMo [22] and GPT [23]. Despite the simplicity and few additional parameters of our model, it still outperforms other BERT based approaches. For instance, our model surpasses BERT Multi-task method by 1.8% in accuracy, where BERT Multi-task is first trained on RACE and then fine-tuned on both OpenbookQA [1] and ARC [24].

3.2.2. Ablation Studies.

In Table 3, we report the results of 5 different ablation settings: (1) remove block fusion layer in equation 6 and evidence filter from our model; (2) remove block fusion layer and keep evidence filter without constraints; (3) remove block fusion layer and keep evidence filter with constraints; (4) keep both block fusion layer and evidence filter with constraints, all the blocks share the same evidence filter. According to the ablative results, we show that our full approach is superior to other variants and achieves the best performance. It demonstrates the effectiveness of evidence filter by comparing (1) with (2)-(4). In (2), we also evaluate the trained model with three sets of shuffled options using different random seed and get varied results: 63.8%, 63.6% and 63%, which exposes the problem mentioned in Section 2.2. In contrast, all the experimental setups based on constraint-attached matrix are not affected by the shuffle operation. The results from (3), (4) and the last one (ours) suggest that block fusion sharing the same evidence filter performs worse than the model without block fusion, while performs better when different evidence filters are applied.

3.2.3. Analysis of Evidence Filter

As discussed in Section 2.2, we expect the α and β of the evidence filter should satisfy $\alpha \times \beta < 0$. From Table 4, we can observe that the values of α and β show opposite sign for all blocks. Another fact is that α are positive values and β are negatives, which also represent the intra-interaction and inter-interaction. We can conclude that the parameters of the well-designed evidence filter are consistent with our intuition, which also provides an explanation for itself.

[3]Note that we do not compare with the models utilizing stronger language models or external knowledge bases.

3.2.4. Error Analysis

In this section we analyze the prediction of instances in OpenbookQA for error analysis. We randomly choose 20 error prediction for further analysis and categorize them to three classes:

I Information incomplete: the information provided in the context is not enough to answer the question.

II Commonsense incomplete: the information provided in the context is enough, however the model is lack of some key commonsense. This class can be seen as the lack of very common knowledge not specialized knowledge as in information incomplete.

III Information complete: from a human point of view, the information is complete, but the model made a wrong prediction.

The statistics of error analysis in shown in table 5.

In the situation of I, the model is not able to do the correct prediction based on the context without related sentences. For example, one of the 11 questions in I is *"A Mola Mola might live where?"*. However, no word *mola* appears in the context. It is reasonable that our model collapses on these kind of questions since it is random guess even for humans without corresponding knowledge. For II, representative question is *"What is different about birth in humans and chickens?"*. Four candidate options are *Father*, *Mother*, *Fertilization*, *the hard shell*. One given sentences is *"Live birth means developing inside the mother instead of an egg."* which already provides enough information. The problem here is that the neural network do not know egg is with hard shell which is obvious to humans. Another example is *"The summer solstice in the northern hemisphere is four months before?"* with four answer of months which is also answered directly in context. The model has neither the ability to solve simple arithmetic problems nor the English expression of the the exact months. In the last class, the model make incorrect predictions with rather direct evidences in the context.

4. CONCLUSIONS

We propose evidence filter to alleviate the effect of unrelated sentences and enhance the saliency of evidences potentially without human efforts. Results on OpenbookQA indicate the effectiveness of our method. Our future work is to enhance the evidence filter by more complex components.

5. ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore under its Strategic Capabilities Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

6. REFERENCES

[1] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal, "Can a suit of armor conduct electricity? a new dataset for open book question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2381–2391.

[2] Abdelghani Bouziane, Djelloul Bouchiha, Noureddine Doumi, and Mimoun Malki, "Question answering systems: survey and trends," *Procedia Computer Science*, vol. 73, pp. 366–375, 2015.

[3] Poonam Gupta and Vishal Gupta, "A survey of text question answering techniques," *International Journal of Computer Applications*, vol. 53, no. 4, 2012.

[4] Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani, "A survey on machine reading comprehension systems," *CoRR*, vol. abs/2001.01582, 2020.

[5] Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita, "Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2335–2345.

[6] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2369–2380.

[7] Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu, "Dynamically fused graph network for multi-hop reasoning," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6140–6150.

[8] Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang, "Cognitive graph for multi-hop reading comprehension at scale," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2694–2703.

[9] Siva Reddy, Danqi Chen, and Christopher D Manning, "Coqa: A conversational question answering challenge," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 249–266, 2019.

[10] Qiyuan Zhang, Lei Wang, Sicheng Yu, Shuohang Wang, Yang Wang, Jing Jiang, and Ee-Peng Lim, "Noahqa: Numerical reasoning with interpretable graph question answering dataset," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Findings*, 2021.

[11] Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth, "Evidence sentence extraction for machine reading comprehension," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, pp. 696–707.

[12] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie, "Improving machine reading comprehension with general reading strategies," 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)*, 2019, pp. 2633–2643.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "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)*, 2019, pp. 4171–4186.

[14] Di Jin, Shuyang Gao, Jiun-Yu Kao, Tagyoung Chung, and Dilek Hakkani-tur, "Mmm: Multi-stage multi-task learning for multi-choice reading comprehension," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 8010–8017.

[15] Jianfei Yu, Minghui Qiu, Jing Jiang, Jun Huang, Shuangyong Song, Wei Chu, and Haiqing Chen, "Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2018, pp. 682–690.

[16] Jimmy Ba, Ryan Kiros, and Geoffrey E. Hinton, "Layer normalization," *CoRR*, vol. abs/1607.06450, 2016.

[17] Jifan Chen and Greg Durrett, "Understanding dataset design choices for multi-hop reasoning," 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)*, 2019, pp. 4026–4032.

[18] Andrzej Białecki, Robert Muir, and Grant Ingersoll, "Apache lucene 4," 2012.

[19] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[20] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy, "Race: Large-scale reading comprehension dataset from examinations," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 785–794.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[22] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana, June 2018, pp. 2227–2237, Association for Computational Linguistics.

[23] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving language understanding by generative pre-training," 2018.

[24] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashutosh Sabharwal, Carissa Schoenick, and Oyvind Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *CoRR*, vol. abs/1803.05457, 2018.