Improving Machine Reading Comprehension with General Reading Strategies

Kai Sun[1]* **Dian Yu**[2] **Dong Yu**[2] **Claire Cardie**[1]
[1]Cornell University, Ithaca, NY, USA
[2]Tencent AI Lab, Bellevue, WA, USA
ks985@cornell.edu, {yudian, dyu}@tencent.com
cardie@cs.cornell.edu

Abstract

Reading strategies have been shown to improve comprehension levels, especially for readers lacking adequate prior knowledge. Just as the process of knowledge accumulation is time-consuming for human readers, it is resource-demanding to impart rich general domain knowledge into a deep language model via pre-training. Inspired by reading strategies identified in cognitive science, and given limited computational resources – just a pre-trained model and a fixed number of training instances – we propose three general strategies aimed to improve non-extractive machine reading comprehension (MRC): (i) BACK AND FORTH READING that considers both the original and reverse order of an input sequence, (ii) HIGHLIGHTING, which adds a trainable embedding to the text embedding of tokens that are relevant to the question and candidate answers, and (iii) SELF-ASSESSMENT that generates practice questions and candidate answers directly from the text in an unsupervised manner.

By fine-tuning a pre-trained language model (Radford et al., 2018) with our proposed strategies on the largest general domain multiple-choice MRC dataset RACE, we obtain a 5.8% absolute increase in accuracy over the previous best result achieved by the same pre-trained model fine-tuned on RACE without the use of strategies. We further fine-tune the resulting model on a target MRC task, leading to an absolute improvement of 6.2% in average accuracy over previous state-of-the-art approaches on six representative non-extractive MRC datasets from different domains (i.e., ARC, OpenBookQA, MCTest, SemEval-2018 Task 11, ROCStories, and MultiRC). These results demonstrate the effectiveness of our proposed strategies and the versatility and general applicability of our fine-tuned models that incorporate these strategies. Core code is available at https://github.com/nlpdata/strategy/.

1 Introduction

Recent years have seen a growing interest in machine reading comprehension (MRC) (Rajpurkar et al., 2016; Choi et al., 2018; Kočiskỳ et al., 2018; Reddy et al., 2018). In this paper, we mainly focus on *non-extractive MRC* (Khashabi et al., 2018; Ostermann et al., 2018; Clark et al., 2018), in which a significant percentage of candidate answers are not restricted to text spans from the reference document or corpus. In comparison to *extractive MRC* tasks (Section 2.1), non-extractive MRC (Section 2.2) requires diverse reading skills and, as a result, the performance of machine readers on these tasks more accurately indicates the comprehension ability of machine readers in realistic settings such as exams (Lai et al., 2017).

Recently, significant progress has been achieved on many natural language processing tasks including MRC by fine-tuning a pre-trained general-purpose language model (Radford et al., 2018; Devlin et al., 2018). However, similar to the process of knowledge accumulation for human readers, it is time-consuming and resource-demanding to impart massive amounts of general domain knowledge from external corpora into a deep language model via pre-training. For example, it takes a month to pre-train a 12-layer transformer on eight P100 GPUs over the BooksCorpus (Zhu et al., 2015; Radford et al., 2018); Devlin et al. (2018) pre-train a 24-layer transformer using 64 TPUs for four days on the BooksCorpus plus English Wikipedia, a feat not easily reproducible considering the tremendous computational resources ($\approx$ one year to train on eight P100 GPUs).

From a practical viewpoint, given a limited number of training instances and a pre-trained

[*] This work was done when K. S. was an intern at the Tencent AI Lab, Bellevue, WA.

Proceedings of NAACL-HLT 2019, pages 2633–2643
Minneapolis, Minnesota, June 2 - June 7, 2019. ©2019 Association for Computational Linguistics

model, can we improve machine reading comprehension during fine-tuning instead of imparting more prior knowledge into a model via expensive pre-training? Inspired by reading strategies identified in cognitive science research that have been shown effective in improving comprehension levels of human readers, especially those who lack adequate prior knowledge of the topic of the text (Mokhtari and Sheorey, 2002; Mokhtari and Reichard, 2002; McNamara et al., 2004), we propose three corresponding domain-independent strategies to improve MRC based on an existing pre-trained transformer (Section 3.1):

- BACK AND FORTH READING (*"I go back and forth in the text to find relationships among ideas in it."*):
consider both the original and reverse order of an input sequence (Section 3.2)
- HIGHLIGHTING (*"I highlight information in the text to help me remember it."*):
add a trainable embedding to the text embedding of those tokens deemed relevant to the question and candidate answers (Section 3.3)
- SELF-ASSESSMENT (*"I ask myself questions I would like to have answered in the text, and then I check to see if my guesses about the text are right or wrong."*):
generate practice questions and their associated span-based candidate answers from the existing reference documents (Section 3.4)

By fine-tuning a pre-trained transformer (Radford et al., 2018) according to our proposed strategies on the largest general domain multiple-choice MRC dataset RACE (Lai et al., 2017) collected from language exams, we obtain a 5.8% absolute improvement in accuracy over the previous best result achieved by the same pre-trained transformer fine-tuned on RACE without the use of strategies (Section 4.2). We further fine-tune the resulting model on a target MRC task. Experiments show that our method achieves new state-of-the-art results on six representative non-extractive MRC datasets that require a range of reading skills such as commonsense and multi-sentence reasoning (i.e., ARC (Clark et al., 2016, 2018), OpenBookQA (Mihaylov et al., 2018), MCTest (Richardson et al., 2013), SemEval-2018 Task 11 (Yang et al., 2017), ROCStories (Mostafazadeh et al., 2016), and MultiRC (Khashabi et al., 2018)) (Section 4.4). These results indicate the effectiveness of our proposed strategies and the versatility and generality of our fine-tuned models that incorporate the strategies.

2 Task Introduction

We roughly categorize machine reading comprehension tasks into two groups: extractive (Section 2.1) and non-extractive (Section 2.2) based on the expected answer types.

2.1 Extractive MRC

Recently large-scale extractive MRC datasets have been constructed (Hermann et al., 2015; Hill et al., 2016; Onishi et al., 2016; Chen and Choi, 2016; Mostafazadeh et al., 2016; Bajgar et al., 2016; Nguyen et al., 2016; Joshi et al., 2017; Ma et al., 2018), such as SQuAD (Rajpurkar et al., 2016) and NewsQA (Trischler et al., 2017). Given a reference document and a question, the expected answer is a short span from the document. In contrast, answers in datasets such as SearchQA (Dunn et al., 2017) and NarrativeQA (Kočiský et al., 2018) are free-form human generated texts based on given documents (Nguyen et al., 2016; Reddy et al., 2018; Choi et al., 2018). However, since annotators tend to directly copy spans as answers, the majority of answers are still extractive (Reddy et al., 2018; Kočiský et al., 2018).

2.2 Non-Extractive MRC

In this section, we primarily discuss multiple-choice MRC datasets, in which answer options are not restricted to extractive text spans. Given a question and a reference document/corpus, multiple answer options are provided, and at least one of them is correct. It involves extensive human efforts to build such a dataset (e.g., MCTest (Richardson et al., 2013), SemEval-2018 Task 11 (Ostermann et al., 2018), MultiRC (Khashabi et al., 2018), and OpenBookQA (Mihaylov et al., 2018)) by crowdsourcing. Besides crowdsourcing, datasets such as RACE (Lai et al., 2017) and ARC (Clark et al., 2018) are collected from language or science exams designed by educational experts (Penas et al., 2014; Shibuki et al., 2014; Tseng et al., 2016) to evaluate the comprehension level of human participants. Compared to questions in extractive MRC tasks, besides surface matching, there are various types of complicated questions such as math word problems, summarization, logical reasoning, and sentiment analysis, requiring advanced read-

	RACE	ARC	OpenBookQA	MCTest	SemEval-2018 Task 11	ROCStories	MultiRC
construction method	exams	exams	crowd.	crowd.	crowd.	crowd.	crowd.
sources of documents	general	science	science	stories	narrative text	stories	mixed-domain
average # of answer options	4.0	4.0	4.0	4.0	2.0	2.0	5.4
# of documents	27,933	14M†	1,326†	660	2,119	3,742	871
# of questions	**97,687**	7,787	5,957	2,640	13,939	–	9,872
non-extractive answer⋆ (%)	87.0	43.3	83.8	45.3	89.9	100.0	82.1

Table 1: Statistics of multiple-choice machine reading comprehension datasets. Some values come from Reddy et al. (2018), Kočiský et al. (2018), and Lai et al. (2017) (crowd.: crowdsourcing; †: regarding each sentence/claim as a document (Clark et al., 2018); ⋆: correct answer options that are not text snippets from reference documents).

ing skills and prior world knowledge. Besides, in most cases, we can adopt an objective evaluation criteria such as accuracy to evaluate system performance (Clark et al., 2016; Lai et al., 2017). As these kind of datasets are relatively difficult to construct or collect, most existing datasets are small in size, which hinders the development of state-of-the-art deep neural models.

In response, in this paper we investigate how to make use of limited resources to improve MRC, using seven representative multiple-choice MRC datasets as case studies. As shown in Table 1, the majority of the correct answer options in most of the datasets (except for ARC and MCTest) are non-extractive. Except for MultiRC, there is exactly one correct answer option for each question. For ARC and OpenBookQA, a reference corpus is provided instead of a single reference document associated with each question.

Here we give a formal **task definition**. Given a reference document d, a question q, and associated answer options $\{o_1, o_2, \ldots, o_m\}$, the goal is to select the correct answer option(s). We can easily adapt our method to an MRC task that only provides a reference corpus (Section 4.4).

3 Approach

We first introduce a neural reader based on a pre-trained transformer (Section 3.1) and then elaborate on the strategies that are applied during the fine-tuning stage — back and forth reading (Section 3.2), highlighting (Section 3.3), and self-assessment (Section 3.4).

3.1 Framework Overview

Our neural reader follows the framework of discriminatively fine-tuning a generative pre-trained transformer (GPT) (Radford et al., 2018). It adapts a pre-trained multi-layer transformer (Vaswani et al., 2017; Liu et al., 2018) language model to a labeled dataset $\mathcal{C}$, where each instance consists of a sequence of input tokens $x^1, \ldots, x^n$, along with a label y, by maximizing:

$$\sum_{x,y} \log P(y \,|\, x^1, \ldots, x^n) + \lambda \cdot L(\mathcal{C}) \qquad (1)$$

where L is the likelihood from the language model, λ is the weight of language model, and $P(y \,|\, x^1, \ldots, x^n)$ is obtained by a linear classification layer over the final transformer block's activation of the language model. For multiple-choice MRC tasks, $x^1, \ldots, x^n$ come from the concatenation of a start token, a reference document, a question, a delimiter token, an answer option, and an end token; y indicates the correctness of an answer option. We refer readers to Radford et al. (2018) for more details.

Apart from placing a delimiter to separate the answer option from the document and question, the original framework pays little attention to task-specific structures in MRC tasks. Inspired by reading strategies, with limited resources and a pre-trained transformer, we propose three strategies to improve machine reading comprehension. We show the whole framework in Figure 1.

3.2 Back and Forth Reading (BF)

For simplicity, we represent the original input sequence of GPT during fine-tuning (Radford et al., 2018) as $[dq \,\$\, o]$, where [, $, and] represent the start token, delimiter token, and end token, respectively. Inspired by back and forth reading, we consider both the original order and the reverse order $[o \,\$\, qd]$. The token order within d, q, and o is still preserved. We fine-tune two GPTs that use $[dq \,\$\, o]$ and $[o \,\$\, qd]$ as the input sequence respectively, and then we ensemble the two models. We also consider other similar pairs of input sequences such as $[qd \,\$\, o]$ and $[o \,\$\, dq]$ in the experiments (Section 4.3).

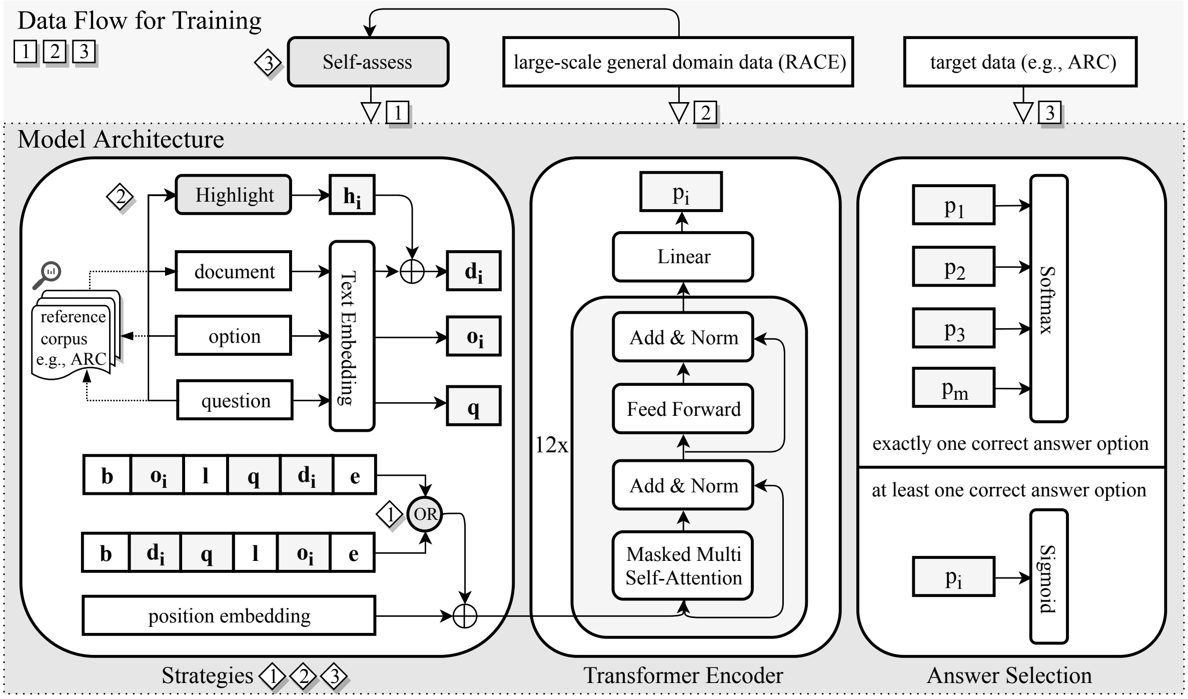

Figure 1: Framework Overview. Strategy 1, 2, and 3 refer to back and forth reading (BF) (Section 3.2), highlighting (HL) (Section 3.3), and self-assessment (SA) (Section 3.4), respectively.

3.3 Highlighting (HL)

In the original implementation (Radford et al., 2018), during the fine-tuning stage of GPT, the text embedding of a document is independent of its associated questions and answer options. Inspired by highlights used in human reading, we aim to make the document encoding aware of the associated question-answer option pair (q, o_i). We focus on the content words in questions and answer options since they appear to provide more useful information (Mirza and Bernardi, 2013), and we identify them via their part of speech (POS) tags, one of: noun, verb, adjective, adverb, numeral, or foreign word.

Formally, we let T be the set of POS tags of the content words. We let $\boldsymbol{d}$ denote the sequence of the text embedding of document d. We use d^j to represent the j^{th} token in d and $\boldsymbol{d^j}$ to denote the text embedding of d^j. Given d and a (q, o_i) pair, we define a *highlight embedding* $\boldsymbol{h_i^j}$ for the j^{th} token in d as:

$$\boldsymbol{h_i^j} = \begin{cases} \boldsymbol{\ell^+} & \text{if the POS tag of } d^j \text{ belongs to } T, \\ & \text{and } d^j \text{ appears in either } q \text{ or } o_i \\ \boldsymbol{\ell^-} & \text{otherwise} \end{cases} \qquad (2)$$

where $\boldsymbol{\ell^+}$ and $\boldsymbol{\ell^-}$ are two trainable vectors of the same dimension as $\boldsymbol{d^j}$.

Following the above definition, the sequence of the highlight embedding $\boldsymbol{h_i} = \boldsymbol{h_i^1}, \boldsymbol{h_i^2}, \ldots, \boldsymbol{h_i^n}$ is of the same length as $\boldsymbol{d}$. We replace $\boldsymbol{d}$ with $\boldsymbol{d_i} = \boldsymbol{d} + \boldsymbol{h_i}$ when we encode a document. More specifically, we use the concatenation of $\boldsymbol{b}$, $\boldsymbol{d_i}$, $\boldsymbol{q}$, $\boldsymbol{l}$, $\boldsymbol{o_i}$, and $\boldsymbol{e}$ as the new input of GPT during fine-tuning (Section 3.1), where $\boldsymbol{b}$, $\boldsymbol{l}$, and $\boldsymbol{e}$ denote the embedding of the start token, delimiter token, and end token, respectively, and $\boldsymbol{q}$ and $\boldsymbol{o_i}$ represent the sequence of the text embedding of q and o_i, respectively.

3.4 Self-Assessment (SA)

While in previous work (Radford et al., 2018), the original GPT is directly fine-tuned on an MRC end task, we instead develop a fine-tuning approach inspired by the self-assessment reading strategy. In particular, we propose a simple method to generate questions and their associated multiple span-based answer options, which cover the content of multiple sentences from a reference document. By first fine-tuning a pre-trained model on these *practice* instances, we aim to render the resulting fine-tuned model more aware of the input structure and to integrate information across multiple sentences as may be required to answer a given question.

Concretely, we randomly generate no more than n_q questions and associated answer options based on each document from the end task (i.e., RACE in this paper). We describe the steps as follows.

- **Input:** a reference document from the end task.
- **Output:** a question and four answer options associated with the reference document.

1. Randomly pick no more than n_s sentences from the document and concatenate these sentences together.
2. Randomly pick no more than n_c non-overlapping spans from the concatenated sentences. Each span randomly contains no more than n_t tokens within a single sentence. We concatenate the selected spans to form the correct answer option. We remove the selected spans from the concatenated sentences and use the remaining text as the question.
3. Generate three distractors (i.e., wrong answer options) by randomly replacing spans in the correct answer option with randomly picked spans from the document.

where n_q, n_s, n_c, and n_t are used to control the number and difficulty level of the questions.

4 Experiment

4.1 Experiment Settings

For most of the hyperparameters, we follow the work of Radford et al. (2018). We use the same preprocessing procedure and the released pretrained transformer. We generate 119k instances based on the reference documents from the training and development set of RACE (Lai et al., 2017), with $n_q = 10$, $n_s = 3$, $n_c = 4$, and $n_t = 4$ (Section 3.4). We first fine-tune the original pretrained model on these automatically generated instances with 1 training epoch (data flow 1 boxed in Figure 1). We then fine-tune the model on a large-scale general domain MRC dataset RACE with 5 training epochs (data flow 2 boxed in Figure 1). Finally, we fine-tune the resulting model on one of the aforementioned six out-of-domain MRC datasets (at max 10 epochs). See data flow 3 boxed in Figure 1. When we fine-tune the model on different datasets, we set the batch size to 8, language model weight λ to 2. We ensemble models by averaging logits after the linear layer. For strategy highlighting (Section 3.3), the content-word POS tagset $T = \{$NN, NNP, NNPS, NNS, VB, VBD, VBG, VBN, VBP, VBZ, JJ, JJR, JJS, RB, RBR, RBS, CD, FW$\}$, and we randomly initialize $\boldsymbol{\ell}^+$ and $\boldsymbol{\ell}^-$.

Approach		#	RACE-M	RACE-H	RACE
MMN (Tang et al., 2019)		9	64.7	55.5	58.2
GPT (Radford et al., 2018)		1	62.9	57.4	59.0
Human performance (Lai et al., 2017)		1	85.1	69.4	73.3
GPT*		1	60.9	57.8	58.7
		2	62.6	58.4	59.6
		9	63.5	59.3	60.6
GPT* + Strategies	SA	1	63.2	59.2	60.4
	HL	1	67.4	61.5	63.2
	BF	2	67.3	60.7	62.6
	SA + HL	1	**69.2**	**61.5**	**63.8**
	SA + HL + BF	2	**70.9**	**63.2**	**65.4**
	SA + HL + BF	9	**72.0**	**64.5**	**66.7**

Table 2: Accuracy (%) on the test set of RACE (#: number of (ensemble) models; SA: Self-Assessment; HL: Highlighting; BF: Back and Forth Reading; $\star$: our implementation).

4.2 Evaluation on RACE

In Table 2, we first report the accuracy of the state-of-the-art models (MMN and original fine-tuned GPT) and Amazon Turkers (Human performance). We then report the performance of our implemented fine-tuned GPT baselines and our models (GPT+Strategies). Results are shown on the RACE dataset (Lai et al., 2017) and its two subtasks: RACE-M collected from middle school exams and RACE-H collected from high school exams.

Our single and ensemble models outperform previous state-of-the-art (i.e., GPT and GPT (9×)) by a large margin (63.8% vs. 59.0%; 66.7% vs. 60.6%). The two single-model strategies – self-assessment and highlighting – improve over the single-model fine-tuned GPT baseline (58.7%) by 1.7% and 4.5%, respectively. Using the back and forth reading strategy, which involves two models, gives a 3.0% improvement in accuracy compared to the ensemble of two original fine-tuned GPTs (59.6%). Strategy combination further boosts the performance. By combining self-assessment and highlighting, our single model achieves a 5.1% improvement in accuracy over the fine-tuned GPT baseline (63.8% vs. 58.7%). We apply all the strategies by ensembling two such single models that read an input sequence in either the original or the reverse order, leading to a 5.8% improvement in accuracy over the ensemble of two original fine-tuned GPTs (65.4% vs. 59.6%).

To further analyze performance, we roughly divide the question types into five categories: detail (*facts and details*), inference (*reasoning ability*), main (*main idea or purpose of a document*), attitude (*author's attitude toward a topic or tone/source of a document*), and vocabulary (*vocabulary questions*) (Qian and Schedl, 2004; Lai et al., 2017) and annotate all the instances of the RACE development set. As shown in Figure 2, compared to the fine-tuned GPT baseline, our single-model strategies (SA and HL) consistently improve the results across all categories. Compared to other strategies, highlighting is likely to lead to bigger gains for most question types.

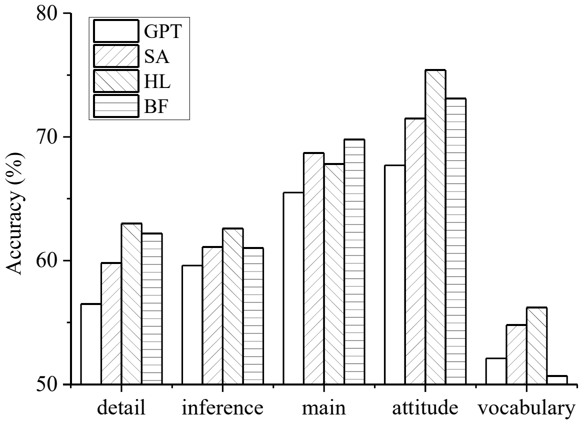

Figure 2: Performance on different question types.

Compared to human performance, there is still a considerable room for improvements, especially on RACE-M. We take a close look at the instances from the RACE-M development set that all our implementations fail to answer correctly. We notice that 82.0% of them require one or multiple types of world knowledge (e.g., negation resolution, commonsense, paraphrase, and mathematical/logic knowledge (Sugawara et al., 2017b,a, 2018)), especially when correct answer options are not explicitly mentioned in the reference document. For example, we need the knowledge — *the type of thing that is written by a writer can probably be a book* — to answer the question *"follow your heart is a _"* from the context *"Follow your heart by Andrew Matthews, an Australian writer, tells us that making our dreams real is life's biggest challenge"*. Besides, 19.7% of these failed instances require coreference resolution. It might be promising to leverage coreference resolvers to connect nonadjacent relevant sentences.

4.3 Further Discussions on Strategies

Besides the strategies introduced in Section 3, we also explore other reading strategies such as SUMMARIZATION (*"I take an overall view of the text to see what it is about before carefully reading it."*) by appending an extractive summary (Boudin et al., 2015) before each reference document, which is shown less effective for machine reading comprehension in our experiments compared to the strategies we focus on. In this section, we further discuss the three strategies.

Back and Forth Reading We notice that the input order difference between two ensemble models is likely to yield performance gains. Besides ensembling two models that use input sequence $[dq\,\$\,o]$ and $[o\,\$\,qd]$ respectively, we also investigate other reverse or almost reverse pairs. For example, we can achieve better results by ensembling $[qd\,\$\,o]$ and $[o\,\$\,dq]$ (61.0%) or $[qd\,\$\,o]$ and $[o\,\$\,qd]$ (61.7%), compared to the ensemble of two original fine-tuned GPTs (both of them use $[d\,\$\,qo]$) on the RACE dataset (59.6% in Table 2).

Highlighting We try two variants to define highlight embeddings (Equation 2 in Section 3.3) by considering the content of questions only or answer options only. Experiments show that using partial information yields a decrease in accuracy (60.6% and 61.0%, respectively) compared to 63.2% (Table 2), achieved by considering the content words in a question and its answer options. We attempt to also highlight the coreferential mentions of the content words, which does not lead to further gains, though.

Self-Assessment We explore alternative approaches to generate questions. For example, we use the Wikipedia articles from SQuAD (Rajpurkar et al., 2016) instead of the general domain documents from the end task RACE. We generate the same number of questions as the number of questions we generate using RACE following the same steps mentioned in Section 3.4. Experiments show that this method also improves the accuracy of the fine-tuned GPT baseline (59.7% vs. 58.7%). As self-assessment can be somehow regarded as a data augmentation method, we investigate other unsupervised question generation methods such as sentence shuffling and paraphrasing via back-translation (Ding and Zhou, 2018; Yu et al., 2018). Our experiments demonstrate that neither of them results in performance improvements on the RACE dataset.

Task	**Metric**	**Previous STOA**		**GPT**	**GPT (2×)**	**GPT+Strategies**	**GPT+Strategies (2×)**
ARC-Easy	Acc.	Clark et al. (2018)	62.6	57.0	57.1	66.6	68.9
ARC-Challenge	Acc.	Ni et al. (2018)	36.6	38.2	38.4	40.7	42.3
OpenBookQA	Acc.	Mihaylov et al. (2018)	50.2	52.0	52.8	55.2	55.8
MCTest-MC160	Acc.	Chung et al. (2018)	76.4	65.4	65.8	80.0	81.7
MCTest-MC500	Acc.	Chung et al. (2018)	72.3	61.5	61.0	78.7	82.0
SemEval	Acc.	Chen et al. (2018)	84.1	88.0	88.6	88.8	89.5
ROCStories	Acc.	Radford et al. (2018)	86.5	87.1	87.5	88.0	88.3
MultiRC	F1_m	Khashabi et al. (2018)	66.5	69.3	70.3	71.5	73.1
	F1_a	Khashabi et al. (2018)	63.2	67.2	67.7	69.2	70.5
	Acc.†	Khashabi et al. (2018)	11.8	15.2	16.5	22.6	21.8
Average	Acc.		60.1	58.1	58.5	**65.1**	**66.3**

Table 3: Performance (%) on the test sets of ARC, OpenBookQA, MCTest, SemEval-2018 Task 11, and ROCStories and the development set of MultiRC (Acc.: Accuracy; F1_m: macro-average F1; F1_a: micro-average F1; †: using the joint exact match accuracy (i.e., EM_0 reported by the official evaluation (Khashabi et al., 2018))). RACE is used as the source task for all our implementations.

Approach	**ARC**	**OpenBookQA**	**MCTest**	**SemEval**	**ROCStories**	**MultiRC**	**Average**
	Easy | Challenge	-	MC160 | MC500	-	-	-	-
	Acc.	Acc.	Acc.	Acc.	Acc.	F1_m | F1_a | Acc.†	Acc.
GPT	54.0 | 30.3	50.0	58.8 | 52.0	87.3	86.7	69.3 | 66.2 | 11.9	53.9
GPT (2×)	53.9 | 30.7	50.0	60.0 | 54.0	88.0	87.0	69.3 | 66.5 | 13.1	54.6
GPT+Strategies	61.9 | 35.0	54.2	67.5 | 64.7	87.6	87.4	68.8 | 67.4 | 16.2	**59.3**
GPT+Strategies (2×)	63.1 | 35.4	55.0	70.8 | 64.8	88.1	88.1	69.7 | 67.9 | 16.9	**60.3**

Table 4: Performance (%) on the test sets of ARC, OpenBookQA, MCTest, SemEval-2018 Task 11, and ROCStories and the development set of MultiRC using the target data only (i.e., without the data flow 1 and 2 boxed in Figure 1) (Acc.: Accuracy; F1_m: macro-average F1; F1_a: micro-average F1; †: using the joint exact match accuracy (i.e., EM_0 reported by the official evaluation (Khashabi et al., 2018))).

4.4 Adaptation to Other Non-Extractive Machine Reading Comprehension Tasks

We follow the philosophy of transferring the knowledge from a high-performing model pre-trained on a large-scale supervised data of a source task to a target task, in which only a small amount of training data is available (Chung et al., 2018). RACE has been used to pre-train a model for other MRC tasks as it contains the largest number of general domain non-extractive questions (Table 1) (Ostermann et al., 2018; Wang et al., 2018a). In our experiment, we also treat RACE as the source task and regard six representative non-extractive multiple-choice MRC datasets from multiple domains as the target tasks.

We require some task-specific modifications considering the different structures of these datasets. In ARC and OpenBookQA, there is no reference document associated with each question. Instead, a reference corpus is provided, which consists of unordered science-related sentences relevant to questions. We therefore first use Lucene (McCandless et al., 2010) to retrieve the top 50 sentences by using the non-stop words in a question and one of its answer options as a query. The retrieved sentences are used to form the reference document for each answer option. In MultiRC, a question could have more than one correct answer option. Therefore, we use a sigmoid function instead of softmax at the final layer (Figure 1) and regard the task as a binary (i.e., correct or incorrect) classification problem over each (document, question, answer option) instance. When we adapt our method to the non-conventional MRC dataset ROCStories, which aims at choosing the correct ending to a four-sentence incomplete story from two answer options (Mostafazadeh et al., 2016), we leave the question context empty as no explicit questions are provided. Since the test set of MultiRC is not publicly available, we report the performance of the model that achieves the highest micro-average F1 (F1_a) on the development set. For other tasks, we select the model that achieves the highest accuracy on the development set and report the accuracy on the test set.

We first fine-tune GPT using our proposed three strategies on RACE and further fine-tune the resulting model on one of the six target tasks (see Table 3). During the latter fine-tuning stage, besides the *highlighting* embeddings inherited from the previous fine-tuning stage, we also apply the strategy *back and forth reading*, and we do not consider *self-assessment* since the model has already benefited from the high-quality RACE instances during the first fine-tuning stage. We compare with the baselines that are first fine-tuned on RACE and then fine-tuned on a target task without the use of strategies, which already outperform previous state-of-the-art (SOTA) on four out of six datasets (OpenBookQA, SemEval-2018 Task 11, ROCStories, and MultiRC). By using the strategies, we obtain a 7.8% absolute improvement in average accuracy over the ensemble baseline (58.5%) and a 6.2% absolute improvement over previous SOTA (60.1%).

To further investigate the contribution of the strategies, we directly fine-tune GPT on a target task without using the labeled data in RACE (i.e., we only keep data flow 3 in Figure 1). Compared to the baseline that is fine-tuned without using strategies (54.6%), we obtain a 10.4% relative improvement in average accuracy (60.3%) and especially large improvements on datasets ARC, OpenBookQA, and MCTest (Table 4).

5 Related Work

5.1 Methods for Multiple-Choice Machine Reading Comprehension

We primarily discuss methods applied to large-scale datasets such as RACE (Lai et al., 2017). Researchers develop a variety of methods with attention mechanisms (Chen et al., 2016; Dhingra et al., 2017; Xu et al., 2018; Tay et al., 2018; Tang et al., 2019) for improvement such as adding an elimination module (Parikh et al., 2018) or applying hierarchical attention strategies (Zhu et al., 2018; Wang et al., 2018b). These methods seldom take the rich external knowledge (other than pre-trained word embeddings) into considerations. Instead, we investigate different strategies based on an existing pre-trained transformer (Radford et al., 2018) (Section 3.1), which leverages rich linguistic knowledge from external corpora and achieves state-of-the-art performance on a wide range of natural language processing tasks including machine reading comprehension.

5.2 Transfer Learning for Machine Reading Comprehension and Question Answering

Transfer learning techniques have been successfully applied to machine reading comprehension (Golub et al., 2017; Chung et al., 2018) and question answering (Min et al., 2017; Wiese et al., 2017). Compared to previous work, we simply fine-tune our model on the source data and then further fine-tune the entire model on the target data. The investigation of methods such as adding additional parameters or an L2 loss and fine-tuning only part of the parameters is beyond the scope of this work.

5.3 Data Augmentation for Machine Reading Comprehension Without Using External Datasets

Previous methods augment the training data for extractive machine reading comprehension and question answering by randomly reordering words or shuffling sentences (Ding and Zhou, 2018; Li and Zhou, 2018) or generating questions through paraphrasing (Yang et al., 2017; Yuan et al., 2017), which require a large amount of training data or limited by the number of training instances (Yu et al., 2018). In comparison, our problem (i.e., question and answer options) generation method does not rely on any existing questions in the training set, and the generated questions can involve the content of multiple sentences in a reference document.

6 Conclusions

Inspired by previous research on reading strategies for improved comprehension levels of human readers, we propose three strategies (i.e., back and forth reading, highlighting, and self-assessment), aiming at improving machine reading comprehension using limited resources: a pre-trained language model and a limited number of training instances. By applying the proposed three strategies, we obtain a 5.8% absolute improvement in accuracy over the state-of-the-art performance on the RACE dataset. By fine-tuning the resulting model on a new target task, we achieve new state-of-the-art results on six representative non-extractive MRC datasets from multiple domains that require a diverse range of reading skills. These results consistently indicate the effectiveness of our proposed strategies and the general applicability of our fine-tuned model that incorporates these strategies.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive suggestions. We thank Hai Wang, Chengzhu Yu, and Chao Weng for their useful discussions. We especially thank Chao for helping us speed up the release of the preprint[1] with technical supports. We thank Rishi Bommasani for proofreading the paper and Saku Sugawara for sharing annotations with us.

References

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: Booktest dataset for reading comprehension. *CoRR*, cs.CL/1610.00956v1.

Florian Boudin, Hugo Mougard, and Benoit Favre. 2015. Concept-based summarization using integer linear programming: From concept pruning to multiple optimal solutions. In *Proceedings of the EMNLP*, pages 17–21, Lisbon, Portugal.

Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Proceedings of the ACL*, pages 2358–2367, Berlin, Germany.

Yu-Hsin Chen and Jinho D Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the SIGDial*, pages 90–100, Los Angeles, CA.

Zhipeng Chen, Yiming Cui, Wentao Ma, Shijin Wang, Ting Liu, and Guoping Hu. 2018. HFL-RC system at SemEval-2018 Task 11: Hybrid multi-aspects model for commonsense reading comprehension. *CoRR*, cs.CL/1803.05655v1.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the EMNLP*, pages 2174–2184, Brussels, Belgium.

Yu-An Chung, Hung-yi Lee, and James Glass. 2018. Supervised and unsupervised transfer learning for question answering. In *Proceedings of the NAACL-HLT*, pages 1585–1594, New Orlean, LA.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *CoRR*, cs.CL/1803.05457v1.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the AAAI*, pages 2580–2586, Phoenix, AZ.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT*, Minneapolis, MN.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the ACL*, pages 1832–1846, Vancouver, Canada.

Peng Ding and Xiaobing Zhou. 2018. Ynu deep at semeval-2018 task 12: A bilstm model with neural attention for argument reasoning comprehension. In *Proceedings of The SemEval*, pages 1120–1123, New Orleans, LA.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new Q&A dataset augmented with context from a search engine. *CoRR*, cs.CL/1704.05179v3.

David Golub, Po-Sen Huang, Xiaodong He, and Li Deng. 2017. Two-stage synthesis networks for transfer learning in machine comprehension. *CoRR*, cs.CL/1706.09789v3.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the NIPS*, pages 1693–1701, Montreal, Canada.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the ICLR*, Caribe Hilton, Puerto Rico.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, cs.CL/1705.03551v2.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the NAACL-HLT*, pages 252–262, New Orleans, LA.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gáabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association of Computational Linguistics*, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the EMNLP*, pages 785–794, Copenhagen, Denmark.

[1] `https://arxiv.org/abs/1810.13441v1`.

Yongbin Li and Xiaobing Zhou. 2018. Lyb3b at semeval-2018 task 11: Machine comprehension task using deep learning models. In *Proceedings of the SemEval*, pages 1073–1077, New Orleans, LA.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by summarizing long sequences. In *Proceedings of the ICLR*, Vancouver, Canada.

Kaixin Ma, Tomasz Jurczyk, and Jinho D Choi. 2018. Challenging reading comprehension on daily conversation: Passage completion on multiparty dialog. In *Proceedings of the NAACL-HLT*, pages 2039–2048, New Orleans, LA.

Michael McCandless, Erik Hatcher, and Otis Gospodnetic. 2010. *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Manning Publications Co., Greenwich, CT.

Danielle S McNamara, Irwin B Levinstein, and Chutima Boonthum. 2004. iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments, & Computers*, 36(2):222–233.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the EMNLP*, pages 2381–2391, Brussels, Belgium.

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. In *Proceedings of the ACL*, pages 510–517, Vancouver, Canada.

Paramita Mirza and Raffaella Bernardi. 2013. Ccg categories for distributional semantic models. In *Proceedings of the RANLP*, pages 467–474, Hissar, Bulgaria.

Kouider Mokhtari and Carla A Reichard. 2002. Assessing students' metacognitive awareness of reading strategies. *Journal of educational psychology*, 94(2):249.

Kouider Mokhtari and Ravi Sheorey. 2002. Measuring esl students' awareness of reading strategies. *Journal of developmental education*, 25(3):2–11.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. In *Proceedings of the NAACL-HLT*, pages 839–849, San Diego, CA.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, cs.CL/1611.09268v2.

Jianmo Ni, Chenguang Zhu, Weizhu Chen, and Julian McAuley. 2018. Learning to attend on essential terms: An enhanced retriever-reader model for open-domain question answering. In *Proceedings of the NAACL-HLT*, Minneapolis, MN.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did What: A large-scale person-centered cloze dataset. In *Proceedings of the EMNLP*, pages 2230–2235, Austin, TX.

Simon Ostermann, Michael Roth, Ashutosh Modi, Stefan Thater, and Manfred Pinkal. 2018. SemEval-2018 Task 11: Machine comprehension using commonsense knowledge. In *Proceedings of the SemEval*, pages 747–757, New Orleans, LA.

Soham Parikh, Ananya Sai, Preksha Nema, and Mitesh M Khapra. 2018. Eliminet: A model for eliminating options for reading comprehension with multiple choice questions. In *Proceedings of the IJCAI-ECAI*, pages 4272–4278, Stockholm, Sweden.

Anselmo Penas, Yusuke Miyao, Alvaro Rodrigo, Eduard H Hovy, and Noriko Kando. 2014. Overview of CLEF QA Entrance Exams Task 2014. In *Proceedings of the CLEF*, pages 1194–1200, Sheffield, UK.

David D Qian and Mary Schedl. 2004. Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, 21(1):28–52.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *Preprint*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the EMNLP*, pages 2383–2392, Austin, TX.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. CoQA: A conversational question answering challenge. *Transactions of the Association of Computational Linguistics*.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the EMNLP*, pages 193–203, Seattle, WA.

Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. 2014. Overview of the NTCIR-11 QA-Lab Task. In *NTCIR*.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the EMNLP*, pages 4208–4219, Brussels, Belgium.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017a. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the ACL*, pages 806–817, Vancouver, Canada.

Saku Sugawara, Hikaru Yokono, and Akiko Aizawa. 2017b. Prerequisite skills for reading comprehension: Multi-perspective analysis of mctest datasets and systems. In *Proceedings of the AAAI*, pages 3089–3096, San Francisco, CA.

Min Tang, Jiaran Cai, and Hankz Hankui Zhuo. 2019. Multi-matching network for multiple choice reading comprehension. In *Proceddings of the AAAI*, Honolulu, HI.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-range reasoning for machine comprehension. *CoRR*, cs.CL/1803.09074v1.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the RepL4NLP*, pages 191–200, Vancouver, Canada.

Bo-Hsiang Tseng, Sheng-Syun Shen, Hung-Yi Lee, and Lin-Shan Lee. 2016. Towards machine comprehension of spoken content: Initial toefl listening comprehension test by machine. In *Proceedings of the Interspeech*, San Francisco, CA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the NIPS*, pages 5998–6008, Long Beach, CA.

Liang Wang, Meng Sun, Wei Zhao, Kewei Shen, and Jingming Liu. 2018a. Yuanfudao at SemEval-2018 Task 11: Three-way attention and relational knowledge for commonsense machine comprehension. In *Proceedings of the SemEval*, pages 758–762, New Orleans, LA.

Shuohang Wang, Mo Yu, Shiyu Chang, and Jing Jiang. 2018b. A co-matching model for multi-choice reading comprehension. In *Proceedings of the ACL*, pages 1–6, Melbourne, Australia.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural domain adaptation for biomedical question answering. In *Proceedings of the CoNLL*, pages 281–289, Vancouver, Canada.

Yichong Xu, Jingjing Liu, Jianfeng Gao, Yelong Shen, and Xiaodong Liu. 2018. Dynamic fusion networks for machine reading comprehension. *CoRR*, cs.CL/1711.04964v2.

Zhilin Yang, Junjie Hu, Ruslan Salakhutdinov, and William Cohen. 2017. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the ACL*, pages 1040–1050, Vancouver, Canada.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. QANet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of the ICLR*, Vancouver, Canada.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the RepL4NLP*, pages 15–25, Vancouver, Canada.

Haichao Zhu, Furu Wei, Bing Qin, and Ting Liu. 2018. Hierarchical attention flow for multiple-choice reading comprehension. In *Proceedings of the AAAI*, pages 6077–6084, New Orleans, LA.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE ICCV*, pages 19–27, Santiago, Chile.